ENGINEERING Management 2026, 13(2): 292–334
https://doi.org/10.1007/s42524-026-5394-x

REVIEW ARTICLE

Yangyang HUANG, Yu TAN, Yuanyuan LI, Yongxiang LI, Kwok-Leung TSUI

AI for quality management: A review

© The Author(s) 2026. This article is published with open access at link.springer.com and journal.hep.com.cn

Abstract Recent advances in artificial intelligence (AI) have significantly enhanced quality management, enabling more effective handling of complex, high-dimensional, and multi-modal data. AI methods, including machine learning (ML) and deep learning (DL), have been pivotal in advancing key areas such as quality optimization, monitoring, and diagnosis. These methods have increased adaptability, efficiency, and scalability, making them particularly suitable for modern industrial applications. This review provides a comprehensive examination of AI methods in quality management, covering the integration of surrogate models, Bayesian optimization (BO), intelligent control charts, change-point detection (CPD), and interpretable quality diagnosis. The review concludes with proposed directions for future research aimed at overcoming existing challenges and enhancing the deployment of AI in real-world quality management implementation.

Keywords quality control, Bayesian optimization, statistical process control, stream of variation, causal inference

1 Introduction

Quality management has long been a central concern in industrial production because it determines product conformance, cost efficiency, and operational safety across diverse domains such as semiconductor manufacturing (Hansen et al., 1997), healthcare (Yang et al., 2020), aerospace maintenance (Eltoukhy et al., 2020), and wind energy (Carroll et al., 2016). Historically, the foundations of quality management trace back to early work in Statistical Process Control (SPC) (Shewhart, 2022; Weibull, 1951). While these classical approaches are still widely applied, the growing complexity of industrial processes calls for new methodologies.

Data-driven methods have greatly advanced quality management, especially in quality optimization, monitoring and diagnosis. First, quality optimization increasingly relies on surrogate modeling and Bayesian optimization to efficiently explore high-dimensional design spaces under limited experimental budgets (Sacks et al., 1989; Jones et al., 1998; Shahriari et al., 2016; Frazier, 2018). Second, quality monitoring focuses on the detection of process anomalies. Classical tools such as Shewhart charts, Cumulative Sum (CUSUM) control chart, and Exponentially Weighted Moving Average (EWMA) control charts are still central to identifying shifts in process behavior (Montgomery, 2020). Beyond control charts, change-point detection (CPD) provides a complementary detection mechanism for identifying process shifts (Basseville and Nikiforov, 1993; Page, 1954; Truong et al., 2020). Third, quality diagnosis translates anomalies into actionable guidance. This step is crucial for identifying the root causes of variations and supporting process recovery and adjustment (Mason et al., 1995; Jardine et al., 2006).

As industrial systems scale and integrate Industrial Internet of Things (IIoT) technologies, they generate increasingly complex data sets, characterized by higher sampling rates, complex distributions, and heterogeneous data types (including time series, images, and logs). These characteristics pose significant challenges for traditional statistical methods, which often assume that the data types are simple and the data stationarily follow well-known distributions. For example, heterogeneous data often exhibit varying structures and characteristics, making it difficult for traditional statistical methods to perform effectively, particularly when dealing with the fusion of multi-modal data and complex interrelationships between data streams. While various extensions of traditional statistical methods have been developed to address complex data sets with high dimensionality and nonlinearity, robust and scalable solutions capable of managing

Received Nov. 19, 2025; revised Feb. 28, 2026; accepted Mar. 5, 2026

Yangyang HUANG, Yu TAN, Yuanyuan LI, Yongxiang LI (✉)
Department of Industrial Engineering and Management, Shanghai Jiao Tong University , Shanghai 200240, China
E-mail: yongxiangli@sjtu.edu.cn

Kwok-Leung TSUI
Department of Industrial, Manufacturing, and Systems Engineering, The University of Texas at Arlington, Arlington TX 76019, USA

This work was supported by the National Natural Science Foundation of China (Grant Nos. 72471142 and 72101147).

nonstationarity, multi-modality, and heteroscedasticity remain a critical challenge. These limitations highlight the need for more adaptive and flexible methods, such as AI-based methods, including conventional machine learning, deep learning with complex model architectures, and emerging foundation models.

Conventional ML approaches provide foundational methods for dimension reduction, classification, regression, and clustering tasks. These techniques, such as principal component analysis (PCA), decision trees, support vector machines (SVMs), and k-nearest neighbors (k-NN), have found broad applications in quality management (Xu and Saleh, 2021). Unlike traditional statistical methods, classical ML methods are more flexible, learning directly from the data without relying on predefined distributions or strong assumptions. However, their performance typically depends on handcrafted features and substantial domain knowledge. In highly complex systems, this reliance on manual feature engineering not only constrains the use of raw (multi-modal) data but also limits the models' ability to generalize and transfer across operating conditions, product variants, and deployment environments.

With the rapid growth of data availability and computational power, DL has emerged as a transformative tool in quality management. DL models, such as Autoencoders (AEs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks, excel in their ability to model complex relationships and dependencies in the data, often uncovering patterns that traditional statistical methods often fail to capture. The ability to process raw data, such as images, sensor readings, and log files, without extensive preprocessing has made DL an end-to-end approach for tackling high-dimensional, multi-modal data sets commonly encountered in modern industrial environments.

Despite their impressive performance, DL models are often considered "black boxes" due to their lack of inherent interpretability. This presents challenges in domains like quality diagnosis, where actionable insights and transparent decision-making are crucial. However, recent advancements in explainable AI (XAI) and causal inference have the potential to address these concerns by providing methods to interpret model predictions. These methods allow for a deeper understanding of the factors driving model decisions, making deep learning models more suitable for quality diagnosis.

Looking ahead, the emergence of large pre-trained models that can be fine-tuned for specific tasks holds significant potential for quality management. These models, which are trained on vast amounts of data, have demonstrated strong generalization capabilities across various applications, including text, image, and time-series tasks. As they become more accessible, their ability to adapt to diverse industrial scenarios could enable substantial advances in quality management.

In light of these advancements, this review aims to systematically survey the applications of AI methodologies—ranging from classical ML and DL to emerging large pre-trained models—within the domain of quality management. While quality management broadly covers the entire product lifecycle, including reliability assessment and maintenance, we limit our scope to three fundamental pillars—quality optimization, monitoring, and diagnosis—which constitute the core workflow of modern quality management. Despite the growing interest and advancements in AI methods for quality management, most existing reviews focus on either classical statistical approaches (Montgomery, 2020) or specific AI applications in particular aspects of quality management, such as CPD (Xu et al., 2025) or anomaly detection (Du et al., 2025). In contrast, our work aims to offer a broader perspective by comprehensively reviewing the applications of AI in quality optimization, monitoring and diagnosis.

To identify the foundational methodologies and representative contributors in quality management, we first summarized the influential scholars and research groups in this domain. Table 1 highlights key pioneers, categorizing their primary research focus into quality optimization, monitoring, and diagnosis. This mapping provides readers with a guide to the leading communities shaping the theoretical foundations of the field.

To provide a quantitative overview of how AI is being adopted in this rapidly evolving field, we curated a list of flagship journals that serve as core publication venues for the pioneers identified in Table 1. These venues comprehensively cover the full spectrum of modern quality management. Subsequently, we conducted a targeted literature search in the Web of Science Core Collection.

Table 1 Representative research groups and key contributors in AI-driven quality management

Representative scholar	Institution	Primary research focus
C. F. Jeff Wu	Georgia Institute of Technology, USA	Experimental design, robust parameter design, computer experiments
Jianjun Shi	Georgia Institute of Technology, USA	Multistage manufacturing quality modeling, stream of variation
William H. Woodall	Virginia Tech, USA	Social network monitoring, profile monitoring, health surveillance
Peihua Qiu	University of Florida, USA	Image and video-based process monitoring, jump regression analysis
Fugee Tsung	HKUST, China	High-dimensional SPC, statistical monitoring of big data streams
William J. Welch	University of British Columbia, Canada	Computer experiments, Gaussian process modeling, screening designs

The search query retrieved articles containing AI-related keywords (e.g., "machine learning," "deep learning," "Bayesian optimization") published over the past decade (2015–2025). Figure 1 illustrates the publication volume across these journals.

