Comprehensive Analysis on Machine Learning Approaches for Interpretable and Stable Soft Sensors

Liang Cao[a], Jingyi Wang[b], Jianping Su[c], Yi Luo[d], Yankai Cao[b], Richard D. Braatz[a*] and Bhushan Gopaluni[b*]

***Abstract*—Soft sensors play a vital role in industrial process monitoring and control by estimating difficult-to-measure quality variables. While significant progress has been made in improving the accuracy of soft sensor models, challenges remain in ensuring their interpretability and stability in dynamic industrial environments. From a measurement science perspective, ensuring transparency and reliable performance under varying process conditions is becoming increasingly critical, particularly in high-stakes industrial applications. This paper provides a comprehensive review of methodologies to enhance the interpretability and stability of soft sensor models. To address interpretability, we analyze various interpretable machine learning techniques applicable to soft sensors and discuss open-source projects that facilitate the implementation of these techniques. For improving stability, we emphasize the role of causal machine learning, detailing methods for causal discovery in industrial processes and highlighting relevant open-source tools. By highlighting current limitations and identifying areas for improvement, we aim to provide valuable insights and practical tools for researchers and practitioners. These insights will guide the development of more transparent and reliable soft sensors, ultimately enhancing industrial process monitoring and control.**

***Index Terms*—Soft Sensor, Interpretable Machine Learning, Causal Machine Learning, Trustworthy Model, Industrial Process Monitoring**

I. Introduction

A. Background

With stringent requirements for product quality and cost, the complexity and degree of automation of industrial processes is continuously increasing [1]–[4]. As the scale of plants grows, it is vital to monitor critical variables that are closely related to process safety and economic benefits. These critical variables are called quality variables. However, some quality variables are difficult or costly to measure in real time, posing a significant challenge for real-time process monitoring. To overcome this challenge, soft sensor technology has been introduced. The basic idea of a soft sensor is to select easily measurable process variables to construct a mathematical relationship that can estimate the values of quality variables.

By enabling real-time estimation of quality variables, soft sensors play a vital role in industrial applications, essential for monitoring, control, and optimization of manufacturing processes. In sectors like oil and gas, pharmaceuticals, and chemicals, they enable real-time monitoring of key quality and safety parameters. For example, in the oil industry, soft sensors estimate the crude oil composition to maintain product quality and process stability [5]. In pharmaceutical manufacturing, they monitor critical quality attributes to enhance the consistency and effectiveness of the drug formulation [6].

The integration of soft sensors with advanced data analytics and the industrial internet of things further amplifies their importance. Modern industrial operations generate vast amounts of data that, when properly analyzed, can provide actionable insights for process improvement. Soft sensors leverage this data to improve process control, support sustainable manufacturing practices, and drive operational efficiencies. As industries continue to evolve toward more automated and data-driven methodologies, the reliance on soft sensors is expected to grow. This trend underscores the indispensable role of soft sensors in both current and future industrial landscapes.

B. Motivation

Soft sensors have become indispensable in modern industrial measurement and control systems. However, many high-performance soft sensors operate as black boxes, making it difficult for operators to understand their predictions. Furthermore, these models may not perform consistently under the varying conditions typical in dynamic industrial environments (as shown in Figure 1). Addressing these challenges is crucial to ensure that soft sensor models are not only accurate but also trustworthy in practical applications [7]–[12].

In the design of soft sensors, *interpretability* refers to the transparency and ease with which a human can understand the decision-making process of the model [8], [13]. Interpretability not only fosters confidence among operators and engineers, but also guarantees adherence to stringent regulatory requirements by enabling transparent and justifiable decision-making processes. *Stability* in soft sensor models refers to the ability to maintain consistent performance under varying operating conditions and data fluctuations [14], [15]. Industrial environments are dynamic, yet many models assume a consistent data distribution between training and testing, which is often

This work was supported by the National Natural Science Foundation of China (Grant No. 62403229). Corresponding author: (Richard D. Braatz, Bhushan Gopaluni.) Liang Cao and Richard D. Braatz are with the Department of Chemical Engineering, Massachusetts Institute of Technology, Boston, 02139, United States (e-mail: liangcao@mit.edu, braatz@mit.edu). Jingyi Wang, Yankai Cao and Bhushan Gopaluni are with the Department of Chemical and Biological Engineering, University of British Columbia, Vancouver, BC, V6T 1Z3, Canada (e-mail: wjy1129@student.ubc.ca, yankai.cao@ubc.ca, bhushan.gopaluni@ubc.ca). Jianping Su is with the China University of Petroleum, Beijing, 102200, China (e-mail: jianping.su@cup.edu.cn). Yi Luo is with the Research Institute of Big Data in Mines, Chinese Institute of Coal Science, 10013, Beijing, China (e-mail: luoyi_ccteg@163.com).

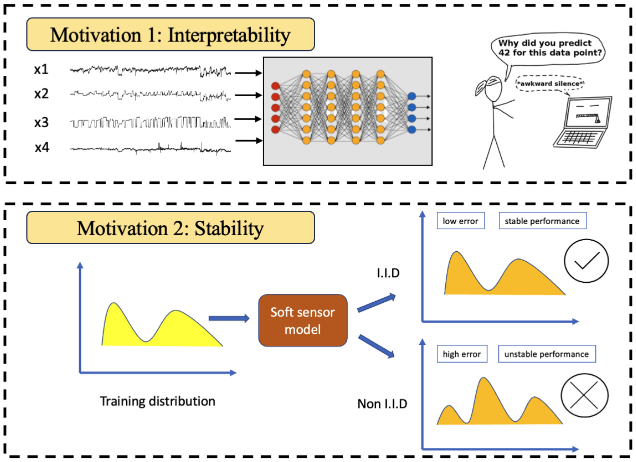

Fig. 1: Motivation for interpretability and stablility in soft sensors

violated in practice. This leads to performance degradation and unreliable predictions under distribution shifts.

Lack of interpretability and stability in soft sensor models may lead to suboptimal process control, wasted resources, and reduced product quality. Addressing these challenges requires methodologies from interpretable machine learning and causal machine learning. By focusing on these aspects, we can advance soft sensor technologies to meet the complex demands of modern industrial processes, ultimately contributing to safer, more efficient, and sustainable operations.

C. Objectives and Structure of the Review

In this paper, our objective is to provide a comprehensive review of methodologies for enhancing the interpretability and stability of soft sensor models in industrial processes. By systematically analyzing current approaches and identifying their limitations, we seek to offer insights that can guide the development of soft sensors that are both high-performing and suitable for deployment in real-world settings.

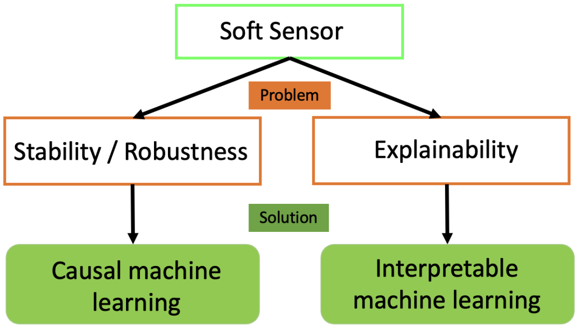

Fig. 2: Challenges and Solutions in Soft Sensor Modeling

Specifically, we focus on the use of interpretable machine learning and causal machine learning methods in soft sensor development, as outlined in Figure 2. We explore how these techniques can be employed to create trustworthy models by reviewing the existing literature and presenting emerging methods. We aim to inspire future research and innovation in this field and believe that addressing these aspects will significantly contribute to the advancement of soft sensor technologies, enabling them to meet the complex demands of modern industrial processes.

The structure of this paper is organized as follows. In Section 2, we provide an overview of soft sensor modeling approaches. Section 3 discusses the challenges of implementing soft sensors, focusing on issues of interpretability, stability, and other practical considerations. In Section 4, we explore methods for enhancing interpretability in soft sensors through interpretable machine learning techniques and relevant open-source projects. Section 5 addresses strategies to ensure the stability of soft sensor models, highlighting the application of causal machine learning and key open-source tools. Section 6 examines future directions and emerging trends in soft sensor technologies. Finally, Section 7 concludes the review, summarizing the importance of advancing interpretability and stability in soft sensor development.

II. Soft Sensor Modeling Approaches

Soft sensor modeling approaches have evolved over time, reflecting advances in both theoretical and practical applications. In the following sections, we first explore the definition and historical evolution of soft sensors, providing a comprehensive understanding of how they have developed. We then delve into the characteristics of various modeling techniques, which can generally be categorized into three main types: first principles models, data-driven models, and hybrid models. This exploration offers insights into how these different approaches are applied in industrial settings and their respective strengths and limitations.

A. Definition and Historical Evolution

The mathematical definition of a soft sensor can be given as follows:

$$y = f\left(d, X\right) \tag{1}$$

where d, X, and y are unknown noise, easy-to-measure variables, and quality variables, respectively. As shown in Equation 1 and Figure 3, developing a soft sensor involves two fundamental components: selecting appropriate input variables X that are easily measured and designing a regression model f that accurately captures the relationship between the inputs and the target variable.

While X represents variables that are convenient to measure in real-time, not all such variables necessarily have a significant or causal impact on the target variable y. Therefore, practitioners must also carefully assess which subsets of these easily measurable variables are truly relevant to the system under study. This assessment often requires rigorous feature selection, causal analysis, and domain expertise to ensure that only those variables most influential on y are included in the soft sensor model.

The early development of soft sensors for process control and monitoring began in the 1970s, with a notable contribution

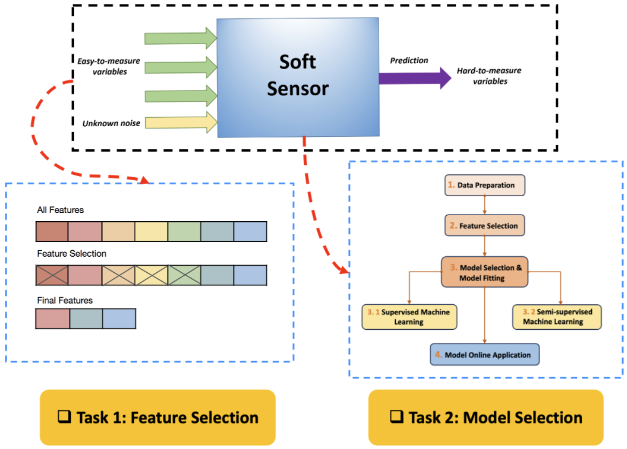
Easy-to-measure variables
Unknown noise
Soft Sensor
Prediction
Hard-to-measure variables
All Features
Feature Selection
Final Features
1. Data Preparation
2. Feature Selection
3. Model Selection & Model Fitting
3. 1 Supervised Machine Learning
3. 2 Semi-supervised Machine Learning
4. Model Online Application
❑ Task 1: Feature Selection
❑ Task 2: Model Selection

Fig. 3: Two major tasks for soft sensor

by Weber and Brosilow in 1972, who systematically studied process control using additional variables, commonly referred to as 'secondary' or 'proxy' variables, to estimate hard-to-measure process parameters [16]. This indirect estimation approach became foundational for the evolution of soft sensors. The term "soft sensors" became formally defined and widely recognized in the mid-1980s. From that point onward, the technology experienced rapid development, attracting considerable attention from both academic and industrial communities worldwide. A key milestone in this evolution was McAvoy, T.J.'s 1992 publication in *Automatica*, solidifying soft sensors as a significant area of research in process control [17].

