A Novel Automated Soft Sensor Design Tool for Industrial Applications Based on Machine Learning

Liang Cao[a], Jianping Su[b], Emilio Conde[c], Lim C. Siang[d], Yankai Cao[e] and Bhushan Gopaluni[e,*]

[a]Department of Chemical Engineering, Massachusetts Institute of Technology, Boston, MA 02139, United States
[b]College of Carbon Neutrality Future Technology, China University of Petroleum, Beijing, 102200, China
[c]Seeq Corporation, Seattle, WA 98104, United States
[d]Parkland Refining (B.C.) Ltd, Department of Process Control Engineering, Burnaby Refinery, Burnaby, BC V5C 1L7, Canada
[e]Department of Chemical and Biological Engineering, University of British Columbia, Vancouver, BC V6T 1Z3, Canada

ARTICLE INFO

Keywords:
Automated Soft Sensors
Process Monitoring
Industrial Applications
Causal Machine Learning
Digital Transformation

ABSTRACT

In modern industrial processes, real-time monitoring and control of key quality variables are crucial but challenging due to measurement limitations and process complexities. Traditional methods for developing soft sensor models are not only time-consuming and labor-intensive but also require substantial expertise in machine learning, and often lack user-friendly interfaces, thereby limiting their accessibility to engineers in the field. To address these issues, this paper introduces an easy-to-use, open and efficient automated soft sensor design tool called Soft Sensor Manager. The Soft Sensor Manager incorporates advanced supervised, semi-supervised, and causal machine learning algorithms to enable effective model development and deployment. It also provides functionalities such as data preprocessing, feature engineering, algorithm selection, hyperparameter optimization, model evaluation and online deployment within a user-friendly interface. The software's effectiveness was demonstrated through its application in predicting light catalytic cracked oil yield using real industrial data. By automating the soft sensor design process, the Soft Sensor Manager enhances modeling efficiency and model quality, ultimately contributing to improved process monitoring and optimization in industrial settings.

1. Introduction

In modern industrial systems, real-time monitoring and control of key quality variables are essential for ensuring safety, efficiency, and product quality (Gopaluni et al., 2020; Lawrence et al., 2024; Cao, 2024). However, measuring many critical variables in real-time with physical sensors is often hindered by high costs or technical limitations (Russell E.L., 2000; Qin, 2014). This measurement gap poses significant challenges for effective process monitoring and optimization, potentially leading to suboptimal operations, increased safety risks, and missed opportunities for process improvement (Lou et al., 2022).

To address these challenges, soft sensors have emerged as a powerful tool in various process industries over the past few decades (Kadlec et al., 2009; Cao et al., 2020). These computational models leverage the rapid accumulation of industrial big data, enhanced computational capabilities, and advanced machine learning theories to estimate difficult-to-measure variables. Soft sensors have been successfully adopted in chemical, petroleum, steel, and pharmaceutical sectors, contributing significantly to process monitoring and optimization (Schaeffer and Braatz, 2022; Fan et al., 2014).

Despite these advancements, developing effective soft sensor models remains a challenging task (Chen et al., 2015; Kadlec and Gabrys, 2009). The process involves multiple complex stages, each requiring substantial domain expertise and data analysis skills. These stages include data acquisition, pre-processing, feature engineering, algorithm selection, and parameter tuning (Chen et al., 2015; Kadlec and Gabrys, 2009). Moreover, traditional soft sensor development is often time-consuming and labor-intensive, with a high potential for human error that affects model quality and reliability (Jiang et al., 2020). This complexity poses a significant barrier to widespread adoption, particularly for engineers who may lack specialized knowledge in machine learning.

Most soft sensor solutions currently available on the market are integrated into large-scale industrial automation and control systems, offered by major companies such as Siemens (Siemens), ABB (ABB), and Honeywell (Honeywell). While these integrated systems provide soft sensor functionalities, they present several limitations. Firstly, they often come with substantial costs, making them less accessible to small and medium-sized enterprises. Secondly, these systems typically exhibit slow algorithm update cycles, potentially lagging behind the latest advancements in the field. Moreover, as closed proprietary systems, they lack the flexibility for customization and expansion to meet specific user requirements.

Consequently, there is a pressing need for an open, efficient, and user-friendly automated soft sensor design tool that simplifies the development process and makes soft sensor technology more accessible(Karmaker et al., 2021; Real et al., 2020; Schaeffer and Braatz, 2022). This tool should have the following features: a user-friendly graphical interface to lower the usage barrier; an open system architecture

*Corresponding author
liangcao@mit.edu (L. Cao); jianping.su@cup.edu.cn (J. Su); emilio.conde@seeq.com (E. Conde); siang.lim@parkland.ca (L.C. Siang); yankai.cao@ubc.ca (Y. Cao); bhushan.gopaluni@ubc.ca (B. Gopaluni)
ORCID(s):

Table 1
Comparison of Soft Sensor Manager with Existing Commercial Solutions

Feature	Existing Commercial Solutions	Soft Sensor Manager
Algorithm Diversity	Limited (basic ML)	Wide range (supervised, semi-supervised, causal)
Openness	Closed source, difficult to customize	Open framework
Update Cycle	Slow (vendor-driven)	Flexible, user-managed updates
Cost	High licensing fees	Free core modules
Causal	Rarely included	Causal feature extraction
Version Control	Basic	Automated logging, multiple model versions

to support user customization and secondary development; a rich algorithm library that includes both traditional methods and intelligent algorithms; and comprehensive project management functions to support the development, maintenance, and iteration of models.

Automated Machine Learning (AutoML) approaches, which streamline end-to-end model development and automate hyperparameter tuning, have shown promise in reducing the time and expertise required for soft sensor development (Hutter et al., 2019; Salehin et al., 2024). Additionally, the incorporation of causal inference capabilities and domain knowledge into AutoML pipelines is expected to yield improved model interpretability and trustworthiness (Cao et al., 2022; Chan et al., 2024). Furthermore, the advent of deep learning, sensor fusion, and hybrid modeling techniques has enabled the development of advanced industrial soft sensors capable of real-time, data-driven estimation of process variables that are either infeasible or costly to measure directly (Yuan et al., 2018; Byrski et al., 2024). This work proposes an innovative automated soft sensor design and visualization software called Soft Sensor Manager. Figure 1 illustrates the overall architecture of Soft Sensor Manager, highlighting its key functional modules and their interactions. Table 1 provides a comparison between the proposed Soft Sensor Manager and existing commercial solutions, highlighting the key advantages of our approach in terms of openness, algorithm diversity, and causal learning capabilities.

The core contributions of this work are threefold. First, it provides a comprehensive open frameworks integrating supervised, semi-supervised, and causal machine learning algorithms for full-process automation of soft sensor design. While some commercial platforms provide limited machine learning capabilities, they often lack causal and semi-supervised methods and are not openly accessible. Our framework addresses this gap by offering an integrated platform for diverse modeling approaches. By incorporating these diverse approaches, the software effectively handles various data scenarios and captures complex relationships in industrial processes. Supervised learning algorithms form the foundation for predictive modeling, while semi-supervised techniques leverage both labeled and unlabeled data to enhance model performance. The inclusion of causal machine learning methods further improves model robustness and interpretability.

Second, it introduces novel unsupervised latent causal feature extraction (UCFE) and supervised causal feature extraction (SCFE) algorithms specifically designed for dynamic industrial processes. By incorporating temporal dynamics and causal relationships, UCFE and SCFE enhance model robustness and interpretability, addressing challenges of non-stationary industrial environments.

Third, from an engineering perspective, the Soft Sensor Manager significantly reduces the complexity of industrial data workflows. Its features, including the integrated data pipeline, automated hyperparameter tuning, model diagnostics, version control and one-click online deployment, lower the barrier for practitioners without specialized expertise to create and maintain high-quality models. This makes it easier for process engineers to use sophisticated soft sensor techniques without requiring extensive coding or statistical knowledge.

