Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería

RIMNI

ARTICLE

Dynamic Variational Autoencoder Regressor for Soft Sensor Modeling

Hao Wu[1], Ping Zhang[1], Shuting Tian[1] and Zhiwei Chen[2,*]

[1]College of Artificial Intelligence, Beijing Polytechnic University, Beijing, China
[2]Department of Automation, Tsinghua University, Beijing, China
*Corresponding Author: Zhiwei Chen. Email: zhiwei-chen@mail.tsinghua.edu.cn

Received: 13 December 2025; Accepted: 02 March 2026

ABSTRACT: Due to the dynamics of industrial production, it is a great challenge to learn robust feature representations from industrial process data for building an accurate soft sensor model. The traditional variational autoencoder (VAE) can learn robust features that can adapt to the dynamic process better, but cannot be directly applied in soft sensor modeling. Therefore, a novel regression-based dynamic VAE (REG-DVAE) is introduced in this paper. Firstly, the encoder of the DVAE, which is constructed by the graph attention and the convolutional neural networks, is utilized to obtain the robust spatio-temporal features. The decoder of the DVAE is responsible for reconstructing the input data via transposed convolution. Secondly, the Transformer is employed to capture the dynamic associations between the robust spatio-temporal features and corresponding outputs. Moreover, with the purpose of guaranteeing the confidence of the proposed method, the Gaussian loss function is used as the optimization target of the REG-DVAE to enhance the confidence of each predicted value. The proposed REG-DVAE is implemented in a real-world melt index modeling of the polypropylene production process. The results show that compared with other baseline methods, the REG-DVAE not only can achieve the best performance, but also can provide a confidence interval, which greatly enhances the credibility of the prediction results.

KEYWORDS: Regression-based dynamic VAE; dynamic features; confidence interval; polypropylene production system

1 Introduction

Soft sensor technology can help reduce the emission of harmful gases through intelligent control [1]. The melt index (MI) serves as a critical rheological parameter for evaluating the process behavior and performance characteristics of polypropylene materials, and plays a key role in quality control and brand switching. However, the MI indicator is obtained by manual sampling and offline analysis in the laboratory, which poses a great difficulty to real-time monitoring of product quality. Therefore, the data-driven soft sensor modeling methods have begun to be used widely due to their flexibility, low cost, and high precision.

Traditionally, the data-driven soft sensor modeling includes shallow neural network-based methods and deep learning-based methods. In the beginning, the shallow neural networks found extensive applications in industrial process modeling. Among them, the extreme learning machine (ELM) is the most prominent since it can accelerate the training of neural networks by using a pseudo-inverse approach [2]. Zhang et al. [3] created a soft sensor model by combining partial least squares regression and ELM within an ensemble learning approach. Geng et al. [4] proposed a self-organizing ELM to solve the problem of improper selection of hidden nodes. He et al. [5] used the Pearson correlation coefficient to construct the interaction between the inputs and the outputs outside of the ELM. Shi et al. [6] derived the L1 norm based on the variational inference, which could

enhance the generalization of input weights. Zhang et al. [7] introduced a soft sensor method by combining the Monte Carlo and the ELM to predict the physical properties of nonwoven fabrics. Despite the good progress made by the ELM, it uses the pseudo-inverse matrix method to train the weights, which only changes the weights of the last layer, resulting in a weaker learning ability of the ELM. Therefore, the multilayer perceptron (MLP) is studied and applied. Souza et al. [8] put forward an iteratively adjusted MLP to select the input variables, which could reduce the number of parameters in the MLP. Sun et al. [9] introduced the least absolute shrinkage and selection operator to choose the process variables. On this basis, the non-negative garrote was proposed to adjust the weight of MLP [10]. da Silva Bispo et al. [11] applied the MLP in deep-sea oil exploration to estimate the apparent viscosity of water-based drilling fluids. However, with the continuous expansion of data scale, the feature extraction ability of shallow neural networks is relatively weak. Recently, owing to the significant advancements of deep learning in natural language processing (NLP) [12] and computer vision (CV) [13], a large number of deep neural networks have been constructed and applied in soft sensor modeling. Yuan et al. [14] put forward a dynamic convolutional neural network (CNN) method as a means to extract hierarchical, local, nonlinear, and dynamic features from process data for modeling. Zhao et al. [15] combined an improved CNN to construct a free calcium oxide content model. Liu et al. [16] proposed a non-local and local fused information extraction method to build the hydrocracking process model. Wu et al. [17] presented a dynamic convolutional model to capture both spatial cross-correlation and temporal autocorrelation. Jia et al. [18] reviewed the applications of graph neural networks in industrial processes. However, these methods are aimed at improving operational efficiency and safety, rather than the dynamic problem. Jia et al. [19] proposed an adversarial relationship graph learning soft sensor modeling method to avoid noise interference. Although the above methods have shown strong local feature extraction ability, they only consider local interaction features and cannot model the global interaction relationship between the process variables. Therefore, a novel Regression-based Dynamic Variational AutoEncoder (REG-DVAE) is established in this work. The encoder of the DVAE is constructed by dynamic graph attention (DGAT) and CNN to extract spatio-temporal features. The decoder of DVAE reconstructs the input data via transposed convolution in a robust Gaussian hidden representation space. Then, the Transformer is derived to learn the evolving associations between the extracted spatio-temporal features and the output variables. Meanwhile, a Gaussian loss function is adopted as the optimization objective of the REG-DVAE to provide confidence for each predicted result. Finally, the REG, which is based on the Markov blanket independence assumption, is used to guarantee the confidence of the proposed method. The effectiveness of the proposed REG-DVAE is verified by the actual polypropylene production process.

The contributions are threefold:

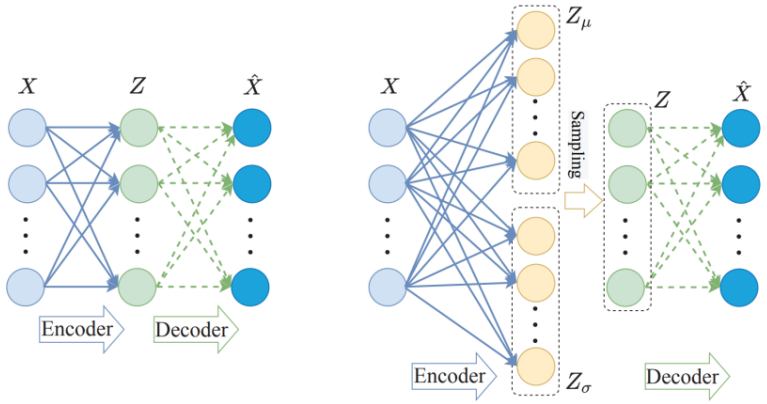
- The DVAE, combined with dynamic gated graph attention and CNN, is used to capture target-related spatio-temporal features for soft sensor modeling.
- The REG takes the spatio-temporal features extracted by the DVAE as its input to establish the dynamic relationships between inputs and outputs.
- The REG-DVAE is applied to the polypropylene production process to predict the MI, and the experiments illustrate that the proposed method can achieve outstanding performance when compared to other models.