Alongside conceptual advances, technological innovations have expanded the range of soft sensor applications. According to Kadlec et al. [18], soft sensor applications can be broadly classified into three primary classes. The first is real-time or near-real-time predictions, which provide instantaneous estimates of process variables that are essential for immediate decision making and control actions. The second is dynamic process monitoring and early fault detection, aimed at enhancing the ability to monitor processes over time and detect faults at an early stage to prevent adverse outcomes. The third class focuses on quality assurance and sensor data calibration, ensuring product quality by calibrating sensors and validating data to maintain accuracy and compliance with standards. Over the years, soft sensor technology has continued to evolve, driven by advances in computational techniques, data acquisition systems, and industrial automation.

B. Classification of Soft Sensors

Soft sensor modeling approaches can be broadly classified into three main categories: first principles models, data-driven models, and hybrid models. Each of these approaches offers distinct advantages and challenges, as summarized in Table I. The selection of the most appropriate method depends on factors such as process complexity, data availability and quality, and domain expertise. Careful consideration of these factors ensures the development of effective and reliable soft sensor models tailored to specific industrial needs.

1) First Principles Models: First principles models are grounded in fundamental physical, chemical, or biological laws that govern the behavior of the system. These models offer a comprehensive understanding of the dynamics of the system by incorporating equations derived from these fundamental principles [19], [20]. For example, in a chemical process, the model would include equations representing reactions, mass transfer, and energy balance. These models are highly reliable when the underlying principles are well understood and can be accurately represented mathematically.

However, the development of first principles models can be challenging due to the inherent complexity of the natural processes they represent. Such models often involve higher-order partial differential equations or systems of equations, which are more difficult to obtain. Additionally, these models often involve many parameters that need accurate estimation to ensure validity and effectiveness. The complexity increases when dealing with non-linear and multivariate systems, which requires sophisticated computational techniques and algorithms to solve and optimize the governing equations.

2) Data-Driven Models: Data-driven soft sensor models learn patterns directly from historical and real-time process data [21]–[27], offering flexibility but requiring careful evaluation of their strengths and weaknesses for industrial applications.

Traditional machine learning techniques form the foundation of data-driven soft sensors. Linear methods like Partial Least Squares (PLS) [26] and ridge regression (RR) offer excellent interpretability and computational efficiency but may struggle with highly nonlinear relationships. Ensemble learning and meta-learning methods further integrate the advantages of multiple models and algorithms to achieve performance beyond that of a single model [28]–[36].

Deep learning approaches, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), autoencoders (AE) and generative adversarial networks (GAN), excel at capturing complex dependencies and nonlinear relationships [37]–[39]. While these models can achieve superior accuracy, they present challenges in interpretability and require substantial training data. Their "black-box" nature often necessitates additional techniques for explanation, potentially complicating deployment in regulated industries.

Reinforcement learning methods [40]–[43] offer unique advantages in adaptive control and optimization but face challenges in industrial deployment due to exploration-exploitation trade-offs and safety concerns. These methods show promise in scenarios requiring continuous adaptation but may be less suitable for critical processes where predictable behavior is essential.

To facilitate systematic comparison, we present a detailed analysis of various data-driven approaches in Table II. Each method category presents distinct trade-offs between accuracy, interpretability, data requirements, and computational complexity. The selection of an appropriate method should consider not only predictive performance but also practical constraints such as data availability, computational resources, and regulatory requirements.

TABLE I: Comparison of Soft Sensor Modeling Methods

Modeling Approach	**Specific Methods**	**Advantages**	**Disadvantages**
First Principles Models	• Differential Equations (e.g., mass and energy balances) • Thermodynamic Models • Kinetic Models • Empirical Correlations	• High accuracy when underlying physical laws are accurately represented • Provides insights into the fundamental mechanisms of the system	• Requires detailed and precise knowledge of the system's physical properties • Time-consuming and complex model development process • High sensitivity to parameter uncertainties and measurement errors
Data-Driven Models	• Traditional Machine Learning • Deep Learning • Reinforcement Learning	• Ability to model complex and nonlinear relationships without explicit physical equations • Capable of handling large-scale datasets with high dimensionality • Can improve performance with more data	• May lack interpretability and may give unrealistic predictions • Requires large amounts of high-quality data for effective training and prone to overfitting • Limited ability when beyond the range of the training data
Hybrid Models	• PINN • Grey-Box Models • Neuro-Fuzzy Systems • Mechanistic-Statistical Models	• Combines the accuracy and flexibility of data-driven models with the interpretability of first principles • Enhances model reliability by incorporating physical laws and constraints • Improves predictive performance in scenarios with limited or noisy data	• Increased complexity in model development and implementation • Requires expertise in both domain knowledge and data-driven techniques • May still inherit limitations from both constituent modeling approaches

TABLE II: Critical Comparison of Different Machine Learning Models for Data-Driven Soft Sensor

Method Category	**Representative Methods**	**Strengths**	**Limitations**	**Best Use Cases**
Linear Methods	• PLS • PCA • Ridge Regression	• High interpretability • Computationally efficient	• Limited nonlinear capability • Sensitive to outliers • May oversimplify relationships	• Well-understood processes • Linear relationships • Regulatory requirements
Tree-based Methods	• Random Forest • XGBoost • LightGBM	• Moderate interpretability • Handle nonlinearity well • Built-in feature importance	• May overfit on small datasets • Require frequent retraining • Memory intensive	• Medium-scale processes • Nonlinear relationships • Feature selection needs
Deep Learning	• CNN • RNN • LSTM	• Superior nonlinear modeling • Handle temporal dependencies • Automatic feature extraction	• Limited interpretability • Large data requirements • Complex training process	• Complex processes • Large data availability • Time-series prediction
Reinforcement Learning	• DQN • DDPG • SAC	• Adaptive optimization • Online learning capability • Handle dynamic environments	• Safety concerns during training • Complex implementation • Unpredictable exploration	• Adaptive control • Process optimization • Non-critical applications

3) Hybrid Models: Hybrid models combine the strengths of first principles and data-driven models, balancing physical insights with the capability to model complex relationships. By embedding physical laws into the model architecture, hybrid models incorporate structured knowledge and enforce constraints, while simultaneously leveraging data-driven components to capture complex nonlinear relationships that are challenging to express through analytical equations [44]–[46]. Among the prominent hybrid modeling techniques are hybrid neural networks [44], grey-box models [47], neuro-fuzzy systems [48], mechanistic-statistical models [49] and physics-informed neural networks (PINNs) [45], [46].

PINNs incorporate physical laws, generally expressed through partial differential equations (PDEs), directly into the neural network's training. Specifically, a PINN trains a neural network $u_\theta(x,t)$ not only to fit the observed data, but also to adhere to the governing physical equations. This is achieved through a composite loss function $\mathcal{L}(\theta)$, which comprises both data-based and physics-based components:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{data}}(\theta) + \mathcal{L}_{\text{physics}}(\theta) \tag{2}$$

$$\mathcal{L}_{\text{data}}(\theta) = \frac{1}{N_d}\sum_{i=1}^{N_d} \left| u_\theta(x_i,t_i) - u_i^{\text{obs}} \right|^2 \tag{3}$$

$$\mathcal{L}_{\text{physics}}(\theta) = \frac{1}{N_r}\sum_{j=1}^{N_r} \left| \mathcal{N}[u_\theta](x_j^r,t_j^r) \right|^2 \tag{4}$$

Here, $\mathcal{N}[u_\theta]$ denotes the differential operator defined by the relevant PDEs, N_d represents the number of data points, N_r is the number of residual points where the PDE residual is evaluated, and u_i^{obs} are the observed data values at spatial and temporal coordinates (x_i,t_i). By minimizing $\mathcal{L}(\theta)$, the PINN ensures that the neural network model aligns with empirical observations while satisfying the underlying physical laws that govern the system.

Hybrid models provide accurate and physically consistent predictions, particularly in scenarios where data may be sparse or noisy. This alignment with physical principles also improves model interpretability and generalizability. Their ability to integrate physical laws with data-driven insights places them at the forefront of soft sensor technology.

III. Challenges in Soft Sensor Implementation

In the context of industrial measurement and instrumentation, soft sensors represent a crucial advancement in process variable estimation and monitoring. The implementation of soft sensors in industrial measurement systems presents unique challenges that extend beyond traditional machine learning considerations. When evaluating soft sensor performance, traditional machine learning metrics must be interpreted within this measurement science context. In real-world industrial environments, these measurement science considerations intersect with practical challenges of deploying advanced sensing solutions. Models must not only be transparent and reliable under varying conditions but also satisfy rigorous metrological requirements. Many advanced models, particularly those based on data-driven methods, struggle to meet these combined demands of measurement science rigor and practical deployability. This section explores these challenges, focusing on how interpretability, stability, and other factors affect the performance of soft sensors.

A. Interpretability

From a measurement science perspective, a highly interpretable soft sensor must provide clear traceability in its measurement chain, allowing users to understand how the model transforms raw sensor inputs into calibrated measurements. This includes quantifying measurement uncertainty at each step of the indirect measurement process and understanding how different input variables and their associated uncertainties contribute to the final measurement result. Beyond just understanding model predictions, interpretability in the measurement context requires establishing clear links to reference standards and documenting the complete measurement equation that relates input quantities to the measurand.

A highly interpretable soft sensor allows users to understand how the model makes predictions based on input data, which input variables significantly impact the prediction results, and how these variables interact with each other. Providing a mathematical definition of interpretability in a soft sensor is challenging, as interpretability is a subjective and context-dependent concept. However, we can attempt to propose a more general definition linking interpretability to the characteristics and behavior of a model. Suppose that we have a model $f : X \to Y$, the interpretability of model f can be defined as:

$$\begin{aligned} \text{Interpretability}(f) &= \alpha \cdot T(f) + \beta \cdot U(f) \\ T &: f \to [0,1], \\ U &: f \to [0,1], \alpha, \beta \in [0,1] \end{aligned} \tag{5}$$

where $T(f)$ and $U(f)$ are the transparency and understandability measure functions of the model f, respectively. α and β are weight coefficients that reflect the relative importance of transparency and understandability in a specific application scenario.