Building on this perspective, this review examines how AI techniques are integrated across the main stages of quality management, forming a continuous chain for design, monitoring, and diagnosis as shown in Fig. 2. We begin with quality optimization, where surrogate modeling and Bayesian optimization serve as the computational backbone of design exploration. The discussion covers both classical Gaussian-process (GP)-based surrogates and their AI-driven extensions, such as deep GPs (DGPs), Bayesian neural networks (BNNs), and physics-informed architectures, highlighting how these models enhance efficiency in high-dimensional or multi-fidelity design spaces. Next, the review addresses quality monitoring, focusing on how ML- and DL-based control charts and advanced CPD frameworks augment traditional detection capabilities in complex data streams. Subsequently, the third part provides a dedicated examination of quality diagnosis, emphasizing how variation propagation, causal inference, and XAI empower root cause analysis to support transparent decision-making. Finally, the review concludes by discussing emerging challenges such as data imbalance, multi-modal uncertainty quantification (UQ), model transferability and interpretability, and AI-empowered decision making. These challenges delineate a clear agenda for future quality management research and practice.

The paper is structured as follows. Section 2 reviews quality optimization using surrogate modeling and Bayesian optimization. Section 3 focuses on quality monitoring, covering AI-enhanced control charts and CPD. Section 4 discusses quality diagnosis, structured around variation propagation, causal inference, and XAI. The paper closes with open challenges in Section 5 and a brief conclusion in Section 6.

2 Quality optimization

Quality optimization encompasses the systematic development of products and processes to achieve high performance while maintaining robustness under uncertainty (Liu et al., 2023; Jin, 2023). This paradigm shift from traditional deterministic engineering to modern, uncertainty-aware methodologies underscores the increasing role of proactive risk management and robust optimization in complex engineering systems (Fernández-Godino, 2023). The overall workflow of quality optimization is illustrated in Fig. 3, highlighting the central roles of surrogate modeling and BO.

In this section, we provide an overview of data-driven frameworks for modern quality optimization. We first review recent advances in GP-based surrogate modeling, covering extensions to multi-output, functional-output, qualitative-quantitative (QQ), distributional-input, and functional-input scenarios. Next, we examine the

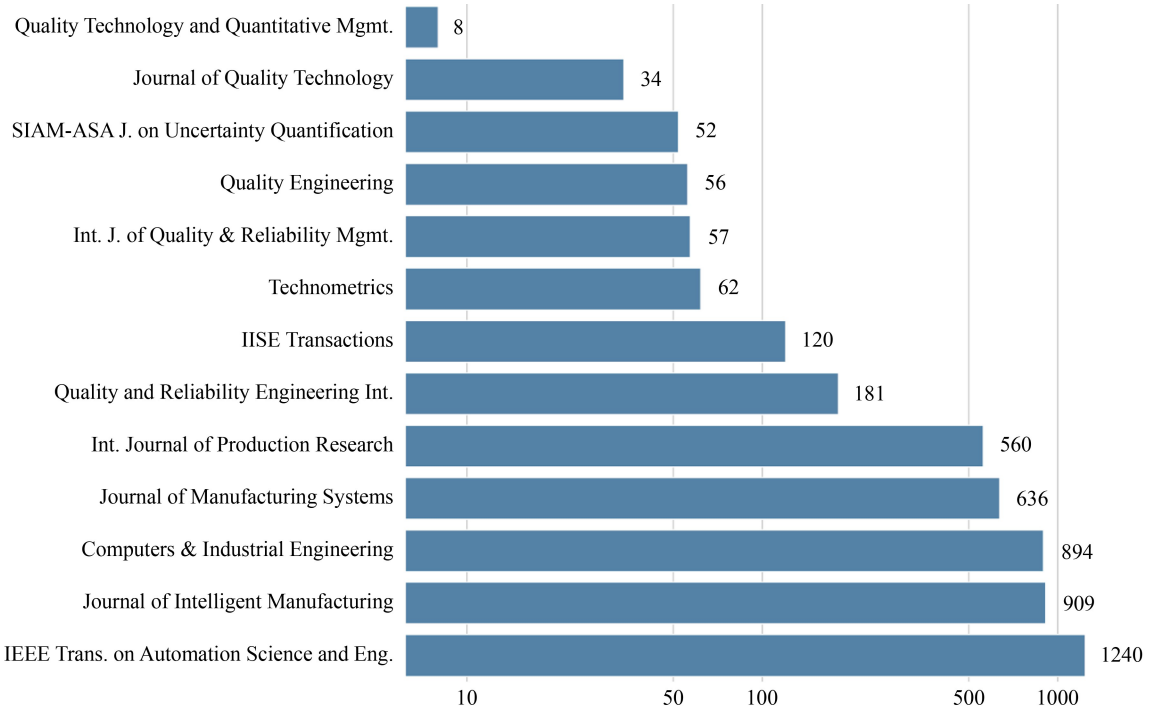

Fig. 1 Distribution of AI-related publications (2015−2025) across the flagship journals in quality management.

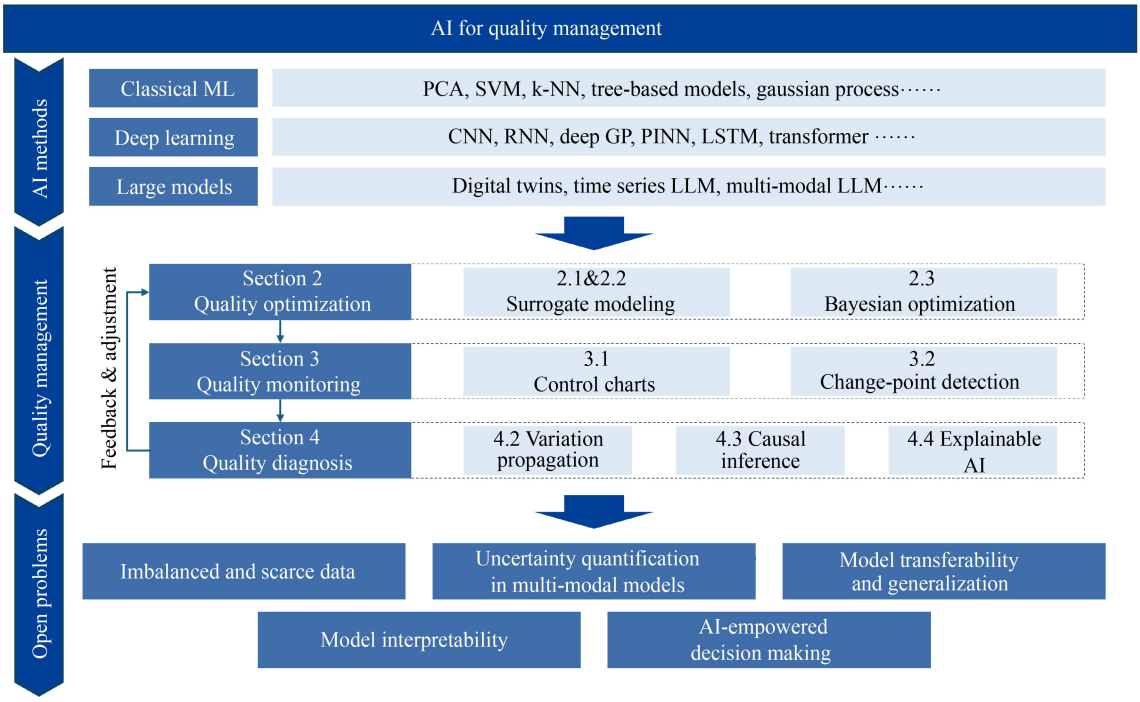

Fig. 2 Structure of AI for quality management.