The remainder of this work is organized as follows. Section II introduces the design of the Soft Sensor Manager, highlighting its core functional modules and technical features. Section III discusses the machine learning techniques integrated into the automated soft sensor design process. Section IV demonstrates the practical application of Soft Sensor Manager in the catalytic cracking unit at the Parkland Refinery. Finally, Section V concludes the study, summarizing key findings, discussing the impact of software on industrial processes, and outlining potential future developments.

2. Automated Soft Sensor Design

To achieve intelligent and automated soft sensor technology, we designed Soft Sensor Manager. The Soft Sensor Manager is developed primarily in Python, leveraging libraries such as PyTorch for deep learning modelsPaszke et al. (2019), Bayesian optimization libraries (Optuna) for hyperparameter tuning(Akiba et al., 2019) and scikit-learn for traditional machine learning algorithms(Pedregosa et al., 2011). This software divides the soft sensor development process into five key stages: data processing, model selection and fitting, model evaluation and visualization, model saving and management, and online application. Figure 2 provides a visualization of the modular architecture of Soft Sensor Manager, which illustrates how the various components interact to streamline the entire soft sensor design process. The modular structure of the software supports the addition of new algorithms and custom preprocessing functions via plugins, fostering further innovation and collaboration within the community. To maximize the tool's impact, the

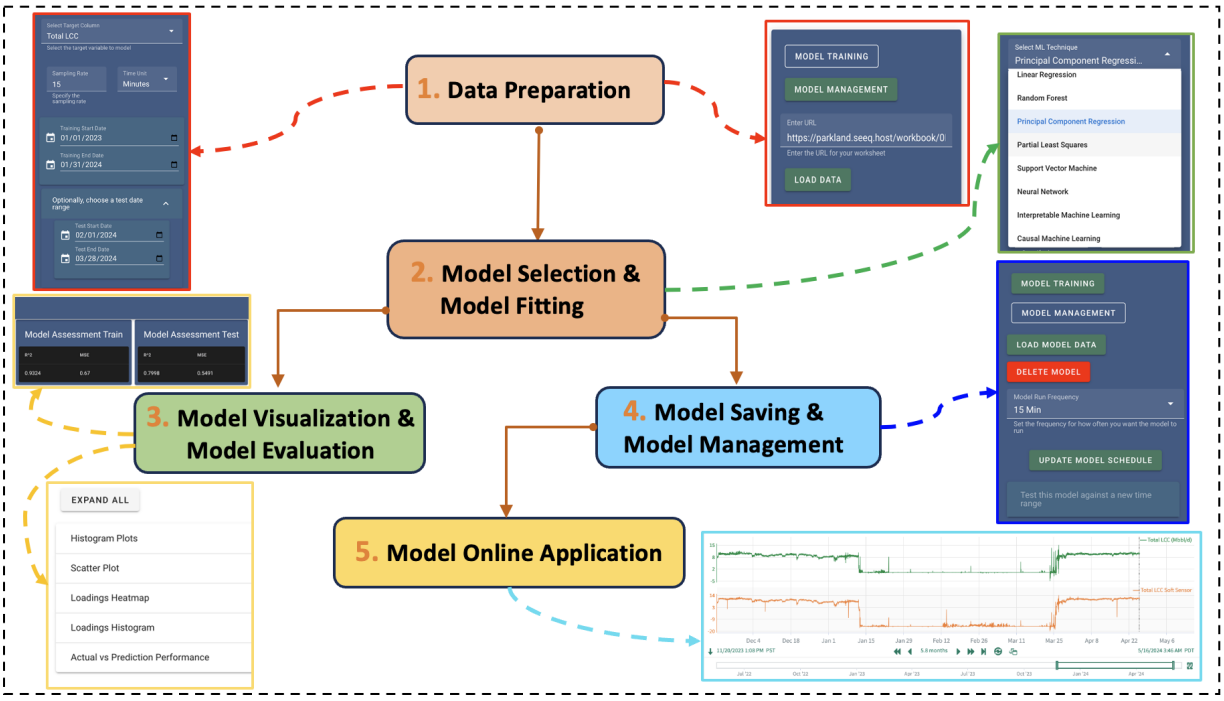

Figure 1: Overview of Automated Soft Sensor Manager's Structure

core framework is released on GitHub as open-source software. In the following section, we will provide a detailed introduction to the core functions and technical details of this software.

2.1. Data Import and Processing

Data import and processing are fundamental steps in constructing a soft sensor model. Soft Sensor Manager offers efficient data processing capabilities, providing reliable data support for subsequent modeling. Soft Sensor Manager supports multiple data import methods, including local file uploads, database connections, and online data acquisition. Users can automatically load and read the data by providing a link to the data source.

After data import, raw data often contain noise, missing values, and outliers, which are unfavorable for modeling. To address these issues, Soft Sensor Manager provides a variety of data cleaning and feature engineering tools. It can automatically identify and handle missing values and outliers. For multi-source data, users can set the time window for data alignment, and the system will automatically process data from different time scales, ensuring consistency on a unified timeline. It integrates various feature selection algorithms to identify the most valuable feature subsets(Tibshirani, 1996; D. Asir Antony Gnana Singh, 2016; Monirul Kabir et al., 2010). The software also allows users to set the sampling rates according to actual needs and to select the time range for the training and testing datasets through the interface.

To manage the challenge of data preprocessing across these diverse algorithms, Soft Sensor Manager employs a unified, yet flexible, approach. A modular data preprocessing pipeline is implemented where all data initially pass through a central preprocessing stage. This stage handles universal tasks, such as data cleaning, outlier detection, missing value imputation, and time-alignment, using a set of predefined "plug-in" functions. After this standardization, each algorithm-specific module can apply further transformations if required. For example, causal learning pipelines may compute lagged features, while autoencoder-based methods automatically normalize and reshape inputs. This design ensures both overall consistency and accommodation of individual algorithm needs.

Soft Sensor Manager seamlessly integrates with existing industrial data management systems. This integration enables direct access to extensive historical datasets, often spanning several months or even years, allowing users to leverage the vast amounts of data already collected and stored in these systems for model training.

2.2. Model Selection and Training

The core module of Soft Sensor Manager is model selection and training. This module integrates a variety of advanced machine learning algorithms, including regularized linear regression(Yu and Yao, 2017), random forest (RF) (Cheng et al., 2023), principal component regression (PCR)(Yuan et al., 2016), partial least squares (PLS)(Liu, 2014), support vector machine (SVM)(Shang et al., 2014), and attention-based neural network (ANN) (Cao et al., 2024a) algorithms, semi-supervised Gaussian process regression (SSGPR)(Esche et al., 2022), semi-supervised auto-encoders regression (SSAER)(Yuan et al., 2020), semi-supervised variational auto-encoders regression (SSVAER)(Zhuang et al., 2023) and causal machine learning(Cao et al., 2020, 2022; Yu et al., 2022). These