In industrial applications, models with high interpretability are usually more popular because they can help operators and engineers better understand the process, diagnose problems, and make decisions. When users clearly understand the working mechanism of the model, they are more likely to trust its predictions and take actions accordingly. However, achieving high interpretability often requires simplifying the model, which can lead to reduced accuracy and performance. In addition, complex data-driven models, such as deep neural networks, inherently lack transparency, making it difficult to elucidate the underlying decision-making process. Balancing interpretability with model complexity and performance remains a significant challenge in the implementation of soft sensors.

B. Stability

A stable soft sensor can provide consistent and reliable predictions when exposed to minor perturbations or changes in the data. Similar to interpretability, we can attempt to propose a more general definition for stability. To define stability, we integrate the concept of uniform stability from statistical learning theory. We first define the average error (AE) and stability error (SE) to quantify the model's stability:

$$\mathrm{AE} = \frac{1}{\mathcal{E}} \sum_{i \in \mathcal{E}} \mathrm{RMSE}^{i} \tag{6}$$

$$\mathrm{SE} = \sqrt{\frac{1}{\mathcal{E}-1} \sum_{i \in \mathcal{E}} (\mathrm{RMSE}^{i} - \mathrm{AE})^{2}} \tag{7}$$

where $\mathcal{E}$ denotes the number of operating conditions, and RMSE^{i} represents the root mean square error under operating conditions i. Uniform stability measures the sensitivity of a learning algorithm to changes in its training data. A learning algorithm A is said to be uniformly β-stable if for any two training sets S and S' differing by a single sample, and for all inputs $z \in Z$:

$$|L_z(h_S) - L_z(h_{S'})| \leq \beta \tag{8}$$

where h_S and $h_{S'}$ are models trained on S and S' respectively, $L_z(h)$ is the loss of model h on input z. By incorporating uniform stability, we link the model's sensitivity to training data perturbations with its performance consistency across different operating conditions. Specifically, a uniformly β-stable soft sensor ensures that:

$$\mathrm{SE} \leq \beta$$

Low SE indicates that the model maintains low and stable prediction errors in different operating conditions. By defining AE and SE and introducing the stability coefficient β, this framework provides a practical assessment of stability for soft sensor models. Stability can be explored in depth from different perspectives, which mainly include algorithmic stability [50], feature selection stability [51], and hyperparameter tuning [52], [53]. Different model architectures, different subset of features, along with varying hyperparameter settings like learning rate and regularization parameters, can significantly influence model stability.

C. Trade-offs Between Interpretability and Stability

In the development of soft sensor models, practitioners often encounter inherent trade-offs between interpretability and stability. Simple, highly interpretable models may lack the complexity required to maintain stable performance under varying conditions, while more complex, stable models often sacrifice transparency. This tension becomes particularly evident in industrial settings where both understanding model decisions and ensuring reliable performance are crucial.

Consider linear models or simple decision trees, which offer clear insights into feature relationships and decision boundaries. While these models excel in interpretability, allowing operators to easily understand how input variables influence predictions, they may struggle to maintain consistent performance when faced with significant process variations or distribution shifts. Conversely, ensemble methods or deep neural networks can achieve remarkable stability across different operating conditions through their complex architectures and robust training procedures, but their intricate decision-making processes often appear as ”black boxes” to users. The balance between these competing objectives depends on several factors specific to the industrial application:

- *Regulatory Requirements*: In highly regulated industries like pharmaceuticals or chemical manufacturing, interpretability may take precedence due to compliance needs, even if it means more frequent model recalibration.
- *Process Dynamics*: Processes with frequent operational changes or significant variability may necessitate prioritizing stability over complete interpretability.
- *Safety Criticality*: Applications where incorrect predictions could lead to safety incidents may require both high interpretability for operator trust and robust stability for reliable operation.
- *Maintenance Resources*: The availability of technical expertise and resources for model maintenance can influence whether to favor simpler, more interpretable models or complex, stable ones.

Several strategies can help mitigate these trade-offs. One approach involves using hybrid models that combine interpretable base models with stability-enhancing techniques. For example, employing ensemble methods with simple, interpretable base learners can provide both clarity in individual predictions and robustness across different operating conditions. Another strategy involves implementing post-hoc interpretation methods for stable, complex models, allowing users to understand specific predictions without compromising the model's robust performance.

D. Feature Selection and Data Analysis

Feature selection is a critical step in soft sensor development, directly impacting the model's interpretability and stability. Industrial processes often generate high-dimensional data, where not all measured variables are relevant to the target quality variable. Selecting the most informative features can reduce model complexity, improve generalization, and enhance interpretability by focusing on variables that directly influence the prediction [54].

Common feature selection techniques in soft sensor applications include filter methods, wrapper methods and embedded methods. Filter methods assess feature relevance based on statistical measures such as correlation coefficients, mutual information, or chi-square tests. They are computationally efficient but may overlook interactions between features [55]. Wrapper methods use a predictive model to evaluate subsets of features, selecting the combination that optimizes model performance. While more accurate, they are computationally expensive, especially for large datasets [54]. Embedded methods incorporate feature selection into the model training process, such as LASSO regression, which penalizes less important features. They balance computational efficiency and accuracy [56].

This article has been accepted for publication in IEEE Transactions on Instrumentation and Measurement. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIM.2025.3556830

Beyond feature selection, data analysis methods are essential for handling the unique characteristics of industrial data. Industrial time-series data often exhibit trends, seasonality, and autocorrelation, which must be accounted for in preprocessing. Key data analysis techniques include data preprocessing, missing value handling, outlier detection, and data drift detection.

Data preprocessing contains normalization, scaling, and transformation techniques that ensure data are suitable for model training [57]. Missing value handling involves imputation methods, such as mean imputation or k-nearest neighbors, to address incomplete data [58]. Outlier detection encompasses identifying and managing outliers, which is crucial as they can skew model predictions. Common methods include Z-scores or machine learning-based anomaly detection [59]. Data drift detection incorporates monitoring changes in data distribution over time, which is vital for maintaining model stability. Statistical tests like the Kolmogorov-Smirnov test can detect drift and trigger model updates [60].

Integrating feature selection with data analysis methods is key to developing soft sensors that perform reliably in industrial environments. By carefully selecting relevant features and preprocessing data to suit its unique characteristics, practitioners can significantly enhance the performance and reliability of soft sensor models.

E. Additional Challenges

The implementation of soft sensors faces several challenges beyond interpretability and stability. Data quality is a critical factor, as the performance of soft sensors, especially data-driven models, is highly dependent on the completeness and representativeness of the input data [61]. Industrial environments often generate noisy, incomplete, or outlier-containing data due to sensor malfunctions, communication errors, or manual recording inaccuracies. Poor data quality can lead to significant errors in soft sensor predictions, reducing system reliability. Techniques such as outlier detection, data cleaning, and imputation methods are employed to mitigate these issues, but they are not infallible.

Real-time processing requirements and scalability present further challenges. Many industrial applications require soft sensors to provide real-time or near real-time predictions for timely decision-making and process control. This poses computational challenges, particularly for complex machine learning algorithms such as deep learning [38]. Achieving real-time performance often requires optimizing both hardware and software components, potentially using lightweight models or edge computing solutions. Scalability becomes crucial in large-scale industrial operations where multiple processes and variables require simultaneous monitoring [62]. Soft sensors must maintain performance when scaled up to handle larger and more complex systems with numerous variables, especially in industries like oil and gas or chemical manufacturing.

IV. Enhancing Interpretability in Soft Sensors

Recent advances in interpretable machine learning have established a robust foundation for developing transparent and trustworthy models [63]–[66]. This foundation rests on understanding two fundamental aspects of interpretability: when explanation occurs during the modeling process (intrinsic vs. post-hoc interpretation) and what scope of model behavior is being explained (global vs. local interpretation). These distinctions guide practitioners in selecting appropriate interpretability methods for their specific applications.

The development of interpretability methods has progressed significantly with the introduction of frameworks that can explain both simple and complex models [67]. These frameworks incorporate dual perspectives: a human-centric approach that ensures explanations are meaningful to domain experts, and a quantitative approach that uses formal metrics to validate explanation accuracy [68]. The theoretical understanding of interpretability has also deepened through research examining fundamental questions about what constitutes meaningful interpretation and explanation [69]. This theoretical work has been complemented by practical advances in incorporating domain expertise into model development [70].

These developments reflect the maturation of explainable AI into a comprehensive field that emphasizes practical deployment, user trust, and regulatory compliance [71], [72]. The growing consensus is that interpretability is not merely desirable but essential for developing trustworthy and responsible machine learning systems.

Building on these advances in general interpretable machine learning, we now focus specifically on adapting these methods for soft sensor applications. The industrial context presents unique challenges that require careful consideration of how interpretability techniques can be effectively implemented while maintaining the high performance standards required in process monitoring and control. In this section, we will explore the significance of interpretability and examine various interpretable machine learning techniques that can be applied to make complex models more transparent. Our goal is to develop soft sensor models that not only achieve accurate predictions but also provide clear, trustworthy explanations of their decision-making processes.

A. Interpretable Machine Learning for Soft Sensors

To systematically understand interpretable machine learning, it is useful to categorize approaches along two key dimensions [8], [9]. The first dimension is based on whether the model inherently possesses interpretability, dividing the methods into *intrinsic interpretability* and *post hoc interpretability*. The second dimension considers whether the explanations focus on the entire dataset (*global interpretability*) or on individual predictions (*local interpretability*).

1) Intrinsic Interpretability and Post-hoc Interpretability: Intrinsic interpretability refers to models that are inherently transparent due to their simple and understandable structures. Examples include linear regression, logistic regression, decision trees, rule-based models, etc. These models have clear mathematical formulations or decision-making logic that allow users to directly understand how input features influence output predictions without the need for additional explanation tools.