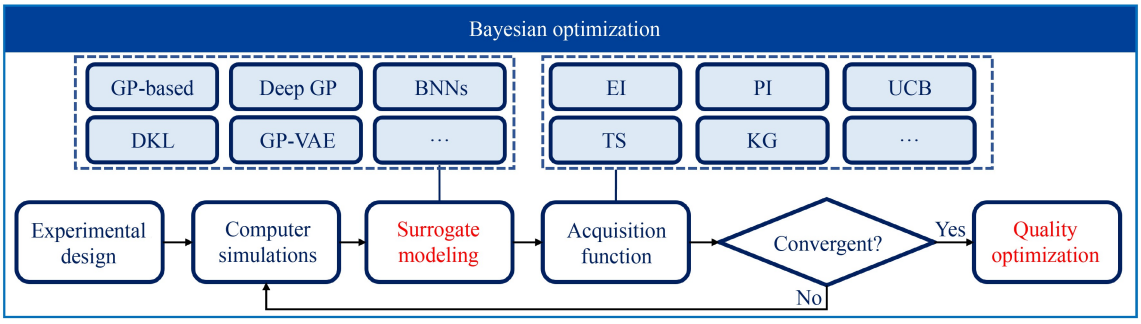

Fig. 3 Paradigm of quality optimization.

integration of AI, including deep learning and hybrid neural-probabilistic models, as scalable surrogates for complex quality-centric tasks. Finally, we survey BO strategies that leverage surrogate models for efficient and principled search for optimal quality designs under uncertainty.

2.1 GP-based surrogate modeling

High-fidelity (HF) computer simulations have become indispensable in modern quality optimization, yet the sheer computational burden can be prohibitive, especially when exploring vast design spaces or quantifying uncertainty. Furthermore, in many engineering domains, each evaluation may require not only hours of simulation but also costly and time-consuming physical experiments. GP-based methods have emerged as a key class of surrogate models for efficient optimization and rigorous UQ (Fernández-Godino, 2023; Li and Wang, 2025b; Pires et al., 2025), offering a principled way to approximate expensive simulators with acceptable loss of fidelity. In this section, we highlight recent methodological advances in GP surrogates and survey their demonstrated and prospective applications in robust product design, process reliability assessment, and intelligent manufacturing systems. Table 2 summarizes representative GP-based surrogate modeling variants and their typical input–output structures and engineering applications.

2.1.1 Single-output GPs

As a probabilistic surrogate with built-in UQ, the classical single-output GP (SOGP) remains a cornerstone for quality optimization tasks involving expensive simulations, limited data, and reliability-aware modeling (Rasmussen and Williams, 2006; Santner et al., 2018; Fernández-

Table 2 Representative GP-based surrogate modeling methods and their engineering applications

Input Type	Output Type	Literature	Application
Vector	Single	Feng et al. (2025)	3D printing
Vector	Single	Pires et al. (2025)	Structural frame
Vector	Single	Lin et al. (2025a)	Turbine rotor
Vector	Multiple	Ghosh et al. (2021)	Low-E glass manufacturing process
Vector	Multiple	Wang et al. (2023d)	Ceramic product density
Vector	Functional	Tan (2018)	Heating of a tissue containing tumor
Vector	Functional	Ma et al. (2022)	NASA's OCO-2 mission
Vector	Functional	Liu et al. (2025)	Fuselage shape control
QQ	Single	Lin et al. (2024b)	Cooling system design
Multi-fidelity	Single	Charisi et al. (2025)	Cantilever beam stress analysis
Multi-fidelity	Single	Wiegand et al. (2025)	High-pressure compressor rotor
Multi-fidelity	Multiple	Lin et al. (2021)	NACA0012 airfoil
Distributional	Single	Li et al. (2023a)	Metro passenger-flow
Distributional	Single	Da Veiga (2025)	Beam bending
Functional	Single	Tan (2019)	Mindlin plate bending
Functional	Single	Damiano et al. (2022)	Structural vibration response prediction
Functional	Single	Sung et al. (2024)	Inverse scattering problems

Godino, 2023; Li and Wang, 2025b). In robust parameter design and reliability modeling, SOGPs are frequently employed to approximate expensive limit-state functions and noisy environmental factors. Recent developments focus on embedding structural constraints and active learning to improve data efficiency. For example, incorporating symmetry constraints and variable selection has been shown to be effective for high-dimensional robustness (Feng et al., 2025). Furthermore, Kriging and denoising GP frameworks (Lin et al., 2025a; Pires et al., 2025) have demonstrated significant computational savings in estimating failure probabilities under noisy simulation outputs, reducing reliance on costly finite-element evaluations.

2.1.2 Multi-output GPs

In quality-centric engineering problems, system performance is often multi-dimensional, involving correlated or competing objectives such as cost, workload, and energy consumption (Zhang et al., 2024b). Modeling each output independently ignores inter-output correlations and can impair trade-off analysis (Lin et al., 2021). Such trade-offs are often formalized through Pareto optimality, where no objective can be improved without degrading another (Miettinen, 1999). Multi-output GPs (MOGPs) address these limitations by jointly modeling correlated outputs, thereby improving predictive accuracy and enabling principled multi-objective quality optimization (Wang et al., 2025d; Ma and Álvarez, 2023).

Traditional MOGP models often assume Gaussian observation noise across all outputs, which limits their applicability to heterogeneous data types, incomplete input domains, or partially related tasks. To address this, representative MOGP extensions introduce output-specific likelihoods, domain alignment, and regularized latent representations to enable selective information sharing across outputs (Moreno-Muñoz et al., 2018; Wang et al., 2023e). Building on these advances, MOGPs naturally accommodate transfer learning and domain adaptation. By selectively transferring information from historical processes or related tasks, methods utilizing partial domain generalization (Li and Wang, 2025a) and nonlinear input warping (Pan et al., 2025) have substantially enhanced modeling efficiency and robustness under data-scarce manufacturing regimes.

2.1.3 Functional-output GPs

With the rapid advancement of sensing technologies, quality management is increasingly confronted with functional data, where quality characteristics are recorded as high-resolution curves or surfaces rather than scalars (Liu et al., 2024a; Brunel et al., 2025). To model such high-dimensional structured outputs, functional-output GP (FOGP) models have emerged, integrating GP regression with functional data analysis (FDA) to enable uncertainty-aware prediction of entire functional responses (Shi and Choi, 2011).

A key challenge is the curse of dimensionality. While classical Functional Principal Component Analysis (FPCA) reduces dimensionality via orthogonal basis

functions (Brunel et al., 2025), recent work has explored dimension reduction strategies explicitly tailored for simulation-based outputs. Two-stage output-adapted projections leveraging linear causal modeling (Marque-Pucheu et al., 2020), as well as one-stage Latent Functional GPs (LFGP) that utilize B-spline bases with output-correlated latent variables (Liu et al., 2025), have successfully overcome traditional bottleneck limitations. For extremely high-dimensional scientific outputs, joint input-output dimension reduction combined with scalable nearest-neighbor inference has proven highly effective (Ma et al., 2022). Furthermore, to incorporate domain knowledge in physics-governed settings, physics-informed FOGP variants explicitly embed boundary conditions or analytical approximations into kernel constructions (Tan, 2018).

2.1.4 GPs for complex and mixed inputs

Traditional GP surrogates typically assume precise, continuous vector inputs. However, practical quality optimization frequently encounters categorical variables, uncertain input distributions, or functional inputs (e.g., dynamic loading curves). Extending GPs to handle these heterogeneous input spaces has been a major focus of recent research.

Quantitative-qualitative and categorical inputs. To jointly model continuous and discrete factors without inflating dimensionality via one-hot encoding, modern approaches embed categorical levels into continuous latent spaces (Zhang et al., 2020; Oune and Bostanabad, 2021). Advanced structured partitioning strategies (Lin et al., 2024b) and interpretable adjustment terms (Xiao et al., 2021) further scale these models to high-cardinality, mixed-variable experiments, overcoming the limitations of fitting separate GPs for each category.