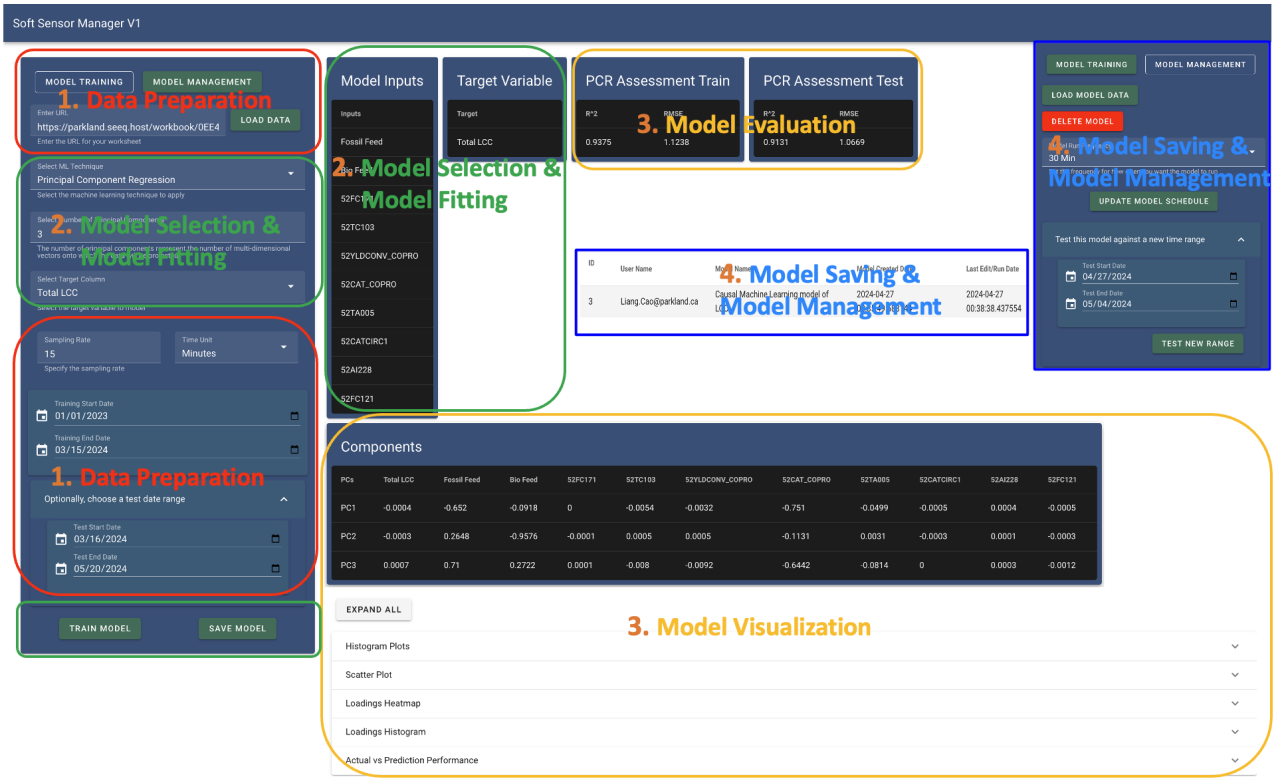

Figure 2: An Example of Modular design for Automated Soft Sensor Manager

algorithms range from simple linear models to complex nonlinear models, capable of addressing industrial modeling tasks with varying levels of complexity and characteristics.

While the base algorithms (such as Random Forest, SVM, and neural networks) integrated into Soft Sensor Manager are classical methods, our contribution lies not in reinventing these fundamental algorithms, but in adapting and integrating them into an automated framework specifically designed for the complexities of industrial process data. This includes incorporating features for time-series alignment, dynamic feature engineering, and outlier handling. For example, our implementation of Random Forest and neural networks allows for the incorporation of time-lagged features and handles partially labeled data, which is crucial for real-world industrial applications. Our implementation of semi-supervised learning algorithms include handling scenarios with a high proportion of unlabeled data and incorporating domain knowledge to guide the learning process, which enhances the model's ability to generalize from limited labeled data.

Users can manually select models and utilize visualization tools to analyze the applicability and performance of the models. This functionality helps users quickly identify the most suitable model for their data and requirements, thereby enhancing modeling efficiency. During the model training process, users can set hyperparameters such as learning rate, number of iterations, batch size, and more to further optimize model performance.

For hyperparameter tuning across these diverse algorithms, a Bayesian optimization engine is integrated(Akiba et al., 2019). This engine systematically searches for optimal hyperparameters (e.g., the number of estimators for Random Forest, kernel parameters for SVM, or the latent dimensions in SCFE). It operates within user-defined or process-specific constraints, such as limiting the maximum number of hidden layers in a neural network due to hardware or execution time limitations. A unified K-fold temporal cross-validation protocol is employed to ensure fair and consistent evaluation across all models while preserving the time-series dependencies inherent in industrial data.

2.3. Model Evaluation and Visualization

Model evaluation and visualization are essential stages in the development of soft sensor models. Through the visual interface, users can view key metrics such as training error and validation error. The software also supports model evaluation and validation, offering various evaluation metrics (e.g., root mean squared error, coefficient of determination) to help users comprehensively assess the predictive capabilities of their models.

Visualization is another major highlight of Soft Sensor Manager. To aid engineers in understanding the causal or critical relationships uncovered by the algorithms, Soft Sensor Manager provides a range of visualization and interpretability tools. For example, feature importance plots and partial dependence graphs help engineers understand the relative impact of different input variables on the model's

predictions. For models incorporating causal discovery techniques, causal graph visually represents the causal dependencies among process variables, providing deeper insights into the underlying process dynamics. Integration of SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) provides local explanations for individual predictions, allowing engineers to understand how specific input values contribute to the model's output. These tools collectively enhance the interpretability of the models, enabling engineers to gain a deeper understanding of the causal and critical relationships in their processes.

2.4. Model Saving and Management

Soft Sensor Manager includes a comprehensive model saving and management module. After model evaluation is completed, users can save the model within the system, which includes the trained algorithm, model parameters, and related metadata (such as time range for training and testing, training time, hyperparameter settings, etc.). The system automatically records training logs, performance metrics, and parameter settings for each model, facilitating model comparison and optimization for users.

Moreover, the software supports model version control and online updates, allowing soft sensor models to continuously learn and improve, adapting to ever-changing industrial environments. With each update or retraining of the model, the system generates a new version, allowing users to easily trace back and compare the performance of different versions. This version control mechanism not only enhances the flexibility of model management, but also provides strong support for model optimization and improvement.

Model management includes models' sharing and collaboration features. Soft Sensor Manager supports multi-user collaboration, allowing different team members to share and discuss models. Model security is also a critical aspect of management. Users can assign different access permissions based on roles, ensuring that only authorized persons can view, modify, and deploy models. These comprehensive features make Soft Sensor Manager a powerful tool for developing, managing, and deploying soft sensor models in various industrial applications.

2.5. Model Loading and Online Application

For models that need to be applied in real production environments, Soft Sensor Manager supports the deployment of models into real-time systems, enabling online predictions. This capability ensures that soft sensor models can respond quickly to changes in actual working conditions. To ensure the reliability and stability of models in production environments, Soft Sensor Manager also provides model monitoring and maintenance tools. Users can set performance indicators and alarm thresholds for the models and the system will automatically monitor the predictive performance of the models. If anomalies are detected, the system will notify users for intervention.

In long-term applications, soft sensor models need continuous updates and maintenance to adapt to changing process conditions and new data inputs. The software employs an incremental learning mechanism, where newly acquired data batches are periodically used to update model parameters. Users can set strategies for periodic training and updates. The system will retrain the models based on the latest data and deploy them to the production environment.

3. Machine Learning Techniques for Automated Soft Sensor Design

The development of soft sensors relies on advanced machine learning techniques capable of modeling the complex and nonlinear relationships inherent in industrial data. As depicted in Figure 3, a range of machine learning approaches, including supervised, semi-supervised, and causal methods, form the backbone of the automated soft sensor design process. This section explores the roles of these techniques in improving the accuracy and robustness of automated soft sensor models.

3.1. Soft Sensor Modeling

At the core of soft sensor development is the modeling of the relationship between measurable input variables and the target output variable. Let $\mathbf{x} \in \mathbb{R}^m$ denote the vector of input features, and $y \in \mathbb{R}$ represent the target variable that we aim to estimate. The objective is to learn a function $f : \mathbb{R}^m \to \mathbb{R}$ that maps the inputs to the output:

$$\hat{y} = f(\mathbf{x}; \boldsymbol{\theta}), \tag{1}$$

where $\hat{y}$ is the estimated output, and $\boldsymbol{\theta}$ is the parameter of the model. The function f can be linear or nonlinear, depending on the complexity of the relationship between $\mathbf{x}$ and y. The goal is to find the optimal parameters $\boldsymbol{\theta}^*$ that minimize a loss function ℓ, capturing the discrepancy between the predicted and true outputs.