In contrast, post-hoc interpretability involves applying techniques after model training to explain predictions from com-

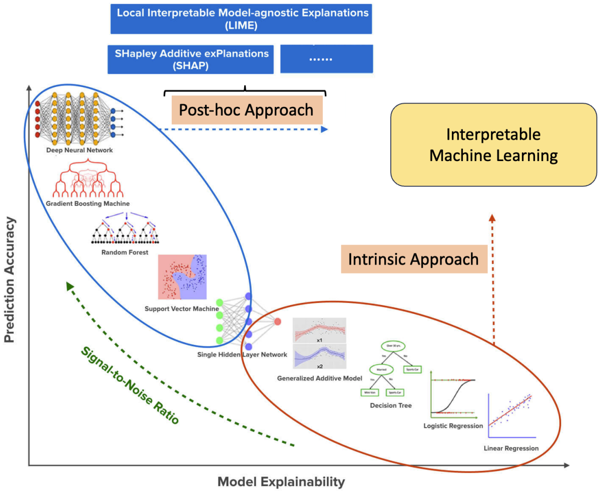

Fig. 4: Classification of interpretable machine learning

plex, often opaque models like deep neural networks or ensemble methods. Post-hoc methods, such as visualization techniques, perturbation analysis [73], and surrogate modeling [74], aim to provide insights into the model's decision-making process without altering its internal structure. Figure 4 shows the detailed classification for this dimension.

2) Local Interpretability and Global Interpretability: Local interpretability focuses on explaining individual predictions or a small subset of predictions. It aims to understand how changes in the input features of a specific instance affect the model's output. For example, if a soft sensor model produces an unexpected prediction at a particular moment, local explanations can help determine the cause by analyzing the influence of each feature on that specific prediction.

Global interpretability seeks to provide a holistic understanding of the model's overall behavior across the entire dataset. It involves analyzing general patterns, feature importance, and feature interactions that influence the model's predictions on average. Global explanations are crucial for validating the model, ensuring that it aligns with domain knowledge, and communicating its behavior to stakeholders. Table III presents a detailed analysis of various interpretable machine learning approaches classified according to these dimensions, highlighting their key features, advantages, and disadvantages.

B. Main Methods of Interpretable Machine Learning

This section explores the main methods of interpretable machine learning relevant to soft sensors, detailing their mechanisms, advantages, and limitations. A comprehensive understanding of these methods enables practitioners to select appropriate techniques tailored to the specific requirements of industrial applications.

Intrinsic interpretability methods, such as linear regression, serve as foundational models by providing direct coefficients that indicate the strength and direction of the relationships between input features and the target variable. These coefficients facilitate easy interpretation, making these models particularly suitable when simplicity and transparency are paramount. Similarly, decision trees construct a tree-like structure where each internal node represents a feature test, each branch signifies the outcome of the test, and each leaf node denotes a class label or continuous value. The decision path from the root to a leaf offers a clear and interpretable sequence of decisions. Rule-based models, including decision rules and association rules, employ simple if-then statements to encapsulate knowledge, ensuring that the captured patterns align with human reasoning and are easily understandable.

However, while these intrinsically interpretable methods excel in simplicity and transparency, they often fall short in capturing complex, nonlinear relationships inherent in dynamic industrial processes. This limitation can impede their predictive performance, particularly in scenarios that require high accuracy and the ability to model intricate dependencies.

To address these challenges, post-hoc interpretability methods are frequently employed, enabling the extraction of explanations from complex black-box models without sacrificing their predictive capabilities. Among post-hoc methods, local interpretable model-agnostic explanations (LIME) [74], [75] and Shapley additive explanations (SHAP) [9], [73], [76] are prominent due to their versatility and effectiveness. LIME approximates the behavior of a complex model locally around a specific instance by fitting an interpretable surrogate model, typically a linear regression. This is formalized by minimizing the following loss function:

$$\hat{g} = \arg\min_{g \in G} \sum_{z' \in Z} \pi_x(z') \left[f(z') - g(z') \right]^2 + \Omega(g) \tag{9}$$

where:

- G represents the family of interpretable models.
- Z is a set of perturbed samples around the instance x.
- $\pi_x(z')$ measures the proximity between x and z'.
- $\Omega(g)$ is a regularization term ensuring the simplicity of g.

By minimizing this loss, LIME identifies a simple model g that locally approximates the complex model f near the instance x, weighted by their proximity $\pi_x(z')$. This approach provides clear, instance-specific explanations, aiding operators in understanding individual predictions, which is crucial for debugging and building trust in critical industrial operations. However, LIME's explanations are inherently local and may not generalize to the model's global behavior.

SHAP, on the other hand, offers a unified framework based on cooperative game theory to compute feature attributions, known as Shapley values. For a model f and an instance x, the Shapley value ϕ_i for feature i is defined as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right] \tag{10}$$

where:

- N is the set of all features.
- S is a subset of features not containing i.
- $f_S(x_S)$ is the model prediction using features in S.

TABLE III: Analysis of Interpretable Machine Learning Approaches

Approach	Methods	Key Features	Advantages	Disadvantages
Intrinsic Interpretability	• Linear Regression • Decision Trees • Rule-Based Models	• Inherently simple and transparent structures • Clear relationships between input features and predictions	• Easily understandable without additional tools • Straightforward implementation and validation • Facilitates trust and transparency	• Limited in capturing complex, nonlinear relationships • May underperform compared to complex models
Post-hoc Interpretability	• LIME • SHAP • PDP • ICE	• Applied after model training to explain predictions • Compatible with complex, black-box models • Provides local or global explanations	• Flexible across model types • Enhances understanding without altering the model • Identifies feature importance and interactions	• Computationally intensive with large datasets • Potential for misleading or biased explanations • May lack consistency across instances
Global Explanations	• Feature Importance Scores • Global SHAP Values	• Provides overarching model behavior insights • Highlights features influencing predictions on average	• Useful for model validation and key driver identification • Simplifies communication with stakeholders	• May overlook nuances in individual predictions • Potential to oversimplify complex interactions
Local Explanations	• LIME • Local SHAP Values • Counterfactual Explanations	• Explains individual predictions • Provides insights into specific feature influences	• Aids in debugging and building trust • Identifies unique patterns or anomalies	• Not comprehensive of overall model behavior • Computationally expensive for multiple instances

- x_S is the instance x restricted to features in S.

This formulation ensures a fair allocation of the prediction to individual features by averaging the marginal contributions of each feature across all possible feature subsets. SHAP's strengths lie in its strong theoretical foundation, model-agnostic applicability, and the provision of consistent feature attributions where the sum of contributions equals the model's output. Additionally, SHAP offers robust visualization tools, such as summary plots and dependence plots, which effectively illustrate feature importance and interactions.

Methods like partial dependence plots (PDP) [77] and individual conditional expectation (ICE) plots [78] enhance the interpretability landscape. PDP visualizes the marginal effect of one or two features on the predicted outcome while holding other features constant, offering global insights into the model's behavior. ICE plots extend this by displaying the individual effects for each instance, revealing heterogeneity in feature impacts. Additionally, counterfactual explanations suggest minimal changes to input features that would alter the prediction outcome, providing actionable insights and helping users comprehend decision boundaries.

While post-hoc methods like LIME and SHAP facilitate the interpretation of complex models without altering their structure, they can be computationally demanding and may introduce approximation errors or biases if not implemented carefully. In industrial applications, selecting appropriate interpretable machine learning methods involves balancing factors such as process complexity, the trade-off between interpretability and performance, available computational resources, and stakeholder requirements. In environments where safety and regulatory compliance are critical, intrinsic interpretability methods may be favored to ensure transparency. Conversely, when capturing intricate process dynamics necessitates complex models, post-hoc methods become indispensable for explaining these models without compromising their predictive performance.

C. Open-Source Projects on Interpretable Machine Learning

Open-source projects play a pivotal role in advancing interpretable machine learning, providing practitioners and researchers with accessible tools to develop and deploy models that are accurate and transparent. In this subsection, we review several prominent open-source projects that are relevant to the development of interpretable soft sensors.

Developed by Microsoft, InterpretML (https://github.com/interpretml/interpret) is a toolkit that provides a suite of interpretable machine learning models and visualization tools, integrating both intrinsically and post-hoc interpretable models. It includes glass-box models like explainable boosting machines [79] that are inherently interpretable and incorporates methods like SHAP and LIME for explaining complex models. The toolkit offers interactive visualization dashboards to explore model explanations. Alibi (https://github.com/SeldonIO/alibi) is an open-source Python library focused on machine learning model interpretability, offering a range of explanation algorithms suitable for various data types. It provides diverse explanation methods, including counterfactual and contrastive explanations, and supports both model-agnostic and model-specific approaches.

Captum (https://github.com/pytorch/captum) is an interpretability library for PyTorch models, providing algorithms to understand and visualize feature importance and model predictions. It integrates seamlessly with PyTorch models and includes attribution methods such as integrated gradients and DeepLIFT [80], [81], along with visualization tools for interpreting results. AIX360 (https://github.com/Trusted-AI/AIX360) is an open-source toolkit from IBM that provides a comprehensive set of algorithms to support interpretability and explainability throughout the machine learning lifecycle. It includes diverse techniques for data, model, and prediction explanations, along with tools for fairness and bias detection. The toolkit offers extensive documentation and examples for various applications.

Several other open-source projects contribute to interpretable machine learning, including DALEX (https://github.com/ModelOriented/DALEX), Eli5 (https://github.com/TeamHG-Memex/eli5) and Fairlearn (https://github.com/fairlearn/fairlearn). DALEX provides tools for visualizing and understanding complex models and is available in both R and Python. Eli5 simplifies the debugging and explanation of machine learning classifiers and is compatible with scikit-learn models. Fairlearn focuses on assessing and mitigating fairness issues in machine learning models, ensuring that ethical considerations are integrated into model development. These projects extend the capabilities of interpretable machine learning by offering specialized tools for various aspects of model explanation, debugging, and fairness assessment.

Although general-purpose interpretability libraries provide foundational tools, adapting them for industrial soft sensor applications necessitates domain-specific customization and rigorous validation. In large-scale industrial processes, the data streams often contain significant noise, correlated process variables, and potential anomalies that can complicate interpretation. Libraries such as InterpretML or Alibi can assist in systematically selecting feature subsets and generating transparent explanations, but practitioners must tailor these frameworks to accommodate unique plant configurations and instrumentation workflows.

A critical initial step involves choosing model architectures that strike an optimal balance between explanatory power and predictive fidelity. For high-stakes quality monitoring, intrinsically interpretable models—like Explainable Boosting Machines (EBMs) or Generalized Additive Models (GAMs)—can yield both clarity and reliability. When deep neural networks are indispensable for their superior accuracy, lightweight interpretability methods such as Integrated Gradients (Captum) allow on-device computation of attributions, minimizing latency and resource consumption in real-time environments.