Distributional inputs. When inputs are subject to aleatory or epistemic uncertainties, representing them as probability distributions is more rigorous. By leveraging Wasserstein geometry (Bachoc et al., 2018), regularized optimal transport (Sinkhorn) divergences (Bachoc et al., 2023), or distributional encoding (Da Veiga, 2025), recent GPs directly construct valid kernels over probability measures, facilitating robust tolerance analysis and optimization under uncertainty.

Functional inputs. For physics-based simulations driven by function-valued inputs, functional data analysis is integrated into the GP kernel. While standard methods rely on basis expansions (e.g., B-splines or PCA) (Tan, 2019; Betancourt et al., 2020), recent innovations introduce automatic dynamic relevance determination to highlight influential subdomains (Damiano et al., 2022) and define kernels directly over function spaces to avoid truncation errors (Sung et al., 2024).

2.1.5 Multi-fidelity GPs

Multi-fidelity (MF) modeling offers an effective strategy for the data scarcity challenge in quality-centric design. The core idea is to synergistically combine a small number of accurate but costly HF evaluations with a larger set of inexpensive low-fidelity (LF) data derived from simplified physics or coarser discretizations (Fernández-Godino, 2023).

A seminal framework is the recursive co-kriging model (Le Gratiet and Garnier, 2014), which established the foundation for efficiently exploiting multiple fidelity levels. Rather than enumerating individual applications, recent literature demonstrates that embedding compositional kernels (Charisi et al., 2025) and integrating adjoint sensitivities (Wiegand et al., 2025) into autoregressive MF schemes can achieve near-HF predictive accuracy while drastically reducing computational costs. Furthermore, MF-GP formulations have been seamlessly extended to jointly model multiple correlated outputs and functional responses, enabling the simultaneous exploitation of fidelity information and output structure (Lin et al., 2021; Brunel et al., 2025).

2.1.6 Discussion

Despite their flexibility and theoretical elegance, GPs suffer from severe computational limitations, primarily due to the $O(N^3)$ complexity of inverting the $N \times N$ kernel matrix. To address this, recent advances incorporate inducing-point approximations and stochastic variational inference frameworks that reduce complexity while maintaining predictive accuracy (Jiang et al., 2025). Complementary strategies such as structured kernel exploitation and distributed GP inference further enable GPs to scale to industrial-scale quality optimization tasks (Li et al., 2024e; Lin et al., 2025b).

Moreover, MOGPs rely on the assumption of meaningful correlations between outputs, but this assumption can lead to negative transfer when spurious dependencies are shared across unrelated tasks. To mitigate this, recent efforts employ output structuring techniques such as directed acyclic graph (DAG)-based conditional independence modeling (Dey et al., 2022) or latent factor regularization (Li and Kontar, 2022), allowing the model to adaptively allocate information only where justified.

In many quality optimization scenarios, domain knowledge is encoded in physical laws such as partial differential equations, conservation rules, or boundary conditions. Physics-informed GPs (PIGPs) aim to embed such constraints into the GP framework, either by designing specialized kernels or by penalizing deviations from known physics during training (Cross et al., 2024; Tan, 2018; Ding et al., 2025). While equality-type constraints are relatively tractable, incorporating complex or

inequality-based priors remains challenging (Pensoneault et al., 2020). Nonetheless, PIGPs represent a promising direction toward hybrid modeling that enhances extrapolation and physical consistency.

In summary, standard homoscedastic GP models provide a coherent probabilistic foundation for quality optimization across diverse input–output structures; however, their practical deployment must explicitly contend with scalability constraints, negative transfer, and the nontrivial incorporation of complex physical priors. This motivates the next subsection, which reviews AI-driven surrogates that trade some of GP's analytic structure for representational flexibility in modern quality-centric applications.

2.2 AI-driven surrogate modeling

Recent developments in AI, particularly deep learning, have introduced powerful alternatives to GP-based surrogates. Models such as DGPs, BNNs, physics-informed neural networks (PINNs), deep kernel learning (DKL), and GP variational autoencoders (GP-VAEs) offer greater flexibility and scalability for complex, high-dimensional quality optimization tasks. This section surveys these emerging methods and their potential in quality optimization. Representative AI-driven surrogate modeling methods and their diverse engineering applications are summarized in Table 3.

2.2.1 Deep GPs

DGPs extend the representational capacity of standard GPs by hierarchically stacking multiple GP layers, forming a deep probabilistic architecture where the output of each layer serves as the input to the next (Dunlop et al., 2018). This hierarchical composition enables the model to capture highly nonstationary behavior and abrupt transitions (Yang et al., 2025), which commonly arise in quality-critical engineering systems such as degradation processes (Toumba et al., 2024).

Beyond their hierarchical expressiveness, DGPs have been effectively combined with Bayesian active learning strategies to improve sample efficiency in computationally intensive experiments. Sauer et al. (2023b) developed an active learning framework for DGPs based on elliptical slice sampling, and showed that carefully designed acquisition functions enabled DGP surrogates to target regions of high complexity or uncertainty, outperforming stationary and treed GPs in both predictive accuracy and data efficiency. To address the computational demands of large-scale, high-dimensional computer experiments, recent advances employ inducing-point variational inference and sparse approximations such as the Vecchia scheme. For example, Sauer et al. (2023a) leveraged local conditional independence to make DGPs scalable for massive simulation data sets while maintaining reliable uncertainty quantification. In practical engineering applications, DGPs have demonstrated particular value in industrial health monitoring and degradation modeling, where systems often display abrupt changes and multi-regime behaviors. Toumba et al. (2024) applied a variational DGP framework to degradation prediction in a real-world manufacturing plant, enabling joint modeling of key dependability metrics and delivering interpretable

Table 3 Representative AI-driven surrogate modeling methods and their engineering applications

Literature	Model	Application
Carbery et al. (2018)	Bayesian network	Fault diagnosis in large-scale manufacturing (Bosch)
Li et al. (2025b)	BNN	Lunar sampling
Wu et al. (2025)	BNN	Diesel engine multi-station assembly
Zhang et al. (2022)	BNN	Online quality prediction for debutanizer column
Toumba et al. (2024)	Deep GP	Mills in semolina plant
Yazdi et al. (2024)	Deep GP	Compact object mergers (binary black holes)
Sauer et al. (2023a)	Deep GP	Satellite drag
Sauer et al. (2023b)	Deep GP	Satellite drag
Chen et al. (2022)	PINN	Bearing remaining useful life (RUL) prediction
Lu et al. (2024)	PINN	RUL prediction
Wang et al. (2024)	PINN	Lithium-ion battery SOH estimation and prognosis
Laugksch et al. (2023)	PINN	Steady-state thermofluid systems
Wang et al. (2025a)	PINN	Shield tail clearance estimation and control
Wang et al. (2022)	DKL	Traffic flow and air quality prediction
Ashman et al. (2020)	GP-VAE	Japanese weather data
Zhu et al. (2023)	GP-VAE	Robot action prediction

uncertainty estimates.

These studies collectively highlight DGPs' advantages in handling nonstationary, data-scarce, and reliability-critical quality management tasks.

2.2.2 Bayesian neural networks

BNNs represent a principled extension of conventional neural networks, wherein model parameters, such as weights and biases, are treated as probability distributions rather than fixed values. By inferring the posterior distribution over these parameters conditioned on observed data, BNNs naturally quantify predictive uncertainty and offer confidence intervals for predictions (Li et al., 2025b; Wu et al., 2025; Zhang et al., 2022). This uncertainty quantification is particularly valuable in quality management, where process variability, limited data, and multi-modal system behaviors are common.

Recent advances highlight several important applications of BNN-based surrogate models in quality management. Li et al. (2025b) developed a BNN-based surrogate model combined with active learning to enhance coring efficiency in lunar sampling. By leveraging predictive uncertainty, the proposed method reduced the number of costly experiments needed to achieve high-precision recovery. For online quality prediction in chemical and process industries, Zhang et al. (2022) proposed a BNN with efficient prior selection to mitigate epistemic uncertainty and improve model generalization under small-sample conditions. In the context of complex assembly manufacturing, Wu et al. (2025) designed an uncertainty-aware BNN framework integrated with Shapley additive explanations (SHAP) analysis for robust probabilistic prediction and feature attribution, enabling engineers to trace root causes when assembly quality deviates. Finally, Bayesian network approaches have been used for fault diagnosis in large-scale manufacturing. Carbery et al. (2018) introduced a Bayesian network learning system for modeling faults in printed circuit board (PCB) micro-drilling processes, achieving superior predictive performance compared to conventional models.