2 Methodology

2.1 Regression-Based Dynamic Variational Autoencoder

The autoencoder (AE) [20], which consists of the encoder and decoder, is a framework for representing high-dimensional inputs as low-dimensional dense features, as shown in Fig. 1(a). The latent features are derived from the input data by the encoder, and then the decoder reconstructs the input data from these features, aiming to minimize the reconstruction loss. Building on the AE, the

variational autoencoder (VAE) [21] extends the original latent feature space to a Gaussian latent space, which can enhance the robustness of the feature representations, as shown in Fig. 1(b) The latent feature vectors are sampled from the Gaussian latent space to reconstruct the input data.



(**a**) (**b**)

Figure 1: The architecture of two different encoders. (**a**) Autoencoder; (**b**) Variational autoencoder.

For the VAE, the input data is transformed into the Gaussian latent representation space by encoder $p(C|X)$. In order to constrain the Gaussian latent representation space, the *KL* divergence between $p(C|X)$ and the prior distribution $q(C)$ needs to be minimized as shown in Eq. (1).

$$KL\left(p(C|X)\|q(C)\right)=\int\frac{1}{\sqrt{2\pi\sigma(x)^2}}e^{\frac{-(c-u(x))^2}{2\sigma(x)^2}}\left(\log\frac{\frac{-(c-u(x))^2}{e^{2\sigma(x)^2}\sqrt{2\pi\sigma(x)^2}}}{e^{\frac{-c^2}{2\sqrt{2\pi}}}}\right)dc$$
$$=\frac{1}{2}\left(-\log\sigma(x)^2+\mu(x)^2+\sigma(x)^2-1\right)d \tag{1}$$

$$L_c=\|x-\mu(c)\|^2 \tag{2}$$

Therefore, the loss function for the *VAE* is presented in Eq. (3).

$$L_{VAE}=\sum_i\left[\frac{1}{2}\left(-\log\sigma(x_i)^2\right)+\left(\mu(x_i)^2+\sigma(x_i)^2-1\right)+\sum_j\|x_i-\mu(c_j)\|^2\right] \tag{3}$$

The traditional *VAE* operates as an unsupervised generative model. Whereas, the soft sensor modeling is a supervised regression task. In order to apply the *VAE* framework to learn robust representation features in the Gaussian latent representation space and construct an accurate soft sensor model, the REG-DVAE is derived based on the Markov blanket independence assumption [22].

Suppose that the given label sample pair (X,Y), where X denotes the input data and Y represents the label. In order to extend the VAE to the regression task, it is necessary to establish a joint distribution $P(X,Y)$. The latent variable C is introduced into the joint distribution $p(X,Y)=\int p(X,Y,C)dC$ and the empirical distribution $q(X,Y,C)$ is used to approximate the joint distribution $p(X,Y,C)$. Therefore, the optimization goal is to minimize the *KL* divergence of $p(X,Y,C)$ and $q(X,Y,C)$ as shown in Eq. (4).

$$
\begin{aligned}
&KL\left(p(X,Y,C)\|q(X,Y,C)\right)\\
&=\iiint p(X,Y,C)In\frac{p(X,Y,C)}{q(X,Y,C)}dXdYdC\\
&=\int\tilde{p}(X)\left[\underbrace{\iint p(Y,C\mid X)In\frac{\tilde{p}(X)P(Y,C\mid X)}{q(X,Y,C)}dYdC}_{①}\right]dX
\end{aligned}
\tag{4}
$$

where the $\tilde{p}(X)$ is the empirical distribution determined by the samples [23]. The formula ① in Eq. (4) can be further simplified as shown in Eq. (5).

$$
\begin{aligned}
①=&\iint\underbrace{p(Y,C\mid X)In\tilde{p}(X)}_{②}\\
&\underbrace{+p(Y,C\mid X)In\frac{p(Y,C\mid X)}{q(X,Y,C)}dYdC}_{③}
\end{aligned}
\tag{5}
$$

where the formulas ② and ③ in Eq. (5) can be further simplified as Eqs. (6) and (7).

$$
\begin{aligned}
②&=\iint p(Y,C\mid X)In\tilde{p}(X)dYdC\\
&=In\tilde{p}(X)
\end{aligned}
\tag{6}
$$

$$
\begin{aligned}
③&=\iint p(Y,C\mid X)In\frac{p(Y,C\mid X)}{q(X,Y,C)}dYdC\\
&=\underbrace{\iint -p(Y,C\mid X)Inq(X,Y\mid C)dYdC}_{④}\\
&\underbrace{+\iint p(Y,C\mid X)In\frac{p(Y,C\mid X)}{q(C)}dYdC}_{⑤}
\end{aligned}
\tag{7}
$$

The graphical models of the VAE and the REG-DVAE are described in Fig. 2(a)and (b). The formulas ④ and ⑤ in Eq. (7) can be further simplified according to the Markov blanket independence assumption as shown in Eqs. (8) and (9). According to the Markov blanket independence assumption [22], given the *C*, the *X* and *Y* are independent. Therefore, $q(x,y\mid c)=q(x\mid c)q(y\mid c)$.

$$
\begin{aligned}
④&=\iint -p(Y,C\mid X)Inq(X\mid Y)dYdC\\
&\quad+\iint -p(Y,C\mid X)Inq(Y\mid C)dYdC
\end{aligned}
\tag{8}
$$

$$
\begin{aligned}
⑤&=\iint p(Y\mid C,X)Inp(C\mid X)In\frac{p(Y\mid C,X)p(C\mid X)}{q(C)}dYdC\\
&=\iint p(C\mid X)p(Y\mid C)Inq(Y\mid C)dYdC\\
&\quad+KL\left(p(C\mid X)\|q(C)\right)
\end{aligned}
\tag{9}
$$

Therefore, the form of formula ③ in Eq. (7) can be simplified as Eq. (10), and the *KL* divergence of $p(X,Y,C)$ and $q(X,Y,C)$ is simplified as Eq. (11).