Once the model is in place, establishing iterative feedback loops is essential for sustained interpretability. Periodically comparing model explanations with empirical observations or known first-principle models helps engineers diagnose discrepancies early. In scenarios where unexpected predictions emerge, counterfactual explanations from Alibi can simulate alternative process conditions to reveal whether predicted outcomes align with operational experience. These "what-if" analyses not only foster trust but also guide proactive interventions—engineers can adjust setpoints or refine control strategies before deviations escalate into costly incidents.

V. Ensuring Stability in Soft Sensor Models

As industries move towards increased automation and reliance on real-time data, the stability of soft sensors becomes a fundamental requirement for successful implementation in dynamic industrial environments. Stability is crucial to avoid costly downtime, maintain product quality, and ensure safety in high-stakes industrial operations. Additionally, a stable soft sensor reduces the need for frequent recalibration or retraining, lowering operational costs. In this section, we discuss strategies to ensure the stability of soft sensor models. We will explore the factors influencing stability, such as environmental changes and data distribution shifts. Using causal machine learning techniques, we discuss methods to enhance the stability of soft sensor models against these challenges.

A. Factors Influencing Stability

Several factors inherent in industrial operations can affect the stability of soft sensor models. Understanding these factors is essential for developing robust models that can operate reliably under varying conditions.

1) Environmental Changes: Industrial processes often occur in environments subject to fluctuations in temperature, pressure, humidity, and other conditions. These changes can directly influence the properties of raw materials, equipment performance, and process variables. For example, temperature variations might affect the viscosity of fluids in a chemical process, impacting reaction rates and the sensor's predictions. Without mechanisms to account for environmental variations, a soft sensor's output may become unstable, leading to inaccurate or inconsistent predictions.

2) Data Distribution Shifts: A significant challenge to stability arises from changes in the underlying data distribution over time, known as data drift. Such shifts occur due to factors such as alterations in raw materials, changes in process dynamics, or sensor degradation. If the soft sensor model is not designed to handle such shifts, its predictions may become inaccurate. Techniques like domain adaptation [10], transfer learning [82] are used to address shifts in data distribution, ensuring that model performance remains stable as data evolve. In addition, factors such as sensor noise and maintenance cycles introduce variability into the system. Stable models should be resilient to these fluctuations and ideally adapt to new conditions with minimal intervention.

B. Causal Machine Learning for Improving Stability

Traditional machine learning models often rely on correlations within data to make predictions. However, models based solely on correlations can suffer from issues with stability and interpretability because they may capture spurious or non-causal relationships that are not robust across different environments or over time.

In industrial settings, causal machine learning has emerged as a powerful tool for enhancing the stability of soft sensor models. By focusing on identifying causal relationships between variables rather than mere correlations, models can provide more reliable and explainable predictions. Understanding the causal structure within the data enables the development of models that are robust to changes in the environment or operational conditions, as causal relationships tend to remain stable across different settings.

Here, we define causality as a relationship between random variables. Assume X and Y are two random variables, X is defined as the cause of Y, which means that the causal relationship $X \rightarrow Y$ exists if and only if the value of Y changes definitely as the value of X changes. Compared to correlation, causality is directional, clearly indicating how a treatment variable (also known as an independent variable) directly influences one or multiple outcome variables (also known as dependent variables) [83]–[85].

When examining the causal relationships between variables in a dataset, two primary approaches are widely used: structural causal models (SCM) [86] and the potential outcomes framework (POF) [85]. SCM utilizes directed acyclic graphs (DAGs) to illustrate causal connections, where each node represents a variable and each arrow indicates a causal influence from one variable to another. By analyzing observational data with SCM, researchers can identify both dependencies and conditional independencies among variables, helping to uncover possible causal structures. Once the causal structure is established, SCM can estimate the causal impact of individual variables on the outcome and also simulate the effects of external interventions, providing a solid theoretical basis for making informed decisions.

The potential outcomes framework, also known as the Rubin causal model, represents causality by assuming that each individual has potential outcomes Y for every possible treatment condition (for example, receiving or not receiving treatment). In POF, causal effects are defined by comparing the potential outcomes of the same individual under different treatment scenarios. However, since multiple potential outcomes for the same individual cannot be observed simultaneously, statistical methods are necessary to estimate these causal effects, such as calculating the average treatment effect. Randomized experiments are considered ideal for these estimations because randomization ensures that the treatment and control groups are comparable. To account for confounders that may influence the results, specialized statistical techniques such as propensity score matching or instrumental variable methods are essential to accurately estimate causal effects [84].

C. Methods for Causal Discovery In Industrial Processes

In this section, we dive into one major area of causal machine learning research: causal discovery. Causal discovery aims to automatically or semi-automatically identify potential causal relationships from observational data. Through causal discovery, we can effectively distinguish causal relationships in the data and use these relationships for prediction.

In general, consider that the observational data is composed of variables X_1, X_2, …, X_d, spanning d dimensions, with n samples in total. The goal of causal discovery is to utilize these observational data to derive a causal graph made up of d nodes, where each node represents one of the variables X_1, X_2, …, X_d. The resulting causal graph from a causal discovery algorithm depicts the causal relationships among the variables in the observational data. This graph is a directed acyclic graph. Moreover, depending on the chosen causal discovery algorithm and the particular problem context, the directed edges may be weighted or unweighted. Causal discovery methods can be broadly classified into three main types (as shown in Figure 5): randomized experiments [87], computer simulation experiments [88], and methods based on observational data [89]–[105]. Table IV presents a detailed overview of causal discovery methods in industrial processes.

Randomized experiments are a conventional method for identifying causality, yet they are both expensive and time-intensive. In this approach, subjects are randomly allocated to different groups, each subjected to different interventions. When the subject pool is large enough, this technique can minimize the influence of both known and unknown confounders within each group. Nevertheless, conducting real experiments in sophisticated large-scale industrial processes is seldom practical. Alternatively, computer simulation tools like Aspen demand considerable expertise and extensive data for physical modeling. Consequently, discerning causal relationships in complex industrial processes is challenging without a comprehensive understanding of the physical model and access to substantial data.

Discovering causal relationships using observational data overcomes these constraints and is currently a prominent area of research in the field of causality. Causal discovery methods leveraging observational data can be categorized into those based on non-temporal data [89]–[98] and those based on temporal data [103]–[105]. Among temporal methods, the Granger causality analysis (GCA) [104] and the transfer entropy (TE) [103] are frequently used to determine pairwise causality between variables. However, pairwise causality is limited when addressing indirect causality and confounders. Although temporal data can offer significant causal insight, findings are often affected by temporal resolution. Typically, it is challenging to infer high-resolution causal relationships from data with lower temporal resolution.

In industrial processes, causal discovery algorithms that utilize non-temporal observational data are widely applicable. These algorithms are mainly divided into three categories: constraint-based [89], [90], [93], [97], causal function-based [91], [92], [99], [100], and score-based [98]. Constraint-based algorithms, such as Peter-Clark (PC) [89], inductive causation (IC) [90], and fast causal inference (FCI) [97], build the causal structure via conditional independence tests and determine the causal direction using predefined rules [97]. These methods typically yield partially undetermined causal directions. To address this issue, researchers have introduced causal function approaches like linear non-Gaussian acyclic model (LiNGAM) [91], [92], additive noise model (ANM) [99], and the post-nonlinear model (PNL) [100], which rely on specific assumptions about the data generation process. Score-based algorithms, on the other hand, employ scoring functions and search algorithms to identify the optimal Bayesian causal network [98]. However, this approach involves a graph search operation with high time complexity and assumes that all confounders are observable, which is often impractical.

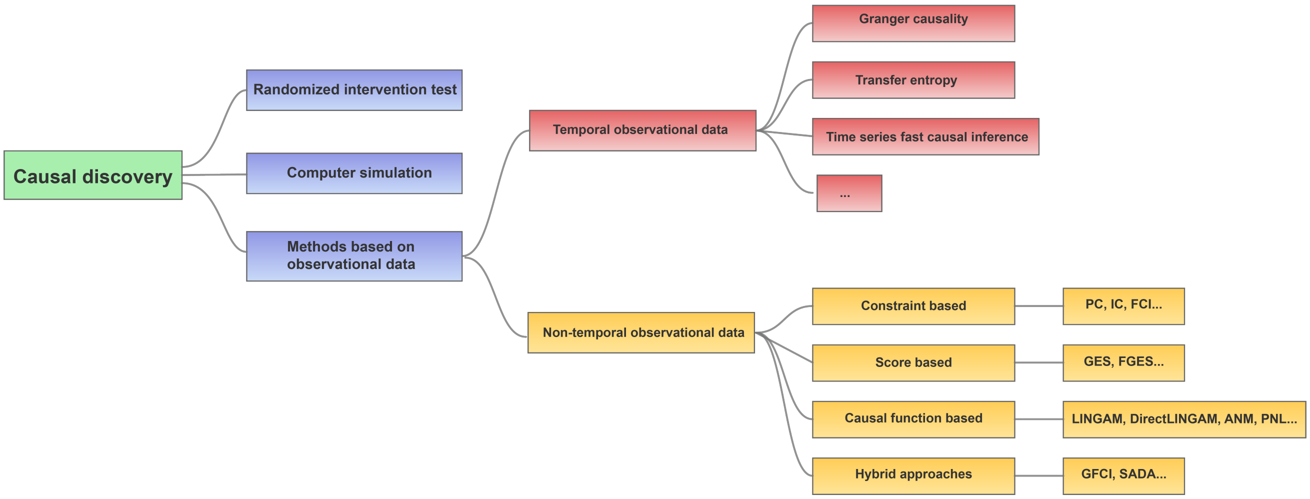
Causal discovery
Randomized intervention test
Computer simulation
Methods based on observational data
Temporal observational data
Granger causality
Transfer entropy
Time series fast causal inference
...
Non-temporal observational data
Constraint based
PC, IC, FCI...
Score based
GES, FGES...
Causal function based
LINGAM, DirectLINGAM, ANM, PNL...
Hybrid approaches
GFCI, SADA...