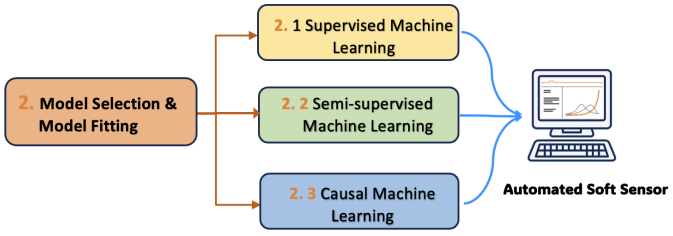

Figure 3: Machine Learning Techniques used in Automated Soft Sensors

3.2. Supervised Learning in Soft Sensor Development

Supervised learning is a fundamental approach in soft sensor modeling, where the model is trained using a labeled dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$. Each input vector $\mathbf{x}_i$ is associated

with a known output y_i. The learning process involves solving the following optimization problem:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, f(\mathbf{x}_i; \boldsymbol{\theta})\right) + \lambda \mathcal{R}(\boldsymbol{\theta}), \tag{2}$$

where $\ell(y_i, \hat{y}_i)$ is the loss function measuring the error between the true output y_i and the predicted output $\hat{y}_i$, $\mathcal{R}(\boldsymbol{\theta})$ is a regularization term to prevent overfitting, and λ is a hyperparameter controlling the regularization strength.

Common choices for the loss function in regression tasks include the mean squared error (MSE): $\ell(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$. The regularization term can be, for example, the L2 norm of the parameters: $\mathcal{R}(\boldsymbol{\theta}) = |\boldsymbol{\theta}|_2^2$. To solve the optimization problem in Equation (2), gradient-based optimization algorithms are commonly used. For instance, the gradient descent update rule is as follows:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^{(k)}), \tag{3}$$

where $\boldsymbol{\theta}^{(k)}$ is the parameter at iteration k, and $\boldsymbol{\theta}^{(k+1)}$ is the updated value. The learning rate η controls the step size, while $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^{(k)})$ represents the gradient of the loss function $\mathcal{L}$, guiding the parameters towards minimizing the empirical loss and regularization terms.

3.3. Semi-Supervised Learning in Soft Sensor Design

In industrial settings, acquiring labeled data can be expensive or impractical, whereas unlabeled data are often abundant. Semi-supervised learning leverages both labeled and unlabeled data to improve model performance (Lu and Chiang, 2018; Zhuang et al., 2023; Yuan et al., 2020; Esche et al., 2022). Let $\mathcal{D}_l = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_l}$ be the labeled dataset and $\mathcal{D}_u = \{\mathbf{x}_j\}_{j=n_l+1}^{n_l+n_u}$ be the unlabeled dataset. The semi-supervised learning objective combines the supervised loss on labeled data with an unsupervised loss on unlabeled data:

$$\begin{aligned}\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} & \frac{1}{n_l} \sum_{i=1}^{n_l} \ell\left(y_i, f(\mathbf{x}_i; \boldsymbol{\theta})\right) \\ & + \alpha \cdot \frac{1}{n_u} \sum_{j=1}^{n_u} \ell_{\text{unsup}}\left(f(\mathbf{x}_j; \boldsymbol{\theta})\right) + \lambda \mathcal{R}(\boldsymbol{\theta}),\end{aligned} \tag{4}$$

where ℓ_{unsup} is an unsupervised loss function encouraging the model to learn from the structure of the input data, and α is a hyperparameter balancing the influence of the unlabeled data. A common choice for the unsupervised loss is the consistency loss(Chang et al., 2021). By minimizing the consistency loss, the model is encouraged to be robust to small perturbations of the input data, i.e., to produce similar outputs for similar inputs:

$$\ell_{\text{unsup}}\left(f(\mathbf{x}_j; \boldsymbol{\theta})\right) = \mathbb{E}_{\boldsymbol{\xi}}\left[\left\| f(\mathbf{x}_j; \boldsymbol{\theta}) - f(\mathbf{x}_j + \boldsymbol{\xi}; \boldsymbol{\theta}) \right\|^2\right], \tag{5}$$

where $\boldsymbol{\xi}$ represents random perturbations applied to the input. By incorporating unlabeled data, semi-supervised learning can learn the underlying structure of the input data and improve the model's generalization ability, especially when labeled data are scarce.

3.4. Causal Machine Learning Based Soft Sensor

In complex industrial processes, traditional soft sensor approaches often struggle with the high dimensionality and multicollinearity of process variables. Although latent feature extraction methods have been widely used to address these issues (Yuan et al., 2016; Liu, 2014), they focus mainly on correlations rather than causal relationships. While existing causal discovery methods, such as the Peter-Clark (PC) algorithm (Spirtes et al., 2000) and the Greedy Equivalence Search (GES) (Chickering, 2002), have been applied to industrial processes, they typically treat variables as instantaneous rather than considering the temporal dynamics inherent in process systems. Furthermore, these methods often struggle with the scale and high noise levels characteristic of industrial data. These limitations can lead to reduced model robustness and poor generalization, especially when process conditions change (Xu et al., 2021).

To overcome these challenges, we propose a novel approach that integrates causal machine learning with latent feature extraction. This integration aims to identify and leverage causal relationships between variables, extracting latent features that are not only highly relevant, but also causally related to the target variable. In doing so, we seek to enhance model robustness, interpretability, and generalization, particularly in dynamic industrial environments. To this end, we introduce two novel methods: unsupervised latent causal feature extraction (UCFE) and supervised latent causal feature extraction (SCFE), which we will detail in the following sections.

Compared to existing methods, our approach offers several advantages. Traditional feature extraction methods like PCA maximize variance or correlation, potentially capturing spurious relationships. Recent deep learning approaches can learn complex nonlinear features but may not preserve causal structure. Our method explicitly optimizes for both predictive power and causal consistency, resulting in features that are both informative and robust under changing process conditions.

3.4.1. Unsupervised Latent Causal Feature Extraction

UCFE aims to extract latent features that reflect the causal structure of the process without using the target variable directly. The latent variables are defined as $\mathbf{Z} = \mathbf{W}^\top \mathbf{X}$, where $\mathbf{X} \in \mathbb{R}^{n \times m}$ is the input data matrix and $\mathbf{W} \in \mathbb{R}^{m \times p}$ is the loading matrix. In UCFE, the soft sensor model is constructed by selecting latent features that are believed to have causal influence on the target variable y. The model can be expressed as:

$$\hat{y} = f(\mathbf{Z}; \boldsymbol{\theta}), \tag{6}$$

where f is a predictive function, and $\boldsymbol{\theta}$ are the model parameters. Figure 4 illustrates the UCFE framework for soft sensor design. While PCA is used as an illustrative example, our Soft Sensor Manager also integrates other techniques such as slow feature analysis, autoencoders, and variational autoencoders. The choice of technique depends on the specific characteristics of the industrial data and the requirements of the application. To clarify the relationship between the latent variables and the target variable y, we note that in a Bayesian network, the Markov blanket of a target node y includes y's parents, children, and the parents of its children. In this example, z_4 is included in the equation $f(z_1, z_6, z_4)$ because it is a child of y in the causal graph, making it part of y's Markov blanket (Pearl, 2014).

While UCFE effectively reduces dimensionality and captures latent structures, it does not directly consider the target variable during feature extraction. This may result in features that are not optimally informative for predicting y.

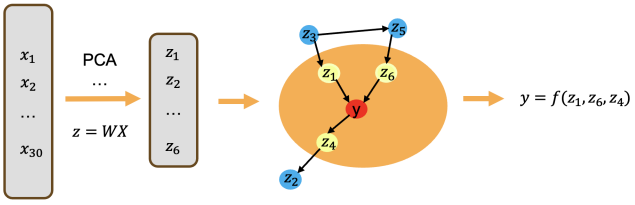

Figure 4: Framework of Unsupervised Latent Causal Feature Extraction for Soft Sensor Design

3.4.2. *Supervised Latent Causal Feature Extraction*

SCFE integrates the target variable y and temporal dynamics directly into the feature extraction process to identify latent features that have the most significant causal impact on y. By incorporating time-lagged relationships, SCFE ensures that the extracted features capture the dynamic behavior of the process and are highly relevant for predicting the target variable. Figure 5 shows the SCFE framework for soft sensor design.