In summary, these studies illustrate that BNNs and related Bayesian methods are emerging as practical, interpretable, and uncertainty-aware surrogates for quality-centric design, online monitoring, and process diagnosis.

2.2.3 Physics-informed neural networks

PINNs integrate physical laws directly into the training of deep neural networks (DNNs), encouraging model predictions to align with established physics (Xu et al., 2023). This hybrid paradigm can enhance the robustness, interpretability, and data efficiency of surrogate modeling for quality management. PINNs are particularly advantageous when labeled data are scarce but domain knowledge is abundant, and they have shown significant promise in lifetime prediction, health monitoring, and other quality-centric tasks (Xu et al., 2023; Lu et al., 2023b; Wang et al., 2024; Al-Adly and Kripakaran, 2024).

Recent advances demonstrate the versatility of PINN-based surrogate models across a range of quality-critical engineering scenarios. For instance, Wang et al. (2024) proposed a PINN for state-of-health (SOH) estimation in lithium-ion batteries, combining empirical degradation models with neural networks to capture battery dynamics under diverse chemistries and usage protocols. Their hybrid PINN achieved high-precision SOH prediction, remained robust to small sample sizes, and supported knowledge transfer across data sets, outperforming conventional multi-layer perceptron (MLP) and CNN approaches in both accuracy and stability. In mechanical systems, Chen et al. (2022) proposed a physics-informed DNN (PDNN), termed the degradation consistent recurrent neural network for bearing prognosis, which incorporated the monotonic degradation knowledge (captured by temperature signals) into the learning process, ensuring that latent features were consistent with the physical degradation state. Their approach combined vibration and temperature signals to achieve accurate and interpretable RUL prediction in run-to-failure experiments. Similarly, Lu et al. (2024) introduced a physics-guided neural network framework for predicting the RUL of rolling bearings, leveraging LSTM networks dynamically weighted by degradation process knowledge to improve prognostic accuracy and interpretability.

Beyond component-level health prediction, PINNs have been applied to complex, integrated engineering systems. Laugksch et al. (2023) demonstrated a PINN-based surrogate modeling methodology for steady-state integrated thermofluid systems, incorporating physical constraints into the loss function to improve the accuracy and generalizability of surrogate models used in simulation and optimization. In the context of large-scale infrastructure, Wang et al. (2025a) developed an online PDNN to predict and control shield tail clearance during tunnel construction. By embedding geometric and mechanical equations into an online-updated DNN framework and coupling with multi-objective optimization, their method achieved superior predictive accuracy and actionable control for real-time engineering decision-making.

Collectively, these studies illustrate that PINNs and their variants are increasingly used as practical tools for robust, physically consistent surrogate modeling in quality management. Their ability to encode domain knowledge and enforce physical constraints directly within data-driven models can improve predictive accuracy and increase trust in model-driven decision support, especially in data-limited or safety-critical applications. Nevertheless, PINNs are known to suffer from training instability and optimization pathologies in practice. These issues often arise from loss imbalance between data-fitting and

physics-based residual terms, stiffness of the resulting optimization landscape, and sensitivity to collocation strategies and hyperparameter choices, which can lead to slow convergence or suboptimal solutions (Karniadakis et al., 2021; Wang et al., 2021; Rathore et al., 2024). As a result, careful loss weighting, adaptive sampling, or hybrid formulations are often required to achieve stable and reliable performance of PINNs.

2.2.4 Deep kernel learning

DKL is a landmark surrogate modeling framework that unites the expressive feature learning capability of DNNs with the nonparametric flexibility and GP-based UQ. Unlike traditional GPs, which are limited by fixed, shallow kernels such as the radial basis function, DKL constructs a composite kernel by feeding the output of a neural network feature extractor into a base kernel, enabling flexible, data-adaptive similarity measures (Wilson et al., 2016a). Formally, the DKL kernel is given by $k_{DKL}(\boldsymbol{x}_i, \boldsymbol{x}_j) = k_{base}(\phi_{NN}(\boldsymbol{x}_i), \phi_{NN}(\boldsymbol{x}_j))$, where ϕ_{NN} denotes a DNN transformation and k_{base} is a learnable kernel function, such as a spectral mixture kernel. All parameters, including network weights and kernel hyperparameters, are jointly optimized by maximizing the GP marginal likelihood, thus providing an integrated, data-driven approach to representation and kernel learning (Wilson et al., 2016b). It is worth noting that DKL's epistemic uncertainty quantification is typically approximate and depends on the quality of the learned neural representations.

Recent years have witnessed a surge of methodological innovations that further enhance the flexibility, robustness, and interpretability of DKL surrogates. Achituve et al. (2023) proposed Guided Deep Kernel Learning (GDKL), which leveraged Neural Network Gaussian Processes (NNGPs), the infinite-width limit of BNNs, to guide DKL training and align its uncertainty quantification. This approach substantially improved the reliability of DKL uncertainty estimates and reduced overfitting, especially in low-data regimes. Huang et al. (2023) rigorously analyzed the expressiveness of hierarchical kernels in deep kernel learning, showing that recursive composition of Gaussian and polynomial kernels expanded the associated reproducing kernel Hilbert space (RKHS), while exponential kernels did not, thereby guiding kernel choice in DKL. Wang et al. (2022) introduced physics-informed DKL (PI-DKL), which incorporated known physical laws or PDE constraints into the DKL framework. By embedding physical priors into the kernel and/or network architecture, PI-DKL achieved improved extrapolation and uncertainty quantification for scientific and engineering tasks governed by physical laws. Beyond static regression, Moss et al. (2024) extended DKL to nonlinear latent force models, allowing uncertainty-aware inference of latent dynamical processes in systems governed by complex differential equations, and broadening the role of DKL surrogates in physics-driven engineering applications.

In summary, Deep Kernel Learning has emerged as a versatile and scalable framework for constructing AI-driven surrogate models in quality-centric engineering, seamlessly combining deep representation learning and Bayesian inference. Its ongoing methodological evolution—including guided, hierarchical, and physics-informed variants—continues to expand the frontiers of robust, interpretable, and uncertainty-aware modeling for design optimization and reliability analysis.

2.2.5 GP-variational autoencoders

GP-VAEs are a class of hybrid deep generative models that combine the expressive power of neural network-based variational autoencoders (VAEs) with the principled, flexible nonparametric priors of GPs. By replacing simple parametric priors with GP priors, GP-VAEs naturally capture temporal or spatial dependencies and provide principled UQ, making them attractive for surrogate modeling under noisy or irregularly sampled conditions (Fortuin et al., 2020; Jazbec et al., 2021; Gondur et al., 2024).

Recent advances have focused on improving the scalability, expressiveness, and interpretability of GP-VAE surrogates. Ashman et al. (2020) introduced the Sparse GP-VAE (SGP-VAE), which combines multi-output sparse GP priors with variational autoencoders via partial inference networks, enabling amortized variational inference and principled treatment of missing observations in spatio-temporal data. However, as with multi-output GP-based generative models, scalability can still be challenging when the output dimensionality becomes large. To improve scalability for high-frequency time series, Zhu et al. (2023) introduced Markovian GP-VAEs, leveraging locally correlated GP priors and Kalman filtering to achieve linear-time training.

Beyond standard settings, GP-VAEs have been extended to handle heterogeneous and multi-modal data as well as to incorporate domain knowledge for improved interpretability and physical consistency. Representative developments include multi-modal GP-VAEs that disentangle shared and modality-specific latent structures, and physics-enhanced or physics-informed GP-VAE variants that embed physical constraints or generators into the latent modeling process; please refer to Gondur et al. (2024); Beckers et al. (2023); Spitieris et al. (2025) for details.