Fig. 5: Classification of Causal Discovery Methods

TABLE IV: Overview of Causal Discovery Methods in Industrial Processes

Method Category	Specific Methods	Advantages	Disadvantages
Randomized Experiments	• A/B Testing • Factorial Experiments • Randomized Controlled Trials	• Establishes causality with high confidence through randomization • Effectively removes confounding variables • Provides clear cause-and-effect links	• Often costly and time-consuming in industrial settings • Ethical issues may restrict experimentation on certain variables • Limited scalability for different processes and large systems
Computer Simulation	• Process Simulation Software (e.g., Aspen Plus, HYSYS) • System Dynamics Models	• Explore many scenarios without physical experiments • Incorporates complex physical, chemical, and biological processes • Test various hypothetical conditions safely	• Depends on the accuracy of simulation models and input data • May miss real-world complexities and emergent behaviors • Requires specialized knowledge to develop and interpret results
Observational Data	• Constraint-Based Methods • Causal Function-Based Methods • Score-Based Methods • Structural Equation Modeling (SEM)	• Uses existing data, reducing need for costly experiments • Uncovers complex causal relationships and interactions • Scales to large datasets with many variables	• May be affected by unmeasured or hidden confounders • Requires large, high-quality, well-curated datasets • Assumptions of specific algorithms may not hold in practice • Difficult to validate causal links without experimental data

Despite the progress made through existing methods in causal discovery, there remains a need for further examination and enhancement of their shortcomings. Researchers are investigating hybrid techniques that merge the benefits of constraints-based, score-based, and causal function approaches to counterbalance the weaknesses of each method. Additionally, the combination of deep learning with causal inference shows great potential for tackling challenges like unobserved confounders and high computational demands. As the discipline advances, future studies should focus on creating more robust, scalable, and broadly applicable algorithms.

D. Open-Source Projects on Causal Machine Learning

In this subsection, we review several prominent open-source projects that are relevant to the development of causal soft sensors. Developed by Microsoft, DoWhy (https://github.com/microsoft/dowhy) is a Python library that provides a principled approach to causal inference by integrating causal graph-based methods with statistical techniques. DoWhy allows users to define causal models using graphical representations, identify causal effects, estimate these effects using various estimators, and perform robustness checks to validate the findings. CausalML (https://github.com/uber/causalml) is an open-source Python package developed by Uber that focuses on estimating heterogeneous treatment effects from observational data. It leverages machine learning algorithms to provide scalable and efficient causal inference solutions, implementing various meta-learners such as the S-learner, T-learner, and X-learner. CausalML is particularly useful for applications like personalized marketing, recommendation systems, and A/B testing, where understanding the differential impact of treatments is crucial.

EconML (https://github.com/microsoft/EconML) is another

Python library from Microsoft that bridges econometric techniques with machine learning for causal inference. It is designed to estimate heterogeneous treatment effects using advanced machine learning methods, including double machine learning and causal forests. EconML provides a flexible framework that integrates seamlessly with scikit-learn, enabling users to incorporate causal inference into existing machine learning pipelines effectively. CausalImpact is particularly useful for evaluating the effectiveness of marketing campaigns, policy changes, and other temporal interventions in industrial settings. CausalNex (https://github.com/quantumblacklabs/causalnex) is a Python library developed by QuantumBlack. It combines causal discovery algorithms with probabilistic modeling to uncover and leverage causal relationships in data. CausalNex supports intervention simulation and provides visualization tools to represent and interpret causal graphs.

Several other open-source projects contribute to causal machine learning, including Tigramite (https://github.com/jakobrun/tigramite), Causal Discovery Toolbox (https://github.com/FenTechSolutions/CausalDiscoveryToolbox), and CausalPy (https://github.com/microsoft/CausalPy). Tigramite specializes in causal discovery for time series data. The Causal Discovery Toolbox offers a comprehensive suite of algorithms for uncovering causal relationships from various data types, supporting constraint-based, score-based, and functional causal models. CausalPy provides tools for causal analysis, including discovery, effect estimation, and counterfactual reasoning, with a user-friendly API that integrates with popular data science libraries.

These projects extend the capabilities of causal machine learning by offering specialized tools for causal discovery, effect estimation, and counterfactual analysis, facilitating their deployment in industrial soft sensors. This alignment with the goals of modern industrial process monitoring and control ensures that soft sensors are effective and reliable, fostering trust in complex industrial environments.

Although existing libraries for causal machine learning were originally developed with general-purpose use, adapting them effectively for soft sensor tasks in industrial settings requires additional domain-specific considerations. In industrial systems, datasets often contain highly correlated process variables or extraneous measurements that can obscure meaningful causal links. Libraries like DoWhy and CausalNex assist in iteratively refining candidate graph structures by testing conditional independence and eliminating variables that contribute noise or confounding. This process maximizes the clarity of causal insights extracted from complex industrial data streams.

Another powerful use of causal toolkits lies in simulating interventions and counterfactuals. Tools such as CausalML provide meta-learners that estimate treatment effects, enabling engineers to anticipate how process modifications may influence key outcomes. This capacity to test interventions in a virtual, risk-free manner significantly reduces uncertainty in operational decision-making.

After the causal structures or effects have been estimated, validating them under real-world process shifts is critical. Factors such as changing feed composition or ambient temperature can challenge the robustness of any learned causal model. Domain-driven hypothesis testing, combined with library-specific stability checks, can quickly detect when causal relationships no longer hold. When inconsistencies arise, these diagnostic methods facilitate targeted model retraining or recalibration, preventing performance drifts that could undermine trust in the soft sensor.

VI. Industrial Application

The practical implementation of interpretable and stable soft sensors demonstrates significant value across industrial sectors. This section presents a comprehensive case study from petroleum refining to illustrate how causal feature selection and SHAP-based explanations can enhance both interpretability and stability of soft sensor.

In modern refineries, maintaining a precise flash point for diesel fuel is paramount for safety and product quality. However, accurately measuring the flash point in real time poses a challenge due to harsh process conditions and the need for frequent laboratory testing. To address this, a data-driven soft sensor was developed within a diesel hydro-treating unit. The primary objective was to provide stable and interpretable flash-point estimates based on routine process measurement, thus enabling operators to adjust key variables promptly in response to any observed deviations.

Over 6000 samples were gathered from a diesel hydro-treating unit in the Parkland refinery in Canada. Each sample contained 24 routinely measured input variables (e.g., temperature, pressure, flow rates), along with the corresponding measured flash point. To rigorously assess robustness, the dataset was divided into a training set (the first 4800 samples) and two separate test sets (600 samples each). These test sets intentionally represented distinctly different operating conditions, helping validate the model's stability.

To ensure that the model captures genuine causal effects rather than spurious correlations, the study employed a combination of FCI and Granger causality. This approach identified 9 key features with direct causal relationships to the flash point, excluding variables whose correlations were found to be indirect or environment-specific. By focusing on causally relevant predictors, the resulting soft sensor is inherently more robust to distribution shifts commonly encountered in refining processes.

Using the selected causal features, an Extra Trees regressor was trained to predict the flash point [36]. Figure 6 shows the performance across the two distinct test distributions. This indicates that the proposed model maintained consistent accuracy despite considerable shifts in process conditions.

To explain the internal decision-making mechanism, SHAP was computed for each feature. These SHAP values quantified how much each feature contributed—positively or negatively—to each prediction, enabling process engineers to pinpoint the root causes of sudden flash-point variations. This interpretability not only bolstered user trust in the soft sensor's decisions but also facilitated timely interventions. For instance, if the SHAP contribution of a particular temperature variable significantly deviated from its normal range, operators could

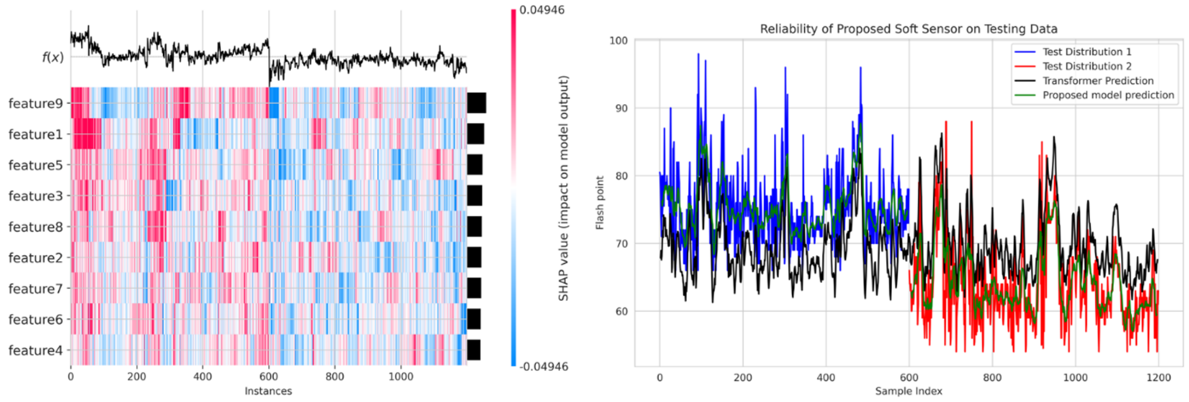

Fig. 6: SHAP Value and Performance of Proposed Soft Sensor on Testing Data

investigate potential equipment malfunctions or feedstock inconsistencies before they escalated into major operational disruptions.

In addition to the diesel hydro-treating unit case, these interpretable and causal machine learning approaches can be applied to a variety of other industrial settings. For example, in pharmaceutical fermentation, real-time monitoring of critical quality attributes (CQAs)—such as biomass concentration, metabolite levels, and product yields—is essential to maintain compliance with stringent regulatory guidelines and ensure product efficacy. However, these CQAs can be both costly and difficult to measure directly. Soft sensors designed with interpretable machine learning techniques, coupled with causal feature selection algorithms, can systematically identify a smaller subset of truly influential variables (e.g., pH, dissolved oxygen, substrate feed rate). By honing in on causal rather than purely correlational factors, these models offer more stable and transparent predictions. This enhanced interpretability empowers process engineers to understand precisely why certain features grow in importance under evolving fermentation conditions, enabling faster root-cause analysis and more informed decision-making—both critical for optimizing operations and simplifying regulatory reporting.

VII. Future Directions and Research Opportunities

Looking toward the future, there are numerous opportunities to push the boundaries of soft sensors by addressing existing challenges and exploring new methodologies. Enhancing interpretability and stability remains a central theme, but several specific research directions can significantly advance the field.

One promising area is the development of adaptive and online learning models that can operate effectively in non-stationary environments. Industrial processes are subject to continuous changes. By employing techniques such as online learning, meta-learning, and reinforcement learning, soft sensors can be designed to update their parameters in real-time, adapting to new data without the need for extensive retraining [33], [34], [40], [41], [43], [106]. This adaptability not only improves model stability but also reduces maintenance costs and downtime associated with model recalibration.

Advanced data fusion methods represent another critical area for innovation. Industrial environments generate a large amount of heterogeneous data, including sensor measurements, images, and textual logs [61]. Effectively integrating these diverse data sources can enrich soft sensor models, leading to more accurate and comprehensive predictions. Research into multimodal data fusion [107], deep representation learning [108], and transfer learning [82] can facilitate the development of models capable of handling complex data types and extracting meaningful features across different modalities.