In SCFE, we start by defining a latent variable t_k at time k as a linear combination of the process variables at that time:

$$t_k = \mathbf{w}^\top \mathbf{x}_k \tag{7}$$

where $\mathbf{x}_k \in \mathbb{R}^m$ is the vector of m process variables at time k, and $\mathbf{w} \in \mathbb{R}^m$ is the weight vector to be determined. To capture the temporal dynamics and causal relationships over time, we construct a latent feature u_k as a combination of the current and past latent variables:

$$u_k = \beta_1 t_k + \beta_2 t_{k-1} + \cdots + \beta_s t_{k-s+1} \tag{8}$$

where β_i are coefficients representing the influence of the latent variables at different time lags on the target variable, and s is the number of time lags considered.

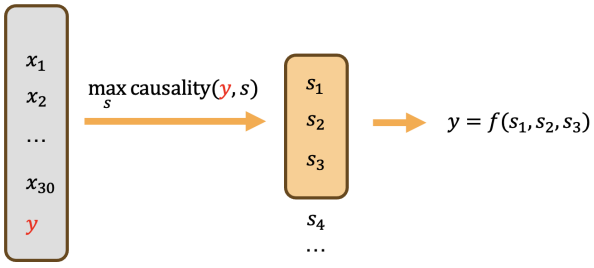

Figure 5: Framework of Supervised Latent Causal Feature Extraction for Soft Sensor Design

The objective of SCFE is to find $\mathbf{w}$ and $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_s]^\top$ that maximize the correlation between u_k and the target variable y_k, ensuring that the extracted features have a direct causal impact on y. This can be formulated as the following optimization problem:

$$\max_{\boldsymbol{\beta}, \mathbf{w}} \quad \frac{\sum_{k=s}^{n} y_k u_k}{\sqrt{\sum_{k=s}^{n} y_k^2} \sqrt{\sum_{k=s}^{n} u_k^2}} \tag{9}$$

subject to $\|\mathbf{w}\| = 1$ and $\sqrt{\sum_{k=s}^{n} u_k^2} = 1$, where n is the total number of samples. By including historical information in u_k, SCFE captures the dynamic causal relationships between the process variables and the target variable over time.

To solve the optimization problem, we reformulate it using data matrices. Let $\mathbf{X} \in \mathbb{R}^{n \times m}$ be the data matrix of the process variables, and $\mathbf{y} \in \mathbb{R}^n$ be the vector of target variables. We define the historical data matrices $\mathbf{X}_i$ as:

$$\mathbf{X}_i = [\mathbf{x}_i, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_{i+n-s}]^\top, \quad \text{for} \quad i = 1, 2, \ldots, s \tag{10}$$

Each $\mathbf{X}_i \in \mathbb{R}^{(n-s+1) \times m}$ contains the process variable data with a time lag of $s - i + 1$. We then construct the matrix $\mathbf{Z}$ that contains all the historical information:

$$\mathbf{Z} = [\mathbf{X}_s, \mathbf{X}_{s-1}, \ldots, \mathbf{X}_1] \in \mathbb{R}^{(n-s+1) \times ms} \tag{11}$$

The latent feature u_k can be expressed in matrix form:

$$\mathbf{u} = \mathbf{Z}(\boldsymbol{\beta} \otimes \mathbf{w}) \tag{12}$$

where $\mathbf{u} = [u_s, u_{s+1}, \ldots, u_N]^\top \in \mathbb{R}^{n-s+1}$, and $\otimes$ denotes the Kronecker product. The optimization objective (9) becomes:

$$\max_{\boldsymbol{\beta}, \mathbf{w}} \quad J = \frac{\mathbf{y}^\top \mathbf{u}}{\|\mathbf{y}\| \|\mathbf{u}\|} = \frac{\mathbf{y}^\top \mathbf{Z}(\boldsymbol{\beta} \otimes \mathbf{w})}{\|\mathbf{y}\| \|\mathbf{Z}(\boldsymbol{\beta} \otimes \mathbf{w})\|} \tag{13}$$

subject to $\|\mathbf{w}\| = 1$ and $\|\mathbf{Z}(\boldsymbol{\beta} \otimes \mathbf{w})\| = 1$. By maximizing J, we ensure that the latent feature $\mathbf{u}$ is highly correlated with $\mathbf{y}$, capturing the most significant causal relationships.

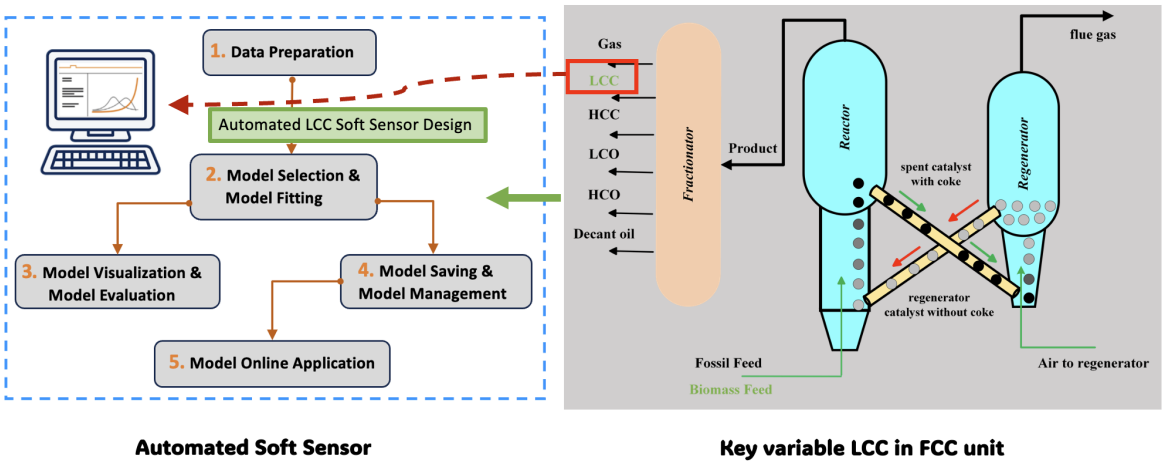

Figure 6: Framework of Automated Soft Sensor Design for LCC

To solve for $\boldsymbol{\beta}$ and $\mathbf{w}$, we use the method of Lagrange multipliers (Rockafellar, 1993). Taking the derivatives of Lagrangian L with respect to $\mathbf{w}$ and $\boldsymbol{\beta}$ and setting them to zero. After obtaining $\mathbf{w}$ and $\boldsymbol{\beta}$, the latent feature $\mathbf{u}$ can be computed. The soft sensor model is then constructed by regressing $\mathbf{y}$ onto $\mathbf{u}$:

$$\hat{\mathbf{y}} = b\mathbf{u} + \boldsymbol{\epsilon} \tag{14}$$

where b is the regression coefficient estimated using the least-squares method, and $\boldsymbol{\epsilon}$ is the residual error vector. To extract multiple latent features, the SCFE algorithm iteratively extracts a set of latent features that are orthogonal and have significant causal impacts on y.

3.5. Algorithm Selection Criteria

To ensure the appropriate selection of algorithms for different industrial scenarios, we provide guidelines for choosing between supervised, semi-supervised, and causal algorithms. Supervised learning methods are prioritized when sufficient labeled data are available. These methods are particularly suitable for processes with relatively stable operating conditions and well-defined input-output relationships.

Semi-supervised learning techniques are applicable in scenarios where labeled data is scarce but unlabeled data is abundant. This is common in industrial settings where manual labeling is expensive or time-consuming. Semi-supervised methods leverage the inherent structure of unlabeled data to enhance predictive performance, even with limited labeled data.

Causal machine learning algorithms are particularly useful for handling complex, high-dimensional industrial data with multicollinearity issues. These methods are beneficial when there are complex interdependencies among process variables, potential confounding factors, or when the plant experiences frequent changes in operating regimes. Causal algorithms can also improve model interpretability and robustness in non-stationary industrial environments.

To further tailor these algorithms for industrial applications, the Soft Sensor Manager incorporates several application-specific optimization strategies. For example, in supervised learning, domain-specific feature selection techniques are applied to enhance model interpretability and efficiency. In semi-supervised learning, the balance between labeled and unlabeled data is dynamically adjusted based on the process conditions, and informative pseudo-labels are generated by leveraging expert process knowledge. For causal machine learning, UCFE and SCFE algorithms integrate time-lag management and process-specific constraints to extract latent features that are both statistically robust and causally relevant.