$$
\begin{aligned}
③=&\int p(C\mid X)KL\left(p(Y\mid C)\|q(Y\mid C)\right)dC\\
&+KL\left(p(C\mid X)\|q(C)\right)-\int p(C\mid X)Inq(X\mid C)
\end{aligned}
\tag{10}
$$

$$
\begin{aligned}
&KL\left(p(X,Y,C)\|q(X,Y,C)\right)\\
&=\mathrm{E}_{\tilde{p}(X)}\left[②+③\right]\\
&=\mathrm{E}_{\tilde{p}(X)}\left[In\tilde{p}(X)\right]\\
&+\mathrm{E}_{\tilde{p}(X)}\left[\int p(C|X)KL\left(p(Y|C)\|q(Y|C)\right)dC\right]\\
&-\mathrm{E}_{\tilde{p}(X)}\left[\int p(C|X)Inq(X|C)\right]dC
\end{aligned}
\tag{11}
$$

where $\mathrm{E}_{\tilde{p}(X)}\left[In\tilde{p}(X)\right]$ is constant, and $KL(p(Y|C)\|q(Y|C))$ represents the distance between the empirical and true distribution. The formal expression of the true distribution $p(X|C)$ is not known. Therefore, this paper approximates the $KL(p(Y|C)\|q(Y|C))$ as a negative log-likelihood estimation of $q(Y|C)$ [24].

The loss function of the REG-DVAE includes the regression loss ℓ_{reg}, the *KL* divergence loss ℓ_{KL} and the reconstructed loss ℓ_{rec} as shown in Eq. (12)

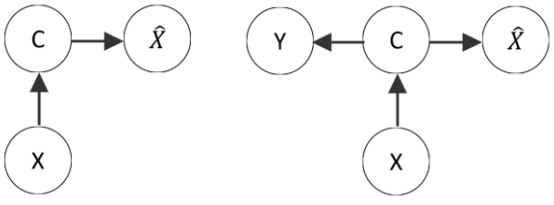
$$
\begin{aligned}
\ell &= \mathrm{E}_{\tilde{p}(X)}\left[\int -\sum_i In\left[q(Y_i|C)\right]p(C|X)dC\right]\\
&+\mathrm{E}_{\tilde{p}(X)}\left[KL\left(p(C|X)\right)\|q(C)\right]\\
&+\mathrm{E}_{\tilde{p}(X)}\left[\int p(C|X)-Inq(X|C)dC\right]\\
&=\ell_{reg}+\ell_{KL}+\ell_{rec}
\end{aligned}
\tag{12}
$$

$$
\begin{aligned}
\ell_{reg} &= \sum_i^M\sum_j^N\sum_k^M\frac{1}{2MNM}\left[\|\frac{y_i-\mu_{reg}\left(c_j^{xk}\right)}{\sigma_{reg}\left(c_j^{xk}\right)}\|^2+In\sigma_{reg}^2\left(c_j^{xk}\right)\right]\\
\ell_{KL} &= \sum_k^M\frac{1}{2M}\left(\mu_{en}^2(x_k)+\sigma_{en}^2(x_k)-In\sigma_{en}^2(x_k)-1\right)\\
\ell_{rec} &= \sum_j^N\sum_k^M\frac{1}{NM}\|x_k-\mu_{de}\left(c_j^{xk}\right)\|^2
\end{aligned}
\tag{13}
$$

where M denotes the number of training data, the c_j^{xk} denotes the samples taken from the conditional distribution $p(C|x_k)$, and N is the number of the c_j^{xk}. In REG-DVAE, $\{\mu_{reg}(C),\ \sigma_{reg}(C)\}$ can establish the mapping between the Gaussian latent representation space and the GL function. The loss function of the REG-DVAE consists of three terms ℓ_{KL}, ℓ_{rec}, and ℓ_{reg}. The ℓ_{KL} guarantees that the REG-DVAE has the ability to generate data. The ℓ_{rec} guarantees that the REG-DVAE can learn a robust Gaussian latent representation space. The ℓ_{reg} is the GL function that can reduce the learning cost of projecting from the Gaussian latent representation space into the target space. Unlike standard regression which assumes a constant variance, the Heteroscedastic Gaussian Regression is employed. the conditional probability distribution of the target variable y is constructed, given the input representation x, as a Gaussian distribution.

$$p(y|x;\theta)=\mathcal{N}\left(y\middle|\mu(x;\theta),\sigma^2(x;\theta)\right)$$

where θ represents the network parameters. Our model predicts both the mean μ and the variance σ^2 in the Regressor layers.



(**a**)VAE (**b**)REG-DVAE

Figure 2: The graphical models of VAE and REG-DVAE.

The architecture of the REG-DVAE is shown in Fig. 3. The encoder of the DVAE is a combination of the DGAT and the CNN. The DGAT employs the temporal gated linear unit for extracting the dynamic spatial interaction features of the inputs. The CNN captures the temporal dependency features from the input variables. The merging layer performs concatenation of temporal features and spatial interaction features to yield spatio-temporal interaction features. The $DGAT_m^E$, the $Conv_m^E$, and the merging layer make up the mth spatio-temporal module. The $DGAT_o$ maps the spatio-temporal features extracted by the last spatio-temporal module into the Gaussian latent representation space. The decoder of the DVAE is established by the transposed convolutional to reconstruct the input data, which can learn a robust Gaussian hidden representation space for spatio-temporal features. The REG constructed by Transformer [25] is utilized to build the dynamic relationships between the extracted spatiotemporal features and output variables, and fully connected layer is utilized to give the output. Moreover, the GL is used as the optimization target of the REG-DVAE to provide confidence for each predicted result. In this paper, the encoder of the DVAE contains 4 spatio-temporal modules and one $DGAT_o^E$, the decoder of the DVAE contains 4 transposed convolutional modules, and the REG contains 4 Transformer modules. The number of layers for each module is shown in Table 1.

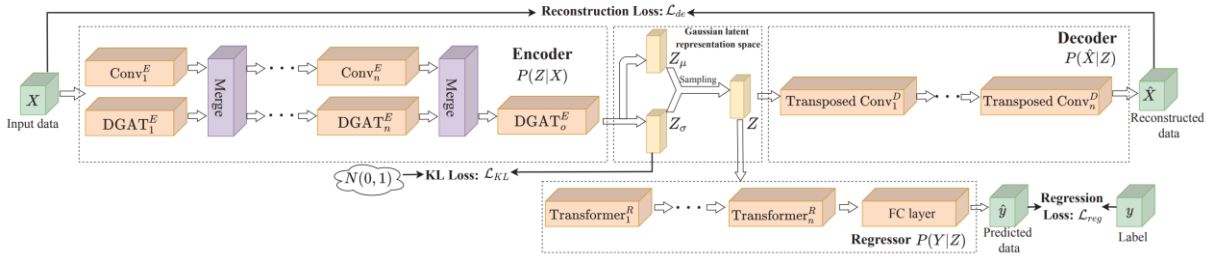


Figure 3: The architecture of the REG-DVAE.