Improving computational efficiency for real-time applications is also a significant research focus. Many existing soft sensor models, especially those based on deep learning, are computationally intensive and may not meet the requirements of time-critical industrial processes. Innovations in model compression techniques, such as pruning and quantization [109], as well as the exploration of lightweight neural network architectures, can help reduce computational demands [110]. Additionally, leveraging advancements in hardware acceleration, such as GPUs and TPUs, and exploring edge computing paradigms can facilitate real-time processing and deployment of soft sensors in industrial settings.

Privacy and security considerations are important, especially with the rise of the Industrial Internet of Things. Developing privacy-preserving machine learning techniques, such as federated learning and differential privacy, can enable collaborative model training without compromising sensitive data [111], [112]. This approach allows multiple industrial sites or organizations to benefit from shared learning without exposing proprietary information. Furthermore, there is potential in integrating soft sensor technologies with digital twin systems. Digital twins are virtual replicas of physical assets, processes, or systems that can be used for simulation, analysis, and control [113], [114]. By coupling soft sensors with digital twins, it is possible to create more accurate and responsive models that reflect real-time changes in the physical system.

The future of soft sensor technologies lies in addressing current gaps. This requires interdisciplinary research that combines advances in machine learning, control engineering, and domain-specific knowledge. Developing models that are interpretable, stable, and adaptable is essential. By leveraging emerging technologies and methodologies, soft sensors can become more reliable and widely adopted tools in industrial process monitoring and control. Continued collaboration between academia and industry is essential. This collaboration will translate innovations into practical solutions that meet the evolving needs of modern industrial operations.

VIII. Conclusion

This review presented a comprehensive analysis of methodologies to enhance the interpretability and stability of soft sensors in industrial processes. For interpretability, we explored various interpretable machine learning techniques that can be applied to complex models to improve transparency and trustworthiness. We also highlighted open-source tools that facilitate the practical implementation of these techniques. On the topic of stability, we emphasized the role of causal machine learning and discussed key causal discovery methods. These approaches provide more robust model predictions by focusing on stable causal relationships rather than mere correlations. In addition, we introduced relevant open-source projects that help to apply these methods to real-world industrial processes. A practical application of interpretable and stable soft sensor in a diesel hydro-treating unit further demonstrated the potential to significantly impact industrial process monitoring and control. As industries continue to evolve towards more automated and data-driven methodologies, the development of sophisticated, reliable, and transparent soft sensors will be crucial. This review serves as a foundation for future research and innovation in this vital field, ultimately contributing to safer, more efficient, and sustainable industrial operations.

References

[1] R. B. Gopaluni, A. Tulsyan, B. Chachuat, B. Huang, J. M. Lee, F. Amjad, S. K. Damarla, J. W. Kim, and N. P. Lawrence, "Modern machine learning tools for monitoring and control of industrial processes: A survey," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 218–229, 2020, 21st IFAC World Congress.

[2] Q. Zhu, S. Joe Qin, and Y. Dong, "Dynamic latent variable regression for inferential sensor modeling and monitoring," *Computers & Chemical Engineering*, vol. 137, p. 106809, 2020.

[3] B. Mehta and Y. Reddy, *Industrial Process Automation Systems Design and Implementation*. Oxford: Butterworth-Heinemann, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128009390000206

[4] G. Fang, Y. Liu, B. Cai, H. Chen, and D. Huang, "A hierarchical soft-sensor using spatiotemporal information transformation and arma with application in wastewater treatment," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–11, 2023.

[5] J. Su, L. Cao, G. Lee, B. Gopaluni, L. C. Siang, Y. Cao, S. van Dyk, R. Pinchuk, and J. Saddler, "Tracking the green coke production when co-processing lipids at a commercial fluid catalytic cracker (fcc): combining isotope 14c and causal discovery analysis," *Sustainable Energy & Fuels*, vol. 6, pp. 5600–5607, 2022.

[6] M. S. Hong, K. A. Severson, M. Jiang, A. E. Lu, J. C. Love, and R. D. Braatz, "Challenges and opportunities in biopharmaceutical manufacturing control," *Computers & Chemical Engineering*, vol. 110, pp. 106–114, 2018.

[7] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Communications of the ACM*, vol. 63, no. 1, pp. 68–77, 2019.

[8] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22 071–22 080, 2019.

[9] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[10] X. Zhang, P. Cui, R. Xu, L. Zhou, Y. He, and Z. Shen, "Deep stable learning for out-of-distribution generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5372–5382.

[11] K. Kuang, P. Cui, S. Athey, R. Xiong, and B. Li, "Stable prediction across unknown environments," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, ser. KDD '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1617–1626.

[12] R. Xu, P. Cui, Z. Shen, X. Zhang, and T. Zhang, "Why stable learning works? A theory of covariate shift generalization," 2021. [Online]. Available: https://arxiv.org/abs/2111.02355

[13] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Communications of the ACM*, vol. 63, no. 1, pp. 68–77, 2019.

[14] K. Kuang, P. Cui, S. Athey, R. Xiong, and B. Li, "Stable prediction across unknown environments," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1617–1626. [Online]. Available: https://doi.org/10.1145/3219819.3220082

[15] L. Cao, F. Yu, F. Yang, Y. Cao, and R. B. Gopaluni, "Data-driven dynamic inferential sensors based on causality analysis," *Control Engineering Practice*, vol. 104, p. 104626, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0967066120301969

[16] R. Weber and C. Brosilow, "The use of secondary measurements to improve control," *AIChE Journal*, vol. 18, no. 3, pp. 614–623, 1972. [Online]. Available: https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.690180323

[17] T. J. McAvoy, "Contemplative stance for chemical process control: An ifac report," *Automatica*, vol. 28, no. 2, pp. 441–442, 1992. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0005109892901342

[18] P. Kadlec, B. Gabrys, and S. Strandt, "Data-driven soft sensors in the process industry," *Computers Chemical Engineering*, vol. 33, no. 4, pp. 795–814, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0098135409000076

[19] A. Torgashov and S. Skogestad, "The use of first principles model for evaluation of adaptive soft sensor for multicomponent distillation unit," *Chemical Engineering Research and Design*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:202880353

[20] T. Bikmukhametov and J. Jäschke, "First principles and machine learning virtual flow metering: A literature review," *Journal of Petroleum Science and Engineering*, vol. 184, p. 106487, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0920410519309088

[21] L. Cao, J. Su, E. Conde, L. C. Siang, Y. Cao, and B. Gopaluni, "A novel automated soft sensor design tool for industrial applications based on machine learning," *Control Engineering Practice*, vol. 160, p. 106322, 2025.

[22] Y.-L. He, Y. Zhao, Q.-X. Zhu, and Y. Xu, "Online distributed process monitoring and alarm analysis using novel canonical variate analysis with multicorrelation blocks and enhanced contribution plot," *Industrial & Engineering Chemistry Research*, vol. 59, no. 45, pp. 20 045–20 057, 2020. [Online]. Available: https://doi.org/10.1021/acs.iecr.0c02209

[23] J. Fan, S. J. Qin, and Y. Wang, "Online monitoring of nonlinear multivariate industrial processes using filtering kica–pca," *Control Engineering Practice*, vol. 22, pp. 205–216, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0967066113001275

[24] Q. Zhu, S. Joe Qin, and Y. Dong, "Dynamic latent variable regression for inferential sensor modeling and monitoring," *Computers & Chemical Engineering*, vol. 137, p. 106809, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0098135419312323

[25] H. Abdi and L. J. Williams, "Principal component analysis," *WIREs Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[26] Y. Dong and S. J. Qin, "Dynamic-inner partial least squares for dynamic data modeling," *IFAC-PapersOnLine*, vol. 48, pp. 117–122, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:61900874

[27] Q. Zhu, S.-Z. J. Qin, and Y. Dong, "Dynamic latent variable regression for inferential sensor modeling and monitoring," *Comput. Chem. Eng.*, vol. 137, p. 106809, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:216368753

[28] B. Khaldi, F. Harrou, S. mohamed Benslimane, and Y. Sun, "A data-driven soft sensor for swarm motion speed prediction using ensemble learning methods," *IEEE Sensors Journal*, vol. 21, pp. 19 025–19 037,

2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:236260561

[29] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[30] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[31] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[32] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv preprint arXiv:1802.03888*, 2018.

[33] X. Gao, Y. ching Liu, X. Yi, and D. Huang, "Novel multimodal data fusion soft sensor modeling framework based on meta-learning networks for complex chemical process," *IFAC-PapersOnLine*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251430871

[34] C. Qiume, "Soft-sensor of water quality based on integrated elm with meta-learning," *Information & Computation*, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:124624199

[35] G. Biau, "Analysis of a random forests model," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1063–1095, 2012.

[36] L. Cao, X. Ji, Y. Cao, Y. Li, L. C. Siang, J. Li, V. K. Pediredla, and R. B. Gopaluni, "Interpretable soft sensors using extremely randomized trees and shap," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 8000–8005, 2023.

[37] B. Y. . H. LeCun, Y., "Deep learning," *Nature*, vol. 521, p. 436–444, 2015.

[38] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[39] X. Wang, "Data preprocessing for soft sensor using generative adversarial networks," *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pp. 1355–1360, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:56596580

[40] J. Xie, O. Dogru, B. Huang, C. Godwaldt, and B. Willms, "Reinforcement learning for soft sensor design through autonomous cross-domain data selection," *Computers & Chemical Engineering*, vol. 173, p. 108209, 2023.

[41] O. Dogru, J. Xie, O. Prakash, R. Chiplunkar, J. Soesanto, H. Chen, K. Velswamy, F. Ibrahim, and B. Huang, "Reinforcement learning in process industries: Review and perspective," *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 2, pp. 283–300, 2024.

[42] W. Hu, C. Du, F. Li, X. Chen, and W. Gui, "Distributed model-free deep reinforcement learning control for unknown nonlinear multi-agent systems," *Journal of the Franklin Institute*, p. 107636, 2025.

[43] E. Skordilis and R. Moghaddass, "A deep reinforcement learning approach for real-time sensor-driven decision making and predictive analytics," *Computers & Industrial Engineering*, vol. 147, p. 106600, 2020.

[44] W. Jia, T. you Chai, and W. Yu, "A novel hybrid neural network for modeling rare-earth extraction process," *IFAC Proceedings Volumes*, vol. 41, pp. 11 427–11 432, 2008.

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

[46] W. Peng, W. Yao, W. Zhou, X. Zhang, and W. Yao, "Robust regression with highly corrupted data via physics informed neural networks," *ArXiv*, vol. abs/2210.10646, 2022.