In essence, GP-VAEs have become a core tool for AI-driven surrogate modeling, providing robust, interpretable, and uncertainty-aware solutions for quality-centric engineering, scientific data analysis, and intelligent experiment design.

2.2.6 Discussion

Reliable UQ is fundamental to quality management, as it underpins trustworthy prediction, risk-aware decision-making, and sample-efficient optimization (Guth et al., 2024; Semenova et al., 2025). In this context, both aleatoric uncertainty (data-inherent variability) and epistemic uncertainty (model uncertainty due to limited data or extrapolation) should be considered, particularly in safety- or reliability-critical applications (Wang et al., 2025c; Mucsányi et al., 2024).

Classical GP models offer principled UQ through closed-form posterior distributions and remain a reference standard for epistemic uncertainty modeling (Li and Wang, 2025b; He et al., 2025). However, standard homoscedastic GPs are inherently limited in modeling input-dependent (heteroscedastic) or multi-modal aleatoric uncertainty, and their tractability does not extend easily to high-dimensional or highly nonlinear problems (Moreno-Muñoz et al., 2018; Sauer et al., 2023a; Yazdi et al., 2024). Advanced variants such as deep GPs enhance model expressiveness but introduce significant computational burdens, limiting their practical scalability (Yazdi et al., 2024).

In contrast, AI-driven surrogate models (e.g., neural networks) offer superior scalability and representational flexibility but typically rely on approximate or heuristic UQ schemes (Wang et al., 2025c; He et al., 2025). Many neural network-based approaches rely on strong distributional assumptions or approximate inference, which can underestimate uncertainty—especially in regions with little data or complex (e.g., multi-modal) noise patterns (Li et al., 2025a; He et al., 2025). To enhance UQ in AI-driven surrogate models, representative strategies have been explored, including deep ensembles for improved empirical robustness and conformal prediction for finite-sample coverage guarantees, yet they do not fully resolve the challenge of modeling conditional uncertainty or disentangling aleatoric and epistemic effects in complex design spaces (Angelopoulos and Bates, 2023; Rahaman, 2021; Lakshminarayanan et al., 2017).

A critical limitation of most AI-based surrogates lies in their treatment of uncertainty merely as pointwise predictive variance, without explicitly modeling the conditional distribution (Abdar et al., 2021; Murphy, 2022). This contrasts with GP surrogates, where posterior covariance naturally encodes data-to-data influence and underpins active learning and Bayesian optimization (Rasmussen and Williams, 2006; Cheng et al., 2025). Extending similar conditional uncertainty representations to AI-driven surrogates remains an open research challenge.

Finally, despite ongoing advances, the limited interpretability of deep surrogate models, driven by complex architectures and non-transparent learned representations, continues to hinder trust and adoption in quality management contexts (Linardatos et al., 2021; Zhang et al., 2021; Saleem et al., 2022). While physics-informed architectures and post-hoc explanation tools can partially alleviate this issue, they do not remove the fundamental non-transparency of deep parametrizations, nor do they guarantee uniquely identifiable or causally grounded explanations in complex quality optimization tasks (Linardatos et al., 2021; Saleem et al., 2022; Saranya and Subhashini, 2023).

In summary, AI-driven surrogate models substantially broaden the feasible scale and complexity of quality optimization, but remain constrained by approximate or calibration-sensitive UQ and limited transparency, which can undermine risk-aware decision-making in safety- or reliability-critical settings. In the next subsection, we review BO frameworks built on both GP-based and AI-driven surrogates, focusing on two core components, namely acquisition function design and the choice of surrogate model, and illustrating their implications for practical performance in quality optimization.

2.3 Bayesian optimization

BO is a sequential design strategy for optimizing expensive black-box functions and is highly relevant to quality management, where each experiment or simulation can be costly and time-consuming (Garnett, 2023). By iteratively updating a probabilistic surrogate to guide the selection of new evaluations, BO supports sample-efficient quality and reliability improvement under uncertainty (Paulson and Tsay, 2025); please refer to Frazier (2018) for a tutorial treatment.

2.3.1 Acquisition function

A central component of BO is the acquisition function (AF), which selects new candidate points based on the surrogate model's predictions and uncertainty (Wang et al., 2023d). The AF balances the trade-off between exploration (sampling in regions of high uncertainty to reduce epistemic risk) and exploitation (focusing on regions predicted to be optimal) (Garnett, 2023). This is particularly important in quality management, where each evaluation can be costly or high-stakes.

Classical AFs typically include expected improvement (EI), probability of improvement (PI), upper confidence bound (UCB), knowledge gradient (KG), and entropy-/information-theoretic criteria (Frazier, 2018; Garnett, 2023). To overcome the limitations of these standard AFs—such as over-greedy exploitation, poor scalability to large batches, or difficulties with high-dimensional constraints—recent literature has introduced sophisticated variants tailored for complex engineering tasks.

Several approaches structurally enhance standard criteria to better balance exploration and exploitation. For instance, Chen et al. (2024b) developed Hierarchical EI

(HEI) by embedding parameter uncertainty into a closed-form AF to prevent premature convergence, while Chen et al. (2025c) proposed Penalized EI (PEI) with a variance-weighted penalty to accelerate early-stage optimization under strict evaluation budgets. Bridging improvement-based and information-theoretic criteria, Cheng et al. (2025) introduced Variational Entropy Search (VES), a unified framework that encompasses EI as a variational inference approximation of Max-value Entropy Search. For scenarios requiring parallel evaluations or robust noise handling, Teufel et al. (2024) formulated BEEBO, a scalable batch AF grounded in statistical physics that seamlessly navigates heteroscedastic environments. Furthermore, to tackle high-dimensional spaces and strict industrial constraints, Fan et al. (2024) introduced MinUCB to bypass local gradient approximations in high-dimensional BO, and Paulson and Lu (2022) developed COBALT, which reformulates constrained expected utility into a tractable nonlinear programming problem for gray-box models.

These advances have substantially broadened the scope of BO for complex, resource-constrained, or reliability-critical engineering problems.

2.3.2 GP-based BO

Another central component of BO is the surrogate model, which provides probabilistic predictions and uncertainty estimates to guide the exploration–exploitation trade-off (Garnett, 2023). Among various surrogates, GPs remain the gold standard due to their nonparametric flexibility and closed-form UQ (Rasmussen and Williams, 2006; Frazier, 2018). In quality management, GP-based BO has enabled the efficient optimization of constrained processes (Song et al., 2024; Li et al., 2026), robust design under uncertainty (Tang et al., 2024; Astudillo and Frazier, 2021), and safety-critical systems with minimal experimental effort (Awasthi et al., 2025).

This subsection reviews recent advances in GP-based BO for quality management, organized into three settings: single-output BO for scalar quality objectives, multi-output BO for correlated attributes or tasks, and Bayesian functional optimization (BFO) for function-valued decision variables. For each setting, we highlight representative methodological developments and engineering applications. Table 4 presents a concise selection of representative GP-BO methods and AFs, together with typical engineering applications.

2.3.2.1 Single-output BO

Recent advances in BO have substantially expanded its applicability in quality management by leveraging structural information and domain constraints. For example, Astudillo and Frazier (2021) studied BO on function networks and proposed the Expected Improvement for Function Networks (EI-FN) AF, which leverages intermediate node observations to accelerate optimization and was shown to be asymptotically consistent in both synthetic benchmarks and multi-step vaccine manufacturing. Cost-aware BO has also attracted attention in settings where switching between configurations is expensive. Representative path-based strategies, such as Sequential BO via Adaptive Connecting Samples (SnAKe) and its variants, explicitly minimize transition costs while maintaining competitive optimization performance; please refer to Folch et al. (2022, 2023).

Beyond structural and cost-aware designs, GP-based BO has been extended to address practical constraints and safety-critical objectives. Contextual and constrained BO frameworks have been applied to sustainable manufacturing and autonomous system validation, enabling efficient optimization under quality, environmental, and safety requirements; please refer to Vincent et al. (2025); Awasthi et al. (2025) for representative applications.