Table 1: The parameter setting of the REG-DVAE.

Component	Name of Layer	Number of Layer
Encoder	Spatio-temporal module	4
Decoder	Transposed convolutional	4
REG	Transformer	4

2.2 Dynamic Graph Attention

The traditional GAT [26] calculates the neighborhood interaction relation through the attention mechanism and aggregates the neighborhood node features as shown in Fig. 4. The input of the GAT is a set of node hidden feature.

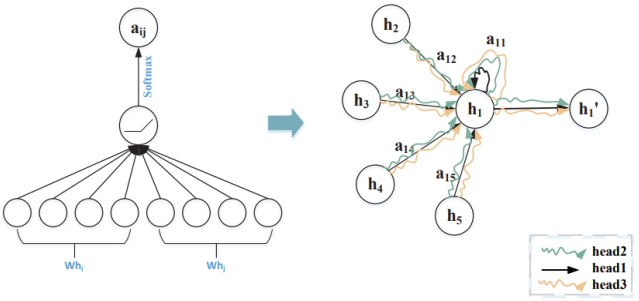


Figure 4: Aggregating the neighbor node features.

The core of the GAT is to compute the relationship weights of the neighborhoods. The GAT utilizes self-attention to calculate the coefficients matrix, which can assign different weights to the neighborhood of each node as shown in Eq. (14).

$$\alpha_{ij}^{loc} = \frac{\exp\left(Leaky\,\mathrm{Re}\,LU\left(W_0^T\left[W_1\vec{h}_i \,\|\, W_1\vec{h}_j\right]\right)\right)}{\sum_{k\in N_j}\exp\left(Leaky\,\mathrm{Re}\,LU\left(W_0\left[W_1\vec{h}_i \,\|\, W_1\vec{h}_k\right]\right)\right)} \tag{14}$$

where W_0 and W_1 denote trainable weights, || represents concatenation operator, and N_j is the set of neighbor nodes of j. $W_0^T\left[W_1\vec{h}_i \,\|\, W_1\vec{h}_j\right]$ is the correlation of the $\vec{h}$ and $\vec{h}_j$. The neighborhood nodes feature is aggregated based on coefficient matrix as shown in Eq. (15).

$$\vec{h}_i' = \sigma\left(\sum_{j\in N_i}\alpha_{ij}W_2\vec{h}_j\right) \tag{15}$$

Similar to Transformer, the GAT also designs multi-head attention to capture richer feature information in different subspaces as shown in Eq. (16).

$$\vec{h}_i' = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in N_i}\alpha_{ij}^k W_2^k\vec{h}_j\right) \tag{16}$$

where the K represents the number of different subspaces.

The traditional GAT is limited to extracting per-time-step interaction features of process variables, ignoring the trend relationship. Thus, the proposed DGAT aims to combine the multi-scale interaction features of the process variables. The computational procedure for DGAT is outlined in Algorithm 1.

Algorithm 1: Dynamic Graph Attention

Input: Auxiliary variable feature vectors $H\in\mathbb{R}^{T\times M\times D}$, and auxiliary variable $X\in\mathbb{R}^{T\times M}$, where T is the time step, M is the number of auxiliary variable, and D is the dimension of the auxiliary variable feature vectors.

Output: The spatial interaction features of the auxiliary variables $H_{spatial}\in\mathbb{R}^{T\times M\times D}$

1: Initialize the spatial feature list spa_{lst}

2: Calculate the trend matrix A_{glob} based on Eq. (18)

3: for $t=0$ to T do

4: $h^t = H[t,:,:]$
5: Calculate the local interaction relation matrix A^t_{loc} at time t based on Eq. (14)
6: Calculate the fused multi-scale interaction matrix A^t based on Eq. (17)
7: Calculate the spatial features $\vec{h}^t_{spa}$ at time t based on Eq. (19)
8: Append $\vec{h}^t_{spa}$ to spa_{lst}
9: end for
10: Calculate the spatial features of the auxiliary variables $\vec{h}_{spa} = cat(spa_{lst})$, which concatenates spa_{lst} in time dimension.
11: return $\vec{h}_{spa}$

The DGAT architecture incorporates a novel dynamic gated linear unit that performs a fusion of the global trend and local relations at each time step, is demonstrated as Eq. (17) and Fig. 5.

$$\begin{aligned} U_t &= \sigma\left(A^t_{loc} W_{loc}\right) \\ A^t &= \left(U_t \odot A_{glob}\right) W_{global} + A_{glob} \end{aligned} \tag{17}$$

where the A_{glob} is the trend matrix of process variables, the A^t_{loc} corresponds to the local relation matrix, which can be obtained by Eq. (14), the A^t is the fused multi-scale interaction matrix, the $\odot$ denotes the Hadamard product. The Pearson correlation coefficient is employed for calculating the trend matrix A^{ij}_{glob} of the process variables in this work, as presented in Eq. (18).

$$A^{ij}_{glob} = \frac{T\sum_{t=0} x^i_t x^j_t - \sum_{t=0} x^t_i \sum_{t=0} x^t_j}{\sqrt{T\sum_t \left(x^t_i\right)^2 - \left(\sum_t x^t_i\right)^2}\sqrt{T\sum_t \left(x^t_j\right)^2 - \left(\sum_t x^t_j\right)^2}} \tag{18}$$

where the T denotes the time step of the input variables, and x^t_i and x^t_j denote the values of the *i*th and the *j*th variables, respectively. The neighborhood node features of the DGAT are aggregated as shown in Eq. (19).

$$\vec{h}'^t_i = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in N_i} A^t_{ij} W^k_2 \vec{h}_j\right) \tag{19}$$

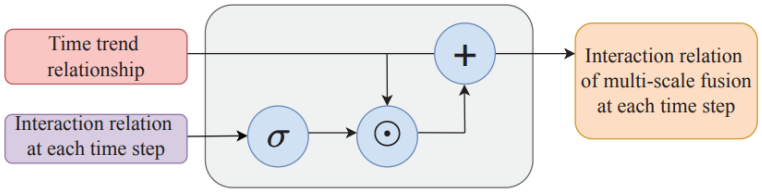


Figure 5: The temporal gated linear unit.