The integration of supervised, semi-supervised, and causal machine learning techniques in the Soft Sensor Manager provides a comprehensive framework for automated soft sensor design. This combination leverages the strengths of each approach, enabling the automated soft sensor design to achieve higher accuracy, better generalization, and greater resilience to changing process conditions. By automating the selection and application of these diverse methods, the Soft Sensor Manager empowers engineers to develop more robust and adaptive soft sensors, ultimately contributing to improved monitoring and control in industrial processes.

4. Automated Soft Sensor Application

To validate the practicality and effectiveness of the proposed Soft Sensor Manager software, we applied it to the fluid catalytic cracking (FCC) unit at the Parkland Refinery in Canada (Cao et al., 2024b). Fluid catalytic cracking is one of the most important processes in modern refining, with the aim of converting heavy oil fractions into high-value lighter intermediates to produce gasoline and diesel. Light catalytic cracked oil (LCC) is a liquid petroleum product in an FCC unit, typically used as a blending feed for gasoline. The accurate measurement of this variable is crucial for optimizing production processes and improving economic benefits. In this project, we used the LCC yield as a key

Table 2
Performance of Different Machine Learning Algorithms in Automated Soft Sensor Tool

Category	Algorithm	RMSE (Train)	R^2 (Train)	RMSE (Test)	R^2 (Test)
Supervised Learning	RF	0.962 ± 0.021	0.942 ± 0.015	1.075 ± 0.032	0.905 ± 0.018
	PCR	*1.106 ± 0.035*	0.938 ± 0.018	*1.221 ± 0.041*	0.890 ± 0.014
	PLS	1.046 ± 0.028	0.927 ± 0.017	1.169 ± 0.036	0.897 ± 0.022
	SVM	0.980 ± 0.024	0.947 ± 0.013	1.089 ± 0.031	0.905 ± 0.020
	ANN	*0.941 ± 0.022*	*0.959 ± 0.014*	1.110 ± 0.033	0.889 ± 0.021
	RNN	0.978 ± 0.025	0.955 ± 0.017	1.092 ± 0.034	0.901 ± 0.015
	LSTM	0.952 ± 0.017	0.958 ± 0.015	1.086 ± 0.032	0.903 ± 0.011
Semi-Supervised Learning	SSGPR	0.950 ± 0.027	0.928 ± 0.015	1.175 ± 0.037	*0.884 ± 0.024*
	SSAER	0.972 ± 0.025	*0.924 ± 0.011*	1.173 ± 0.038	0.891 ± 0.019
	SSVAER	0.974 ± 0.026	0.931 ± 0.017	1.146 ± 0.035	0.895 ± 0.022
Causal Machine Learning	UCFE	0.971 ± 0.014	0.945 ± 0.007	1.084 ± 0.015	0.904 ± 0.009
	SCFE	0.950 ± 0.016	0.955 ± 0.012	*1.074 ± 0.013*	*0.906 ± 0.008*

quality variable. Figure 6 illustrates the framework for the automated LCC soft sensor design.

4.1. Data Preparation in LCC

For LCC soft sensor design, the training dataset spans from January 1, 2023, to March 15, 2024, and the testing dataset spans from March 16, 2024, to May 30, 2024, with a sampling frequency of 15 minutes. It should be noted that the data was split temporally rather than randomly. This approach preserves the temporal order of the data and reflects the practical scenario of deploying soft sensors for future predictions. We used the data import module to read raw data from the database into the software. This data includes process variables such as reaction temperature, pressure, feed rate, and catalyst activity. Next, we used the software's data preprocessing module to clean and align the raw data, automatically removing outliers and anomalies. The software used the mutual information-based feature selection method to automatically select important variables from hundreds of process variables (Jiang et al., 2019).

4.2. Model Selection and Evaluation in LCC

During the model selection phase, we evaluated three categories of machine learning algorithms. For each category, multiple algorithms were tested to provide a comprehensive comparison. We employed a rolling window cross-validation approach, which involves creating multiple training-testing splits while preserving temporal order. This method allows us to assess model stability across different time periods. Additionally, we have included uncertainty estimates for each performance metric to provide a more nuanced evaluation of model performance.

For supervised learning, we implemented random forest (RF), principal component regression (PCR), partial least squares (PLS), support vector machine (SVM), attention-based neural network (ANN) algorithms, recurrent neural network (RNN) and long short-term memory (LSTM) network. To address the customization of neural network architectures for different real-world processes, the Soft Sensor Manager provides a default setting with two hidden layers of 64 neurons each. This configuration has been found to be a good starting point for medium-scale industrial datasets. Beyond this default setting, the software supports auto-tuning through Bayesian optimization, which explores a range of network depths (2–9 hidden layers) and numbers of neurons per layer (32–256). The optimal architecture is determined through systematic cross-validation to balance model accuracy and computational efficiency. Among these, RF demonstrated the best performance on the test set with an RMSE of 1.075 and an R^2 of 0.905, indicating its strong generalizability. ANN showed the best performance on the training set, but exhibited some overfitting as evidenced by the slightly lower test set performance.

For semi-supervised learning, we employed SSGPR,SSAER and SSVAER. To simulate a real-world scenario where labeled data are scarce, we deliberately removed the 20% labels from the training data. This approach allowed us to evaluate the algorithms' ability to leverage unlabeled data, although it expectedly resulted in a slight decrease in overall performance compared to supervised methods using all available labels. Among the semi-supervised methods, SSVAER showed the best performance on the test set (RMSE: 1.146, R^2: 0.895). While this performance is lower than the best supervised learning results, it demonstrates the effectiveness of SSVAER in utilizing unlabeled data to improve model generalization under limited labeled data conditions.

In the causal machine learning category, we implemented the proposed UCFE and SCFE algorithms. SCFE outperformed UCFE and showed competitive results compared to the best supervised learning methods, achieving an RMSE of 1.074 and an R^2 of 0.906 on the test set. This indicates that incorporating causal relationships in feature extraction can lead to robust and generalizable models.

Table 2 provides a comprehensive summary of performance metrics with uncertainty estimates across various machine learning techniques. It should be noted that while some models showed excellent performance on the training set (e.g., NN with RMSE of 0.941 and R^2 of 0.959),