[47] R. Romijn, L. Özkan, S. Weiland, J. Ludlage, and W. Marquardt, "A grey-box modeling approach for the reduction of nonlinear systems," *Journal of Process Control*, vol. 18, no. 9, pp. 906–914, 2008.

[48] S. Mitra and Y. Hayashi, "Neuro-fuzzy rule generation: survey in soft computing framework," *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 748–768, 2000.

[49] A. Moser, C. Appl, S. Brüning, and V. C. Hass, *Mechanistic Mathematical Models as a Basis for Digital Twins*. Springer International Publishing, 2021, pp. 133–180.

[50] T. Liu, G. Lugosi, G. Neu, and D. Tao, "Algorithmic stability and hypothesis complexity," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2159–2167.

[51] S. Nogueira, K. Sechidis, and G. Brown, "On the stability of feature selection algorithms," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6345–6398, 2017.

[52] J. Ahn, "A stable hyperparameter selection for the gaussian rbf kernel for discrimination," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 3, no. 3, pp. 142–148, 2010.

[53] M. Feurer and F. Hutter, "Hyperparameter optimization," *Automated machine learning: Methods, systems, challenges*, pp. 3–33, 2019.

[54] L. Cao, J. Su, J. Saddler, Y. Cao, Y. Wang, G. Lee, L. C. Siang, Y. Luo, R. Pinchuk, J. Li *et al.*, "Machine learning for real-time green carbon dioxide tracking in refinery processes," *Renewable and Sustainable Energy Reviews*, vol. 213, p. 115417, 2025.

[55] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, 1994.

[56] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[57] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2011.

[58] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[59] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review*, vol. 22, no. 2, pp. 85–126, 2004.

[60] G. I. Webb, L. K. Lee, B. Goethals, and F. Petitjean, "Understanding concept drift," *arXiv preprint arXiv:1704.00362*, 2016.

[61] J. Fan, F. Han, and H. Liu, "Challenges of Big Data analysis," *National Science Review*, vol. 1, no. 2, pp. 293–314, Feb 2014.

[62] J. Zhu, M. Jiang, G. Peng, L. Yao, and Z. Ge, "Scalable soft sensor for nonlinear industrial big data via bagging stochastic variational gaussian processes," *IEEE Transactions on Industrial Electronics*, vol. 68, no. 8, pp. 7594–7602, 2021.

[63] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[64] R. Guidotti *et al.*, "A survey of methods for explaining black box models," *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2019.

[65] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Lulu. com, 2019.

[66] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[67] M. Scott, L. Su-In *et al.*, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, pp. 4765–4774, 2017.

[68] D. Nguyen, "Comparing automatic and human evaluation of local explanations for text classification," pp. 1069–1078, 2018.

[69] A. Sudjianto and A. Zhang, "Designing inherently interpretable machine learning models," *arXiv preprint arXiv:2111.01743*, 2021.

[70] K. Beckh, S. Müller, M. Jakobs, V. Toborek, H. Tan, R. Fischer, P. Welke, S. Houben, and L. von Rueden, "Explainable machine learning with prior knowledge: an overview," *arXiv preprint arXiv:2105.10172*, 2021.

[71] P. J. Lisboa, S. Saralajew, A. Vellido, R. Fernández-Domenech, and T. Villmann, "The coming of age of interpretable and explainable machine learning models," *Neurocomputing*, vol. 535, pp. 25–39, 2023.

[72] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22 071–22 080, 2019.

[73] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[74] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144.

[75] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling lime and shap: Adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 180–186.

[76] L. S. Shapley, "Notes on the n-person game—ii: The value of an n-person game," 1951.

[77] J. Moosbauer, J. Herbinger, G. Casalicchio, M. Lindauer, and B. Bischl, "Explaining hyperparameter optimization via partial dependence plots," in *Advances in Neural Information Processing Systems*, M. Ranzato,

A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 2280–2291.

[78] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, 2015.

[79] V. Dsilva, J. Schleiss, and S. Stober, "Trustworthy academic risk prediction with explainable boosting machines," in *Artificial Intelligence in Education*, N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, and V. Dimitrova, Eds. Cham: Springer Nature Switzerland, 2023, pp. 463–475.

[80] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for pytorch," *CoRR*, vol. abs/2009.07896, 2020.

[81] M. Shajalal, A. Boden, and G. Stevens, "Forecastexplainer: Explainable household energy demand forecasting by approximating shapley values using deeplift," *Technological Forecasting and Social Change*, vol. 206, p. 123588, 2024.

[82] Y. Wang, J. Zhu, L. Cao, B. Gopaluni, and Y. Cao, "Long short-term memory network with transfer learning for lithium-ion battery capacity fade and cycle life prediction," *Applied Energy*, vol. 350, p. 121660, 2023.

[83] G. Imbens, "Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics," *NBER Working Paper Series*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:197431019

[84] P. C. Austin, N. Thomas, and D. B. Rubin, "Covariate-adjusted survival analyses in propensity-score matched samples: Imputing potential time-to-event outcomes," *Statistical Methods in Medical Research*, vol. 29, pp. 728 – 751, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:58662109

[85] G. Imbens and D. B. Rubin, "Causal inference for statistics, social, and biomedical sciences: An introduction," 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:52036833

[86] J. Pearl, "Causality: Models, reasoning and inference," 2000. [Online]. Available: https://api.semanticscholar.org/CorpusID:12575481

[87] H. Hu, Z. Li, and A. R. Vetta, "Randomized experimental design for causal graph discovery," *Advances in Neural Information Processing Systems*, vol. 27, pp. 1–10, 2014. [Online]. Available: https://proceedings.neurips.cc/paper/2014/file/e53a0a2978c28872a4505bdb51db06dc-Paper.pdf

[88] N. J. Matiasz, J. Wood, W. Wang, A. J. Silva, and W. Hsu, "Computer-aided experiment planning toward causal discovery in neuroscience," *Frontiers in Neuroinformatics*, vol. 11, p. 12, 2017. [Online]. Available: https://www.frontiersin.org/article/10.3389/fninf.2017.00012

[89] P. Spirtes and C. Glymour, "An algorithm for fast recovery of sparse causal graphs," *Social Science Computer Review*, vol. 9, no. 1, pp. 62–72, 1991.

[90] T. Verma and J. Pearl, "Equivalence and synthesis of causal models," in *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, ser. UAI '90. USA: Elsevier Science Inc., 1990, p. 255–270.

[91] S. Shimizu, P. O. Hoyer, A. Hyvarinen, and A. Kerminen, "A linear non-gaussian acyclic model for causal discovery," *Journal of Machine Learning Research*, vol. 7, no. 72, pp. 2003–2030, 2006. [Online]. Available: http://jmlr.org/papers/v7/shimizu06a.html

[92] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen, "Directlingam: A direct method for learning a linear non-gaussian structural equation model," *J. Mach. Learn. Res.*, vol. 12, no. null, p. 1225–1248, Jul. 2011.

[93] P. Spirtes, C. Meek, and T. Richardson, "Causal inference in the presence of latent variables and selection bias," in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, ser. UAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, p. 499–506.

[94] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in Genetics*, vol. 10, p. 524, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fgene.2019.00524

[95] J. M. Ogarrio, P. Spirtes, and J. Ramsey, "A hybrid causal search algorithm for latent variable models," in *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, ser. Proceedings of Machine Learning Research, vol. 52. Lugano, Switzerland: PMLR, 06–09 Sep 2016, pp. 368–379.

[96] R. Cai, Z. Zhang, and Z. Hao, "Sada: A general framework to support robust causation discovery," in *Proceedings of the 30th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, S. Dasgupta and D. McAllester, Eds., vol. 28, no. 2. Atlanta, Georgia, USA: PMLR, 17–19 Jun 2013, pp. 208–216. [Online]. Available: https://proceedings.mlr.press/v28/cai13.html

[97] G. F. Cooper and C. Glymour, *Computation, Causation, and Discovery*. AAAI Press: California, 05 1999. [Online]. Available: https://doi.org/10.7551/mitpress/2006.001.0001

[98] J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour, "A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images," *International Journal of Data Science and Analytics*, vol. 3, no. 2, p. 121–129, 2017.

[99] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," *Advances in Neural Information Processing Systems*, vol. 689, no. 2009, pp. 689–696, 2008.

[100] K. Zhang and A. Hyvärinen, "On the identifiability of the post-nonlinear causal model," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, ser. UAI '09. Arlington, Virginia, USA: AUAI Press, 2009, p. 647–655.

[101] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, and J. Zscheischler, "Inferring causation from time series in earth system sciences," *Nature Communications*, vol. 10, no. 1, 2019.

[102] J. Runge, "Causal network reconstruction from time series: From theoretical assumptions to practical estimation," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 28, no. 7, p. 075310, 2018. [Online]. Available: https://doi.org/10.1063/1.5025050

[103] T. Schreiber, "Measuring information transfer," *Phys. Rev. Lett.*, vol. 85, pp. 461–464, Jul 2000. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.85.461

[104] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[105] L. Barnett, A. B. Barrett, and A. K. Seth, "Granger causality and transfer entropy are equivalent for gaussian variables," *Physical Review Letters*, vol. 103, no. 23, p. 238701, 2009.

[106] S. Spielberg, R. Gopaluni, and P. Loewen, "Deep reinforcement learning approaches for process control," in *2017 6th International Symposium on Advanced Control of Industrial Processes (AdCONIP)*, 2017, pp. 201–206.

[107] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[108] X. Yan, J. Wang, and Q. Jiang, "Deep relevant representation learning for soft sensing," *Information Sciences*, vol. 514, pp. 263–274, 2020.

[109] A. Kuzmin, M. Nagel, M. Van Baalen, A. Behboodi, and T. Blankevoort, "Pruning vs quantization: which is better?" *Advances in neural information processing systems*, vol. 36, 2024.

[110] D. Yao, H. Liu, J. Yang, and X. Li, "A lightweight neural network with strong robustness for bearing fault diagnosis," *Measurement*, vol. 159, p. 107756, 2020.

[111] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," *Knowledge-Based Systems*, vol. 216, p. 106775, 2021.

[112] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. Vincent Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.

[113] W. Xiang, K. Yu, F. Han, L. Fang, D. He, and Q.-L. Han, "Advanced manufacturing in industry 5.0: A survey of key enabling technologies and future trends," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 2, pp. 1055–1068, 2024.

[114] D. Zhang, K. Shang, Y. Zhang, and L. Feng, "Soft sensor for blast furnace temperature field based on digital twin," *IEEE Transactions on Instrumentation and Measurement*, 2024.