2.3 The Transformer

The Transformer [25] is a strong sequence feature extractor, which inputs are formally defined as three parts with query, key and value. When the query, key and value are the same, the attention is called self-attention. Otherwise, it is called vanilla attention. The attention function is shown as Eq. (20).

$$Attention\left[Q,K,V\right]=soft\max\left(\frac{QK^{T}}{\sqrt{d_{k}}}\right)V \tag{20}$$

The softmax function is shown in Eq. (21).

$$soft\max\left(y_{i}\right)=\frac{\exp\left(y_{i}\right)}{\sum_{j}\exp\left(y_{j}\right)} \tag{21}$$

where Q, K and V is query, key and value. d_k is the dimension of the input feature.

The query, key and value are projected by multi-head attention into multiple different spaces through multiple projection matrices. The query, key and value of different spaces are fed into the attention model as shown in Eq. (22). The multi-head attention is proved to be more beneficial for extracting features.

$$\begin{aligned} &MultiHead\left(Q,K,V\right)=Concat\left(head_{1},\cdots head_{h}\right)W^{o}\\ &head_{i}=Attention\left(QW_{i}^{Q},KW_{i}^{K},VW_{i}^{V}\right)\end{aligned} \tag{22}$$

The multi-head attention model concatenates the output vectors of the attention, which are the features of the query, key and value extracted in different spaces. Then, the vector of this connection is mapped again to get the final output.

2.4 The Overall Workflow of the Proposed Method

The overall workflow of the REG-DVAE based soft sensor modeling process consists of 5 steps:

Step1: The spatial features are captured by DGAT.

Step2: The temporal features are captured by CNN.

Step3: The robust spatio-temporal features are combined by DVAE which consist of DGAT and CNN.

Step4: The Transformer is utilized to build regression model by capturing the dynamic associations between the robust spatio-temporal features and corresponding outputs.

Step5: The REG which is based on the Markov blanket independence assumption is used to guarantee the confidence of the proposed method.

The clear framework diagram is outlined in Algorithm 2.

Algorithm 2: The overall workflow of the REG-DVAE

Input: Auxiliary variables $X\in\mathbb{R}^{T\times M}$, where T is the time step, M is the number of auxiliary variables.
Output: The soft sensor model.
1: Obtained spatial feature spa_{list} by the DGAT
2: Obtained temporal feature spa_{list} by the CNN
3: Combined the spatial feature and temporal feature as spatio-temporal features
4: Taken the spatial feature and temporal feature into the DVAE to obtain the robust spatio-temporal features.
5: Taken the robust spatio-temporal features to the Transformer to build the soft sensor model.

6: Guaranteed the confidence of the proposed method by REG.

3 Case Study

3.1 Polypropylene Production Process

Three widely adopted processes for polypropylene synthesis include the bulk, gas-phase, and slurry techniques. In the gas-phase polymerization variant, gaseous propylene polymerizes via direct contact with a catalyst held in suspension, which is visually represented in Fig. 6.

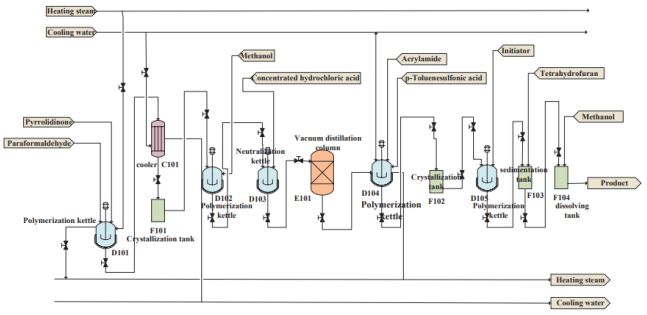


Figure 6: Flowchart for polypropylene production.

Central to the polypropylene manufacturing process is a reaction loop comprising the circulating gas compressor, gas-phase fluidized bed, and circulating gas cooler. The feed materials introduced into the reactor via the circulating gas include the T2 catalyst, liquid propylene, and an electron donor. Hydrogen is utilized in the polymerization reaction to regulate polypropylene's molecular weight; specifically, a higher hydrogen input leads to lower molecular weight and a corresponding rise in the MI. The MI quantifies the mass, in grams, of a thermoplastic material that extrudes through a standardized capillary tube (2.095 mm diameter) over a 10-min period under specific temperature and pressure conditions. Enhanced processing fluidity of the plastic corresponds directly to higher MI values. Utilizing data from the polypropylene production process, a soft sensor model was established for MI. The involved measurable process variables as inputs for the REG-DVAE model are shown in Table 2, the output is MI value.

Table 2: The parameter setting of the REG-DVAE.

Input Number	Description
No.1	Propylene dryer temperature
No.2	Propylene feed flow
No.3	Propylene feed temperature
No.4	Propylene degassing tower reflux flow
No.5	Inlet pressure of electron feed pump
No.6	Electron body feed pump outlet pressure
No.7	Density in the upper part of the bed
No.8	Density in the middle of the bed
No.9	Density in the lower part of the bed
No.10	Hydrogen to propylene concentration ratio
No.11	Propylene into vaporizer flow
No.12	Reactor bed temperature
No.13	Dew point temperature
No.14	Circulating water pump flow
Output	Melt index

Real-time process variable data is gathered directly via the distributed control system, whereas MI data originates from manual sampling procedures coupled with laboratory testing. The experimental data is collected from 2021-06 to 2021-11 in this work. The sliding window size is set to 40 in this experiment. Therefore, the size of input data X is Batch × 40 × 14, and the size of output data Y is Batch × 1.

3.2 Evaluation Metrics

The Root Mean Square Error (*RMSE*), Mean Absolute Error (*MAE*), Relative Squared Error (*RSE*) and Mean Absolute Percentage Error (*MAPE*) are taken as evaluation indicators, which are shown in Eqs. (23)–(26).

$$RMSE = \sqrt{\sum_{i}^{L} \frac{(y_i - \hat{y}_i)^2}{L}} \tag{23}$$

$$MAE = \sum_{i}^{L} \frac{|y_i - \hat{y}_i|}{L} \tag{24}$$

$$RSE = \frac{\sqrt{\sum_{i=1}^{L} (y_i - \hat{y}_i)^2}}{\sqrt{\sum_{i=1}^{L} (\bar{y} - y_i)^2}} \tag{25}$$

$$MAPE = \frac{1}{L} \sum_{i}^{L} \frac{|y_i - \hat{y}_i|}{|y_i|} \tag{26}$$

where the y_i denotes label, $\hat{y}_i$ denotes predicted value, $\bar{y}$ denotes the mean, and L represents the number of test sample sets. The $RMSE$, MAE, RSE, and the $MAPE$ denote the prediction errors from different perspectives. The smaller their values, the better the soft sensor performance. In order to ensure fair comparisons across all baseline methods, all experiments are conducted on CentOS v7.0, CPU (Intel (R) Xeon (R) Silver 4210@2.20 GHz) and GPU (TITAN XP, 12 GB). Apart from this, there are no other hyperparameters in the experimental.