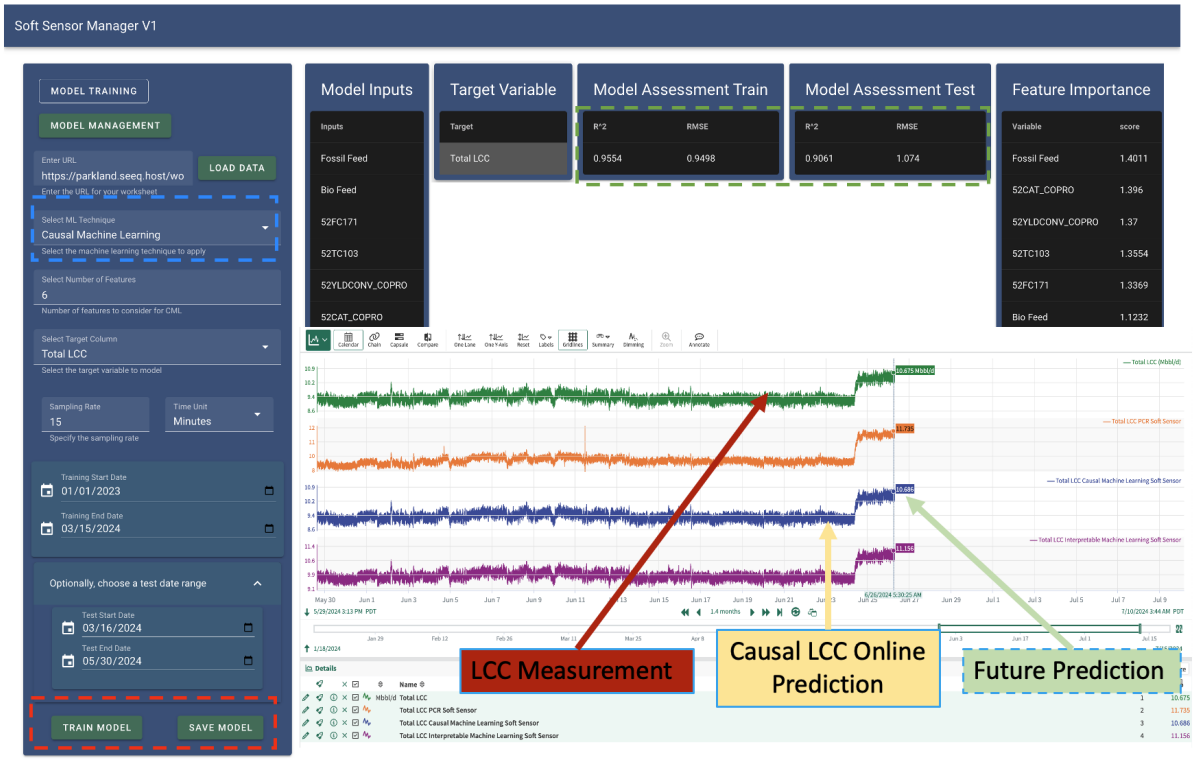

Figure 7: Application of Causal Machine Learning for Automated LCC Soft Sensor Design

their test set performance was relatively lower, indicating potential overfitting. In contrast, methods like RF and SCFE maintained more consistent performance between training and test sets, suggesting better generalization. These results highlight the importance of evaluating multiple algorithms. The Soft Sensor Manager allows engineers to easily compare these different approaches, enabling them to select the most suitable model for their specific application based on both predictive performance and other considerations such as interpretability and generalization capability.

4.3. Model Saving and Online Application

For the LCC automated soft sensor, the causal machine learning approach (SCFE) shows the best performance among various machine learning techniques tested. Once the model evaluation is complete, Soft Sensor Manager provides a straightforward method to save trained models. Users can save the optimized SCFE model by simply clicking the 'Save Model' button on the software interface. This function stores not only the model parameters, but also the training configuration, hyperparameters, and metadata associated with the model, ensuring that all necessary information is preserved for future use or retraining.

For online deployment, once the fitting results of SCFE saved, the model can be loaded directly into production environments via the software online deployment module. The deployment process is fully automated, requiring minimal user intervention, and the software monitors model performance in real time. Figure 7 presents the fitting and online application results of the SCFE method based on causal machine learning. In addition, the system allows for scheduled retraining, where models can be updated with new data, further enhancing their robustness and long-term reliability in dynamic industrial environments.

4.4. Discussion

The primary objective of this study is not to determine the superiority of any single algorithm, as the efficacy of different methods can vary significantly across industrial processes, key variables, and datasets. Instead, Soft Sensor Manager is designed as a platform that allows engineers to rapidly evaluate and implement a wide range of machine learning techniques without requiring extensive prior knowledge or preparation.

This approach offers several key advantages. The software accommodates various industrial scenarios, allowing users to select the most appropriate model for their specific application. By automating much of the modeling process, it significantly reduces the time and effort required to develop and deploy soft sensors. In addition, the user-friendly interface and automated workflows make advanced machine learning techniques accessible to engineers who may not have specialized data science expertise. The ability to easily test and compare multiple models also encourages ongoing optimization of soft sensor performance, fostering continuous improvement.

From a scientific and technical perspective, one of the significant challenges addressed by this paper is the integration of diverse machine learning algorithms. This platform integrates supervised, semi-supervised, and causal learning

methods, providing engineers with a comprehensive toolkit to address diverse industrial challenges. By developing unified modules for data preprocessing, model training, and evaluation that cater to these varied algorithms, the Soft Sensor Manager overcomes compatibility issues and simplifies the user experience.

Another technical difficulty tackled is the automation of hyperparameter tuning and model selection. Selecting optimal hyperparameters is critical for model performance but is often a complex and time-consuming task requiring expert knowledge. The software incorporates automated hyperparameter optimization techniques that streamline this process, enabling non-expert users to achieve high-quality models without extensive experimentation.

Moreover, the inclusion of causal machine learning methods addresses the challenge of model interpretability and robustness in changing process conditions. Traditional soft sensors may perform well under static conditions but can degrade when process dynamics shift. By leveraging causal relationships, the Soft Sensor Manager enhances the stability and generalization capabilities of the models, ensuring more reliable performance in dynamic industrial environments. Specifically, the UCFE and SCFE algorithm introduced in this study incorporate causal relationships and temporal dynamics, leading to more robust and interpretable models.

While the results demonstrate the potential of Soft Sensor Manager, it is crucial to recognize its current limitations and areas for improvement. The performance of the software may fluctuate across different industrial processes, data characteristics, and operational complexities. Factors such as data quality, quantity, and intrinsic complexity of the monitored processes can significantly influence the tool's effectiveness. Additionally, the software's adaptability to highly specialized or rapidly evolving industrial environments may require further refinement.

Future research directions should focus on enhancing the robustness and interpretability of the Soft Sensor Manager. This could involve integrating advanced techniques such as transfer learning (Li et al., 2023) and meta-learning (Sun et al., 2019) to address challenges related to data scarcity and improve the tool's performance in new or data-limited scenarios. Furthermore, exploring the incorporation of domain-specific knowledge and hybrid modeling approaches could enhance the software's applicability across various industrial sectors (Jia et al., 2008; Subramanian et al., 2022; Peng et al., 2022).

5. Conclusion

In this study, we developed an innovative automated soft sensor design tool to address the challenges associated with real-time monitoring and control in complex industrial processes. The Soft Sensor Manager simplifies data preprocessing, feature engineering, model selection, model evaluation, and online deployment, making advanced soft sensor technology more accessible. The results from the industrial light catalytic cracked oil yield demonstrate that the software not only reduces the complexity and time required for soft sensor development, but also enhances the accuracy and reliability of the developed models. The proposed automated soft sensor design tool facilitates the integration of advanced machine learning techniques into industrial process monitoring systems, representing a significant advancement in soft sensor technology. Future work will focus on improving user experience, and extending its applications to other industrial processes, thus supporting the intelligent and digital transformation of the industry.

References

ABB, . Abb 800xa dcs distributed control system. https://new.abb.com/control-systems/system-800xa/800xa-dcs. Accessed: 2024-06-27.

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A next-generation hyperparameter optimization framework, in: The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2623–2631.

Byrski, W., Drapała, M., Byrski, J., Noack, M., Reger, J., 2024. Comparison of lqr with mpc in the adaptive stabilization of a glass conditioning process using soft-sensors for parameter identification and state observation. Control Engineering Practice 146, 105884.

Cao, L., 2024. Interpretable and stable soft sensor modeling for industrial applications. Ph.D. thesis. University of British Columbia. URL: https://open.library.ubc.ca/collections/ubctheses/24/items/1.0440692, doi:http://dx.doi.org/10.14288/1.0440692.

Cao, L., Ji, X., Cao, Y., Luo, Y., Wang, Y., Siang, L.C., Li, J., Gopaluni, R.B., 2024a. Interpretable industrial soft sensor design based on informer and shap. IFAC-PapersOnLine 58, 73–78.

Cao, L., Su, J., Saddler, J., Cao, Y., Wang, Y., Lee, G., Siang, L.C., Pinchuk, R., Li, J., Gopaluni, R.B., 2024b. Real-time tracking of renewable carbon content with ai-aided approaches during co-processing of biofeedstocks. Applied Energy 360, 122815.

Cao, L., Su, J., Wang, Y., Cao, Y., Siang, L.C., Li, J., Saddler, J.N., Gopaluni, B., 2022. Causal discovery based on observational data and process knowledge in industrial processes. Industrial & Engineering Chemistry Research 61, 14272–14283. doi:10.1021/acs.iecr.2c01326.

Cao, L., Yu, F., Yang, F., Cao, Y., Gopaluni, R.B., 2020. Data-driven dynamic inferential sensors based on causality analysis. Control Engineering Practice 104, 104626.