Table 4 Overview of representative GP-based BO frameworks and engineering applications

Setting	AF	Literature	Application
Single-output	EI	Astudillo and Frazier (2021)	Multi-step vaccine manufacturing
Single-output	EI/UCB	Folch et al. (2022)	Water quality monitoring
Single-output	EI	Vincent et al. (2025)	FDM 3D printing
Single-output	TS, PI	Awasthi et al. (2025)	Autonomous driving
Single/Multi-output	EI/UCB	Folch et al. (2023)	Water quality monitoring
Multi-output	EI	Tang et al. (2024)	HDPE production
Multi-output	EI	Lin et al. (2024a)	Metamaterial vibration isolator
Multi-output	EI/UCB	Lübsen et al. (2024)	Laser synchronization (LbSync)
Multi-output	UCB	Wang et al. (2025b)	Material type determination
Multi-output	UCB	Shen et al. (2023)	Raw material properties
Functional variable	UCB	Vellanki et al. (2019)	Recirculating fiber manufacturing

2.3.2.2 Multi-output BO

Many quality management problems demand the simultaneous optimization of multiple, often correlated quality attributes or process tasks. This shift has motivated the development of multi-output BO frameworks. Representative approaches extend GP-based BO with multi-objective, multi-fidelity, or safety-aware acquisition strategies to address trade-offs, constraints, and coupling across tasks. For example, Lin et al. (2024a) proposed a constrained multi-fidelity, multi-output BO framework that integrates multi-objective acquisition design and parallel optimization, demonstrating improved efficiency on both benchmark and metamaterial vibration isolator problems.

Beyond multi-objective settings, safety and robustness considerations have driven the development of safe and adaptive multi-task BO methods. Safe multi-task BO frameworks provide probabilistic safety guarantees under task and hyperparameter uncertainty, while competitive or adaptive task-selection strategies enable efficient optimization when the primary task is unknown (Lübsen et al., 2024; Wang et al., 2025b). Multi-task BO has also been applied to simulation-based optimization and high-dimensional quality control, where task coupling and target specifications are explicitly incorporated into the surrogate or optimization strategy; please refer to Shen et al. (2023); Tang et al. (2024) for representative examples.

Despite these advances, multi-output/multi-task BO methods critically rely on the assumption of meaningful inter-task correlations. As discussed in Section 2.1.6, misspecified or weak correlations may lead to negative transfer, which can misguide AFs and degrade optimization efficiency in practice.

2.3.2.3 Bayesian functional optimization

While BO has achieved remarkable success for finite-dimensional parameters and multi-output spaces, many quality management applications require optimizing function-valued decision variables, such as time-dependent control laws, spatially varying process profiles, or entire policies and schedules. This challenge has motivated the emergence of BFO, which extends classical BO from vector spaces to infinite-dimensional function spaces. A foundational framework was introduced by Vien et al. (2018), who modeled the function domain as a RKHS endowed with a GP prior. This formulation enabled classical acquisition functions, such as EI and GP-UCB, to be generalized to functional objectives and optimized via analytic functional gradients, providing theoretical regret guarantees in infinite-dimensional settings and enabling flexible optimization without restrictive parameterizations.

Building on this foundation, Vellanki et al. (2019) proposed a functional BO framework based on adaptive Bernstein polynomial representations, which facilitates incorporating shape constraints and expert priors (e.g., monotonicity or unimodality) through dynamic basis adaptation for under-specified control laws. This line of work illustrates how domain-informed structure can substantially improve the practicality of BFO in engineering design and control. Other extensions, such as kernel functional optimization in hyper-RKHS spaces and dynamic BFO with mechanisms to handle non-stationary environments, have also been studied; please refer to Anjanapura Venkatesh et al. (2021); Bardou et al. (2024).

2.3.3 AI-driven BO

Recent advances in machine learning have greatly expanded the landscape of BO, moving beyond the classical GP surrogate to a new generation of AI-driven models (Binois and Wycoff, 2022; Li et al., 2024f; Rodemann and Augustin, 2024). These approaches leverage DNNs (Snoek et al., 2015), ensemble tree models (Lei et al., 2021), BNNs (Brunzema et al., 2024), and large language models (Chang et al., 2025) to address GP's limitations in scalability, flexibility, and handling complex or high-dimensional design spaces (Binois and Wycoff, 2022). This section reviews major methodological innovations in AI-driven BO, with an emphasis on their motivations and potential for large-scale and engineering-oriented optimization tasks.

Among AI-driven surrogates, DNNs and BNNs have received significant attention for their scalability and expressive power. Snoek et al. (2015) introduced the Deep Networks for Global Optimization (DNGO) framework, which combined deep networks for nonlinear feature extraction with Bayesian linear regression on the learned representations. This approach allowed BO to scale linearly with sample size, enabling efficient optimization in settings with large data sets or substantial parallelism, such as hyperparameter tuning of deep learning models. Recent studies have systematically benchmarked BNN-based surrogates, deep kernel learning, and related methods. Li et al. (2024f) comprehensively compared GPs, BNNs (using various inference algorithms), deep kernel learning, and deep ensembles in a range of BO tasks. Their results revealed that no single surrogate dominated across all problem settings: high-quality Bayesian inference, such as Hamiltonian Monte Carlo, yielded robust performance for fully stochastic BNNs but was computationally intensive; deep kernel learning offered a scalable alternative with competitive results; deep ensembles were less effective in sample-efficient BO; and infinite-width BNNs, which are equivalent to neural tangent kernel GPs, showed particular promise in high-dimensional regimes. These findings underscore the importance of tailoring surrogate model

choice to the problem structure and illustrate how modern neural surrogates expand the applicability of BO to large-scale and complex engineering optimization challenges. More recently, Brunzema et al. (2024) proposed a variational Bayesian last-layer (VBLL) approach, which applied variational inference only to the final layer of a neural network surrogate. This design achieved analytic Gaussian predictive distributions, significantly reduced computational cost, and enabled BO to handle high-dimensional and nonstationary objectives through continual learning and robust AFs. Empirical results demonstrated that VBLL achieves state-of-the-art performance on challenging high-dimensional and nonstationary benchmark problems.

Tree-based models, such as Bayesian Additive Regression Trees (BART) and Bayesian Multivariate Adaptive Regression Splines (BMARS), have demonstrated significant advantages in BO, particularly when handling high-dimensional and non-smooth objective functions. Lei et al. (2021) showed that BMARS and BART, unlike traditional GPs, offered greater flexibility and efficiency in managing complex, sparse data sets. BMARS and BART are both nonparametric approaches that use product spline basis functions and ensemble learning, respectively, to automatically select relevant features. These models have shown superior performance in materials discovery tasks, outperforming several GP variants in terms of search efficiency and robustness. Additionally, van Hoof and Vanschoren (2021) introduced Hyperboost, a novel approach that leveraged Gradient Boosting Decision Trees (GBDT) to enhance BO, particularly for high-dimensional or discrete design spaces. Hyperboost integrated quantile regression for uncertainty estimation and Manhattan distance-based exploration to optimize hyperparameters more efficiently. This method showed competitive performance in hyperparameter optimization tasks, particularly in larger, more complex configuration spaces. BART, BMARS, and Hyperboost highlighted the growing importance of tree-based surrogates in optimizing complex, high-dimensional problems, offering both flexibility and computational efficiency.