3.3 Experimental Result

Compared with the MLP [27], the ELM [28], the LSTM [29], the CNN+LSTM [30], the Gated Convolutional neural network-based Transformer (GCT) [31], the causality-informed variational autoencoder (CIVAE) [32], the Temporal Convolutional Network (TCN) [33], the maximal relevance and minimal redundancy-based representation learning (MRMRRL) [34] and the gated stacked target-supervised VAE with variable weights (W-GSTVAE) [35]. The REG-DVAE shows excellent prediction error, which are denoted in Table 3. Compared with the MLP, the LSTM has 16%, 13%, 28%, and 12% drops in the RMSE, the MAE, the RSE, and the MAPE, respectively. Therefore, the deep learning models are better than the shallow neural network models. Compared with the CNN+LSTM, the TCN and the GCT, the REG-DVAE has 12.9%, 11% and 14.5% drops in the RSE, which illustrates the proposed model is preceding the general sequence models. Compared with the CIVAE, W-GSTVAE and more resent representative baseline MRMRRL, the REG-DVAE achieves state-of-the-art results and shows about 1.6%, 1.1% and 0.5% drops in the RMSE, respectively. The comparative experiment proves the effectiveness of the REG-DVAE. The statistical histogram of RSE of the REG-DVAE are shown in Fig. 7.

Table 3: Comparative experimental results for the REG-DVAE.

Model	RMSE	MAE	RSE	MAPE
ELM	0.275 ± 0.05	0.229 ± 0.04	0.625 ± 0.08	0.074 ± 0.03

MLP	0.134 ± 0.06	0.097 ± 0.05	0.490 ± 0.09	0.031 ± 0.03
LSTM	0.112 ± 0.03	0.084 ± 0.02	0.349 ± 0.04	0.027 ± 0.03
CNN+LSTM	0.110 ± 0.02	0.084 ± 0.05	0.349 ± 0.06	0.027 ± 0.02
TCN	0.105 ± 0.03	0.072 ± 0.02	0.330 ± 0.04	0.023 ± 0.03
GCT	0.114 ± 0.03	0.087 ± 0.03	0.365 ± 0.05	0.027 ± 0.02
CIVAE	0.103 ± 0.04	0.079 ± 0.02	0.333 ± 0.06	0.025 ± 0.01
MRMRRL	0.092 ± 0.01	0.087 ± 0.01	0.240 ± 0.04	0.026 ± 0.02
W-GSTVAE	0.098 ± 0.02	0.090 ± 0.01	0.220 ± 0.06	0.018 ± 0.01
Ours	0.087 ± 0.01	0.060 ± 0.01	0.220 ± 0.04	0.019 ± 0.01

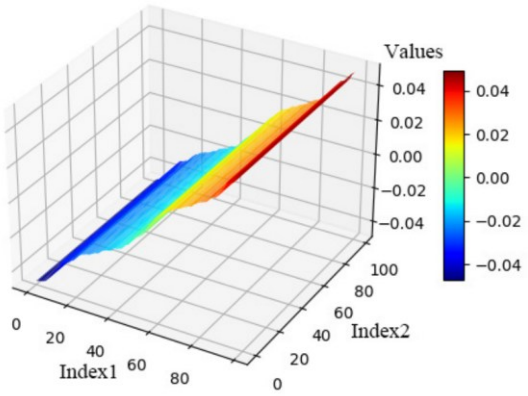


Figure 7: The errors of the REG-DVA.

The errors of the REG-DVAE are between −0.04 and 0.04, which shows the accuracy and effectiveness of the REG-DVAE. The significant improvement shows that the proposed model can successfully learn effective dynamic features of industrial data and accurately predict key indicators.

As illustrated in Fig. 8, the orange line represents actual values and the blue line represents the corresponding predicted values. Unlike standard regression, this approach models heteroscedasticity, allowing the model to dynamically output wider intervals in regions with higher data noise. The confidence interval is notably narrower when the prediction error is small, demonstrating that the learned variance serves as a reliable indicator of prediction confidence. However, it is important to acknowledge the limitations of this Gaussian uncertainty modeling. It relies on the assumption that the conditional distribution of the target is unimodal and Gaussian. Consequently, it may not fully capture complex multimodal distributions where multiple distinct future outcomes are possible. Despite this limitation, the results confirmed the effectiveness of the proposed Gaussian Likelihood (GL) method in the current scenario.

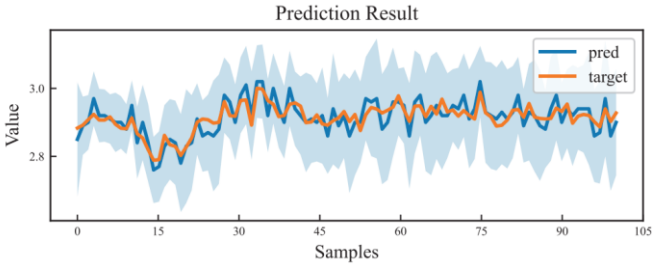


Figure 8: The prediction results of the REG-DVAE.

In Fig. 9, the input data are taken to compare with the reconstructed data. From the Fig. 9, It must be admitted that the reconstructed and original data distributions are nearly indistinguishable,

which indicates that the REG-DVAE can learn a robust Gaussian latent representation space with the GL as the optimization objective.

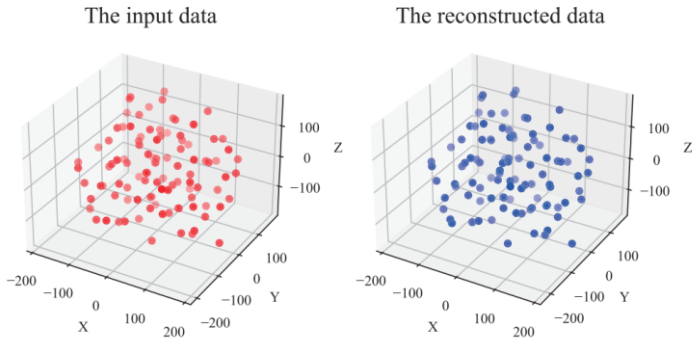


Figure 9: Comparison between the original input and reconstructed input.