Chan, G., Claassen, T., Hoos, H., Heskes, T., Baratchi, M., 2024. Autocd: Automated machine learning for causal discovery algorithms .

Chang, S., Zhao, C., Li, K., 2021. Consistent-contrastive network with temporality-awareness for robust-to-anomaly industrial soft sensor. IEEE Transactions on Instrumentation and Measurement 71, 1–12.

Chen, K., Castillo, I., Chiang, L.H., Yu, J., 2015. Soft sensor model maintenance: A case study in industrial processes. IFAC-PapersOnLine 48, 427–432. 9th IFAC Symposium on Advanced Control of Chemical Processes ADCHEM 2015.

Cheng, Q., Chunhong, Z., Qianglin, L., 2023. Development and application of random forest regression soft sensor model for treating domestic wastewater in a sequencing batch reactor. Scientific Reports 13, 9149.

Chickering, D.M., 2002. Optimal structure identification with greedy search. Journal of Machine Learning Research 3, 507–554.

D. Asir Antony Gnana Singh, S. Appavu Alias Balamurugan, E.J.L., 2016. Literature review on feature selection methods for high-dimensional data. International Journal of Computer Applications 136, 9–17.

Esche, E., Talis, T., Weigert, J., Rihm, G.B., You, B., Hoffmann, C., Repke, J.U., 2022. Semi-supervised learning for data-driven soft-sensing of biological and chemical processes. Chemical Engineering Science 251, 117459.

Fan, J., Han, F., Liu, H., 2014. Challenges of Big Data analysis. National Science Review 1, 293–314.

Gopaluni, R.B., Tulsyan, A., Chachuat, B., et al., 2020. Modern machine learning tools for monitoring and control of industrial processes: A

survey. IFAC-PapersOnLine 53, 218–229.

Honeywell, . Industrial automation. https://automation.honeywell.com/us/en. Accessed: 2024-06-27.

Hutter, F., Kotthoff, L., Vanschoren, J., 2019. Automated machine learning: methods, systems, challenges. Springer Nature.

Jia, W., you Chai, T., Yu, W., 2008. A novel hybrid neural network for modeling rare-earth extraction process. IFAC Proceedings Volumes 41, 11427–11432.

Jiang, B., Luo, Y., Lu, Q., 2019. Maximized mutual information analysis based on stochastic representation for process monitoring. IEEE Transactions on Industrial Informatics 15, 1579–1587. doi:10.1109/TII.2018.2853702.

Jiang, Y., Yin, S., Dong, J., Kaynak, O., 2020. A review on soft sensors for monitoring, control, and optimization of industrial processes. IEEE Sensors Journal 21, 12868–12881.

Kadlec, P., Gabrys, B., 2009. Soft sensors: where are we and what are the current and future challenges? IFAC Proceedings Volumes 42, 572–577. 2nd IFAC Conference on Intelligent Control Systems and Signal Processing.

Kadlec, P., Gabrys, B., Strandt, S., 2009. Data-driven soft sensors in the process industry. Computers & Chemical Engineering 33, 795–814.

Karmaker, S.K., Hassan, M.M., Smith, M.J., Xu, L., Zhai, C., Veeramachaneni, K., 2021. Automl to date and beyond: Challenges and opportunities. ACM Computing Surveys (CSUR) 54, 1–36.

Lawrence, N.P., Damarla, S.K., Kim, J.W., Tulsyan, A., Amjad, F., Wang, K., Chachuat, B., Lee, J.M., Huang, B., Bhushan Gopaluni, R., 2024. Machine learning for industrial sensing and control: A survey and practical perspective. Control Engineering Practice 145, 105841.

Li, D., Liu, Y., Huang, D., Lui, C.F., Xie, M., 2023. Development of an adversarial transfer learning based soft sensor in industrial systems. IEEE Transactions on Instrumentation and Measurement .

Liu, J., 2014. Developing a soft sensor based on sparse partial least squares with variable selection. Journal of Process Control 24, 1046–1056.

Lou, Z., Wang, Y., Si, Y., Lu, S., 2022. A novel multivariate statistical process monitoring algorithm: Orthonormal subspace analysis. Automatica 138, 110148.

Lu, B., Chiang, L., 2018. Semi-supervised online soft sensor maintenance experiences in the chemical industry. Journal of Process Control 67, 23–34. Big Data: Data Science for Process Control and Operations.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Monirul Kabir, M., Monirul Islam, M., Murase, K., 2010. A new wrapper feature selection approach using neural network. Neurocomput. 73, 3273–3283.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems 32.

Pearl, J., 2014. Probabilistic reasoning in intelligent systems: networks of plausible inference. Elsevier.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: Machine learning in python. the Journal of machine Learning research 12, 2825–2830.

Peng, W., Yao, W., Zhou, W., Zhang, X., Yao, W., 2022. Robust regression with highly corrupted data via physics informed neural networks. ArXiv abs/2210.10646.

Qin, S.J., 2014. Process data analytics in the era of big data. AIChE Journal 60, 3092–3100.

Real, E., Liang, C., So, D., Le, Q., 2020. Automl-zero: Evolving machine learning algorithms from scratch, in: International conference on machine learning, PMLR. pp. 8007–8019.

Rockafellar, R.T., 1993. Lagrange multipliers and optimality. SIAM review 35, 183–238.

Russell E.L., Chiang L.H., B.R., 2000. Data-driven Methods for Fault Detection and Diagnosis in Chemical Processes. Springer, London.

Salehin, I., Islam, M.S., Saha, P., Noman, S., Tuni, A., Hasan, M.M., Baten, M.A., 2024. Automl: A systematic review on automated machine learning with neural architecture search. Journal of Information and Intelligence 2, 52–81.

Schaeffer, J., Braatz, R.D., 2022. Latent variable method demonstrator — software for understanding multivariate data analytics algorithms. Computers Chemical Engineering 167, 108014.

Shang, C., Gao, X., Yang, F., Huang, D., 2014. Novel bayesian framework for dynamic soft sensor based on support vector machine with finite impulse response. IEEE Transactions on Control Systems Technology 22, 1550–1557.

Siemens, . Industry software. https://www.siemens.com/global/en/products/automation/industry-software.html. Accessed: 2024-06-27.

Spirtes, P., Glymour, C., Scheines, R., 2000. Causation, Prediction, and Search. MIT Press.

Subramanian, S., Kirby, R.M., Mahoney, M.W., Gholami, A., 2022. Adaptive self-supervision algorithms for physics-informed neural networks, in: European Conference on Artificial Intelligence.

Sun, Q., Liu, Y., Chua, T.S., Schiele, B., 2019. Meta-transfer learning for few-shot learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58, 267–288.

Xu, R., Cui, P., Shen, Z., Zhang, X., Zhang, T., 2021. Why stable learning works? A theory of covariate shift generalization URL: https://arxiv.org/abs/2111.02355.

Yu, C., Yao, W., 2017. Robust linear regression: A review and comparison. Communications in Statistics - Simulation and Computation 46, 6261–6282. doi:10.1080/03610918.2016.1202271.

Yu, F., Xiong, Q., Cao, L., Yang, F., 2022. Stable soft sensor modeling based on causality analysis. Control Engineering Practice 122, 105109.

Yuan, X., Huang, B., Ge, Z., Song, Z., 2016. Double locally weighted principal component regression for soft sensor with sample selection under supervised latent structure. Chemometrics and Intelligent Laboratory Systems 153, 116–125.

Yuan, X., Huang, B., Wang, Y., Yang, C., Gui, W., 2018. Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted sae. IEEE Transactions on Industrial Informatics 14, 3235–3243.

Yuan, X., Ou, C., Wang, Y., Yang, C., Gui, W., 2020. A novel semi-supervised pre-training strategy for deep networks and its application for quality variable prediction in industrial processes. Chemical Engineering Science 217, 115509.

Zhuang, Y., Zhou, Z., Alakent, B., Mercangöz, M., 2023. Semi-supervised variational autoencoders for regression: application to soft sensors, in: 2023 IEEE 21st International Conference on Industrial Informatics (INDIN), IEEE. pp. 1–8.