4 Ablation Study

Table 4 summarizes the ablation experiment, which designed to evaluate the impact of individual REG-DVAE components on the prediction outcomes. The -ST, the -GL, and the -VAE mean to remove the spatio-temporal module, the GL function, and the VAE. After removing the spatio-temporal module, the encoder of the DVAE only contains the MLP to extract features of the auxiliary variables. Compared with the REG-DVAE, the -ST has 50%, 46%, 242%, and 52% increase in the RMSE, the MAE, the RSE, and the MAPE, respectively. After removing the GL function, the REG needs to learn the mapping from Gaussian latent space to Euclidean space, which increases the learning cost of different feature space transformations. Therefore, in comparison to the REG-DVAE, the -GL has 13%, 25%, 152%, and 26% increase in the RMSE, the MAE, the RSE, and the MAPE, respectively. Once the VAE module was excluded, the -VAE contains only the encoder of the DVAE and the REG. In comparison to the REG-DVAE, the -VAE has 25%, 33%, 126%, and 36% increase in the RMSE, the MAE, the RSE, and the MAPE, respectively. The ablation study revealed that removing the spatio-temporal module from the REG-DVAE resulted in the most substantial degradation in performance. Therefore, the spatio-temporal features play a crucial role in soft sensor modeling. The RMSE of the -GL is larger than that of -VAE, which shows that learning the feature representation in the Gaussian space is more effective for the soft sensor model than implementing the transformation in the same domain.

Table 4: Performance evaluation of REG-DVAE through ablation studies.

Ablation Par	RMSE	MAE	RSE	MAPE
-ST	0.131	0.088	0.753	0.029
-GL	0.099	0.075	0.555	0.024
-VAE	0.109	0.080	0.499	0.026
Ours	0.087	0.060	0.220	0.019

5 Conclusion

The novel REG-DVAE is proposed to build an MI soft sensor model in this paper. The encoder of the DVAE is constructed by the dynamic graph attention (DGAT) and CNN to extract the spatio-temporal features. The decoder of the DVAE reconstructs the input data by a transposed convolution to learn the spatio-temporal features in a robust Gaussian hidden representation space. Then, the REG is derived based on the Markov blanket independence assumption, which captures the time-varying associations linking the extracted spatio-temporal features to the output values. The fully

connected layer of the REG-DVAE is utilized to estimate the key indicator. Meanwhile, a Gaussian loss function is used as the optimization objective of the REG-DVAE to provide confidence for each predicted result. Finally, the model is utilized to forecast the MI in an operational polypropylene manufacturing environment. Compared with the CIVAE, the REG-DVAE achieves the best performance, with the RMSE, MAE, RSE, and MAPE reduced by 15%, 31%, 35%, and 34%, respectively. The ablation experiment shows that the spatio-temporal module can get spatio-temporal interaction features of industrial process data and play a crucial role in improving the model accuracy. Simultaneously, learning the feature representation in the Gaussian space is more effective than implementing the transformation in the same domain.

In the future, we will focus on strategies for transferring soft sensor models across various operating regimes, with the goal of establishing a unified modeling framework that incorporates data from multiple working conditions.

Acknowledgement: This work is supported by the School-level key projects of Beijing Polytechnic University and the School-level major projects of Beijing Polytechnic University

Funding Statement: This work is supported by the School-level key projects of Beijing Polytechnic University (2025X007-KXZ) and the School-level major projects of Beijing Polytechnic University (2025X019-KXD).

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Hao Wu and Ping Zhang; methodology, Hao Wu; software, Hao Wu and Zhiwei Chen; validation, Hao Wu, Ping Zhang and Shuting Tian; formal analysis, Hao Wu and Zhiwei Chen; investigation, Hao Wu; resources, Hao Wu; data curation, Hao Wu; writing—original draft preparation, Hao Wu; writing—review and editing, Hao Wu; visualization, Hao Wu; supervision, Hao Wu; project administration, Hao Wu; funding acquisition, Hao Wu. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: "Not applicable." for studies not involving humans or animals.

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

Nomenclature

Symbol	Description
X	Input data
$\hat{X}$	Reconstructed input data
Y	True label/output target (melt index)
$\hat{y}$	Predicted value
Z	Latent feature variable in the VAE
(Z_{μ}, Z_{σ})	Mean and variance vectors in the Gaussian latent space
C	Latent variable introduced in the joint distribution
N_i	The set of neighbor nodes for node *i*
$P(Z \mid X)$	Encoder of DVAE
$P(X \mid Z)$	Decoder of DVAE
$P(Z)$	Prior distribution of the latent variable
$\tilde{P}(X)$	Empirical distribution of the input data
$D_{KL}(\cdot\|\cdot)$	Kullback-Leibler (KL) divergence
L_{reg}	Regression loss
L_{KL}	KL divergence loss
L_{de}	Reconstruction loss
$\sigma(\cdot)$	Activation function
θ	Network parameters

(μ, σ^2)	Mean and variance parameters predicted by the model
W	Trainable weight matrix in neural networks
a_{ij}	Attention coefficient between node *i* and node *j*
A_{glob}	Trend matrix of process variables
Q, K, V	Query, key, and value matrices in Transformer
d_k	Dimension of the input feature in Transformer
λ	A constant weight factor in the loss function
$\odot$	Hadamard product (element-wise multiplication)

References

1. Tan X, Rodrigue D. A review on porous polymeric membrane preparation. part II: production techniques with polyethylene, polydimethylsiloxane, polypropylene, polyimide, and polytetrafluoroethylene. Polymers. 2019;11(8):1310. doi:10.3390/polym11081310.
2. Huang GB, Zhu QY, Siew CK. Extreme learning machine: theory and applications. Neurocomputing. 2006;70(1–3):489–501. doi:10.1016/j.neucom.2005.12.126.
3. Zhang X, Zhu Q, Jiang ZY, He Y, Xu Y. A novel ensemble model using PLSR integrated with multiple activation functions based ELM: applications to soft sensor development. Chemom Intell Lab Syst. 2018;183:147–57. doi:10.1016/j.chemolab.2018.10.016.
4. 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. doi:10.1016/j.engappai.2017.03.011.
5. He YL, Geng ZQ, Zhu QX. Data driven soft sensor development for complex chemical processes using extreme learning machine. Chem Eng Res Des. 2015;102:1–11. doi:10.1016/j.cherd.2015.06.009.
6. Shi X, Kang Q, An J, Zhou M. Novel L1 regularized extreme learning machine for soft-sensing of an industrial process. IEEE Trans Ind Inf. 2022;18(2):1009–17. doi:10.1109/tii.2021.3065377.
7. Zhang J, Fan Y, Zhang L, Xu C, Dong X, Liu L, et al. Seepage time soft sensor model of nonwoven fabric based on the extreme learning machine integrating Monte Carlo. Sensors. 2021;21(7):2377. doi:10.3390/s21072377.
8. Souza FAA, Araújo R, Matias T, Mendes J. A multilayer-perceptron based method for variable selection in soft sensor design. J Process Control. 2013;23(10):1371–8. doi:10.1016/j.jprocont.2013.09.014.
9. Sun K, Huang SH, Jang SS, Wong DS. Development of soft sensor with neural network and nonlinear variable selection for crude distillation unit process. In: Proceedings of the 26th European Symposium on Computer Aided Process Engineering; 2016 Jun 12–15; Protoroz, Slovenia. p. 337–42.
10. Sun K, Sui L, Wang H, Yu X, Jang SS. Design of an adaptive nonnegative garrote algorithm for multi-layer perceptron-based soft sensor. IEEE Sens J. 2021;21(19):21808–16. doi:10.1109/JSEN.2021.3102586.
11. da Silva Bispo VD, Scheid CM, Calçada LA, da Cruz Meleiro LA. Development of an ANN-based soft-sensor to estimate the apparent viscosity of water-based drilling fluids. J Petrol Sci Eng. 2017;150:69–73. doi:10.1016/j.petrol.2016.11.030.
12. Otter DW, Medina JR, Kalita JK. A survey of the usages of deep learning for natural language processing. IEEE Trans Neural Netw Learning Syst. 2021;32(2):604–24. doi:10.1109/tnnls.2020.2979670.
13. O'Mahony N, Campbell S, Carvalho A, Harapanahalli S, Hernandez GV, Krpalkova L, et al. Deep learning vs. traditional computer vision. In: Advances in computer vision. Cham, Switzerland: Springer International Publishing; 2019. p. 128–44.
14. Yuan X, Qi S, Wang Y, Xia H. A dynamic CNN for nonlinear dynamic feature learning in soft sensor modeling of industrial process data. Control Eng Pract. 2020;104:104614. doi:10.1016/j.conengprac.2020.104614.
15. Zhao Y, Ding B, Zhang Y, Yang L, Hao X. Online cement clinker quality monitoring: a soft sensor model based on multivariate time series analysis and CNN. ISA Trans. 2021;117:180–95. doi:10.1016/j.isatra.2021.01.058.
16. Liu C, Wang Y, Wang K, Yuan X. Deep learning with nonlocal and local structure preserving stacked autoencoder for soft sensor in industrial processes. Eng Appl Artif Intell. 2021;104:104341. doi:10.1016/j.engappai.2021.104341.
17. Wu P, He Y, Li Y, Wang Y, Wang S. Online prediction of cutting temperature using self-adaptive local learning and dynamic CNN. IEEE Trans Ind Inform. 2022;18(12):8629–40. doi:10.1109/TII.2022.3152578.
18. Jia M, Yao Y, Liu Y. Review on graph neural networks for process soft sensor development, fault diagnosis, and process monitoring. Ind Eng Chem Res. 2025;64(17):8543–64.

19. Jia M, Yang C, Pan Z, Liu Q, Liu Y. Adversarial relationship graph learning soft sensor via negative information exclusion. J Process Control. 2025;145:103354. doi:10.1016/j.jprocont.2024.103354.
20. Meng Q, Catchpoole D, Skillicom D, Kennedy PJ. Relational autoencoder for feature extraction. In: Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN); 2017 May 14–19; Anchorage, AK, USA. p. 364–71.
21. Xie R, Jan NM, Hao K, Chen L, Huang B. Supervised variational autoencoders for soft sensor modeling with missing data. IEEE Trans Ind Inf. 2020;16(4):2820–8. doi:10.1109/tii.2019.2951622.
22. Khan W, Kong L, Noman SM, Brekhna B. A novel feature selection method via mining Markov blanket. Appl Intell. 2023;53(7):8232–55. doi:10.1007/s10489-022-03863-z.
23. Mohamed A, Qian K, Elhoseiny M, Claudel C. Social-STGCNN: a social spatio-temporal graph convolutional neural network for human trajectory prediction. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 14424–32.
24. Karp RM, Luby M, Madras N. Monte-Carlo approximation algorithms for enumeration problems. J Algoritms. 1989;10(3):429–48. doi:10.1016/0196-6774(89)90038-2.
25. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA. p. 5998–6008.
26. Chen Z, Ge Z. Knowledge automation through graph mining, convolution, and explanation framework: a soft sensor practice. IEEE Trans Ind Inform. 2022;18(9):6068–78. doi:10.1109/TII.2021.3127204.
27. Fan Y, Tao B, Zheng Y, Jang SS. A data-driven soft sensor based on multilayer perceptron neural network with a double LASSO approach. IEEE Trans Instrum Meas. 2020;69(7):3972–9. doi:10.1109/tim.2019.2947126.
28. Yao L, Ge Z. Deep learning of semisupervised process data with hierarchical extreme learning machine and soft sensor application. IEEE Trans Ind Electron. 2018;65(2):1490–8. doi:10.1109/TIE.2017.2733448.
29. Yuan X, Li L, Shardt YAW, Wang Y, Yang C. Deep learning with spatiotemporal attention-based LSTM for industrial soft sensor model development. IEEE Trans Ind Electron. 2021;68(5):4404–14. doi:10.1109/tie.2020.2984443.
30. Zheng J, Ma L, Wu Y, Ye L, Shen F. Nonlinear dynamic soft sensor development with a supervised hybrid CNN-LSTM network for industrial processes. ACS Omega. 2022;7(19):16653–64. doi:10.1021/acsomega.2c01108.
31. Geng Z, Chen Z, Meng Q, Han Y. Novel transformer based on gated convolutional neural network for dynamic soft sensor modeling of industrial processes. IEEE Trans Ind Inform. 2022;18(3):1521–9. doi:10.1109/TII.2021.3086798.
32. Jiang X, Ge Z. Improving the performance of just-in-time learning-based soft sensor through data augmentation. IEEE Trans Ind Electron. 2022;69(12):13716–26. doi:10.1109/TIE.2021.3139194.
33. Li B, Li J, Wu X, Tang H. Gated recurrent unit and temporal convolutional network with soft thresholding and attention mechanism for tool wear prediction. Measurement. 2025;240:115546. doi:10.1016/j.measurement.2024.115546.
34. Jin H, Dong X, Qian B, Wang B, Yang B, Chen X. Soft sensor modeling using deep learning with maximum relevance and minimum redundancy for quality prediction of industrial processes. ISA Trans. 2025;159:293–311. doi:10.1016/j.isatra.2025.02.010.
35. Qiu C, Huang FQ, Zhong YJ, Wu JZ, Li QL, Zhan CH, et al. Comparative analysis and application of soft sensor models in domestic wastewater treatment for advancing sustainability. Environ Technol. 2025;46(11):1959–80. doi:10.1080/09593330.2024.2415722.