Recent advancements in generative models and large language models (LLMs) are introducing a new paradigm for BO, moving beyond traditional surrogates that predict outcomes to models that can intelligently propose novel candidate designs. The core philosophy of this hybrid approach is to combine the contextual reasoning strengths of LLMs for early exploration with principled statistical models for efficient exploitation. Yin et al. (2024) proposed the ADO-LLM (Analog Design Optimization) framework, which leveraged the LLM's ability to infuse domain knowledge to rapidly generate viable design points to remedy BO's inefficiency in finding high-value design areas. The framework ran a GP-BO proposer in parallel with an LLM agent that proposed candidates based on high-quality examples from a shared data set. This method was validated on real-world circuits, including a two-stage differential amplifier and a hysteresis comparator, demonstrating notable improvements in design efficiency and effectiveness compared to traditional methods. Similarly, Chang et al. (2025) developed the LLINBO (LLM-in-the-Loop BO) framework to address the risks of relying solely on LLMs for optimization. The authors proposed a hybrid framework that leveraged LLMs for early exploration while relying on statistical surrogates for exploitation, and introduced three mechanisms (Transient, Justify, and Constrained) that enable this collaboration. They further established theoretical regret bounds for the proposed mechanisms, showing that the hybrid approach controlled long-term risk. The framework was empirically validated on synthetic benchmarks, hyperparameter tuning tasks, and a 3D printing case study aimed at reducing stringing, demonstrating early lead and overall competitive performance versus traditional BO. These models collectively highlight a significant trend in which AI's creative power and domain knowledge are fused with the statistical rigor of BO, paving the way for more efficient and trustworthy engineering optimization.

Building on the foundation of GP-based multi-fidelity BO, recent research has explored more flexible and scalable surrogate models to tackle complex, real-world engineering challenges. Li et al. (2020) introduced DNN-MFBO, a multi-fidelity BO framework that used a stacked neural network architecture to capture complex, nonlinear, and nonstationary relationships between fidelity levels. This was a key methodological contribution, as it allowed the model to capture intricate inter-fidelity dependencies more effectively than traditional methods that assume simple correlation structures. To enable efficient optimization, the authors further developed a computationally tractable AF based on mutual information, approximating the objective using sequential, fidelity-wise Gauss-Hermite quadrature and moment-matching. Their approach demonstrated significant speed and cost advantages in both synthetic and real-world engineering design applications, including mechanical plate vibration design and thermal conductor design. Following this trend, Thebelt et al. (2022b) proposed using tree ensembles as scalable surrogates for multi-objective constrained optimization in energy applications. Their method was well-suited for nonlinear, multi-objective problems with heterogeneous variable spaces, which traditional GP-based methods could struggle with. The authors designed a framework that effectively handled input constraints and efficiently explored the Pareto frontier, demonstrating its advantages on challenging energy-related benchmarks like wind farm layout optimization and lithium-ion battery design.

Beyond these advancements, researchers are also extending BO to handle non-traditional constraints and alternative surrogate models. A key challenge in physical

experiments is to manage sequences of actions when each new state is dependent on the previous one, which is known as a transition constraint. To address this, Folch et al. (2024) introduced a transition-constrained BO framework based on Markov Decision Processes (MDPs). This method modeled the sequential optimization problem as an MDP and proposed a new utility function derived from hypothesis testing, which was then optimized via a Frank-Wolfe algorithm and an inner reinforcement learning (RL) subproblem. The framework was theoretically proven to converge and showed superior performance in applications, such as the Knorr pyrazole synthesis reaction, free-electron laser tuning, and automatic monitoring of Lake Ypacarai. In parallel, other efforts have focused on using non-GP models as surrogates to improve online optimization efficiency. Zhou et al. (2024) developed the BO sequential support vector regression (SVR) based on the online robust parameter design (BOSSVR-RPD) framework, which used sequential ε-SVR as the response surface for online robust parameter design. The key innovation was a nested BO loop that automatically optimized the SVR's hyperparameters using a GP prior and an EI AF with each new sample. This approach was validated on synthetic benchmarks and a real-life case study on signal quality research for color TV signal transmission, demonstrating improved accuracy and convergence speed in online scenarios where data arrives sequentially and the model must be updated with minimal computational overhead.

2.3.4 Discussion

BO has evolved into a cornerstone of modern quality management, offering an efficient and principled means to optimize expensive black-box systems under uncertainty (Garnett, 2023; Lu et al., 2023a). Methodological innovations across AF design, surrogate modeling, and hybrid AI-BO integration have greatly broadened BO's applicability, enabling successful deployment in diverse domains such as advanced manufacturing, process optimization, and safety-critical system validation (Garnett, 2023; Frazier, 2018; Astudillo and Frazier, 2021; Vincent et al., 2025; Awasthi et al., 2025). The growing adoption of AI-driven surrogates, such as DNNs, BNNs, tree ensembles, and LLMs, has further enriched the BO toolkit, enabling more flexible modeling of complex, high-dimensional, and heterogeneous design spaces (Li et al., 2024f; Binois and Wycoff, 2022; Snoek et al., 2015; Lei et al., 2021; Thebelt et al., 2022b; Chang et al., 2025; Yin et al., 2024).

A key observation is that the boundary between classical GP surrogates and AI-driven models is increasingly blurred (Semenova et al., 2025; Lu et al., 2023a). Many AI-inspired surrogate structures, such as tree ensembles (Thebelt et al., 2022a), neural network embeddings (Westermann and Evins, 2021), and deep kernels (Lu et al., 2023a), can be directly implemented within the GP framework via specialized kernel design, thereby unifying probabilistic inference with expressive model architectures. For instance, Thebelt et al. (2022a) showed that tree ensemble kernels enabled GPs to emulate the hierarchical and discrete representations typically associated with tree-based AI models, facilitating BO in mixed-feature and constraint-rich settings. This methodological convergence suggests that the distinction between "GP-based" and "AI-driven" BO is often a matter of implementation detail rather than a fundamental divide. Hybrid approaches can combine the uncertainty quantification of GPs with the expressive power of modern AI surrogates.

Robust and interpretable UQ remains central to BO's effectiveness, particularly in quality management, where risk-aware and sample-efficient decision-making is essential (Li et al., 2025a; Li and Wang, 2025b; Westermann and Evins, 2021). Classical GPs naturally provide closed-form posterior variance and enable principled acquisition optimization (Rasmussen and Williams, 2006). However, many AI-driven surrogates require additional calibration mechanisms to ensure trustworthy UQ, such as deep ensembles (Yang and Yee, 2024; Thuy and Benoit, 2025; Li et al., 2024f), variational inference (Brunzema et al., 2024), conformal prediction (Angelopoulos and Bates, 2023), or quantile regression (van Hoof and Vanschoren, 2021). As hybrid frameworks and generative models become more prevalent, developing unified and theoretically sound approaches to UQ across diverse surrogate classes remains an active and practically significant research direction.

Beyond numerical optimization, recent advances utilizing Google's PaLM (Anil et al., 2023) and large language models as optimizers (OPRO) (Yang et al., 2023) have demonstrated the potential to solve optimization problems through natural language prompting. Similarly, domestic coding models like CodeGeeX (Zheng et al., 2023) are increasingly applied to automate the generation of simulation scripts for complex experimental designs, reducing the barrier for quality engineers.

Looking forward, key challenges and opportunities include scaling BO to high-dimensional, data-intensive, and dynamic design problems; integrating domain knowledge, multi-fidelity, and multi-modal data; and achieving formal safety and robustness guarantees in complex, constrained environments. The rapid advancement of generative models and LLMs offers new possibilities for creative, context-aware candidate proposal and intelligent experimental design, but also raises critical questions regarding interpretability, theoretical guarantees, and long-term reliability (Chang et al., 2025; Yin et al., 2024). Addressing these problems will be essential for realizing the full potential of BO as a foundational tool for intelligent, trustworthy, and scalable quality management.

3 Quality monitoring

Once the optimal process parameters are established using the surrogate modeling and BO techniques discussed in Section 2, the focus shifts to maintaining operational stability. However, even optimally designed processes are subject to stochastic disturbances and environmental shifts during actual production. Consequently, quality monitoring serves as the vigilant eye of quality management, dedicated to the real-time assessment of process stability and product conformance. Its primary objective is the timely detection of deviations from the target in-control (IC) state (Montgomery, 2020; Qiu, 2013). While classical statistical methods provide the foundation for stability analysis, modern monitoring increasingly integrates data-driven and AI methods to enable intelligent detection of anomalies in high-dimensional, multi-modal, and nonlinear data streams.

With the advancement of AI methods, the modern monitoring workflow has evolved into a continuous, data-driven process. The core task is to monitor process data continuously to identify potential anomalies or distributional changes using control charts and CPD. When a significant deviation or change point is detected, the system raises an alarm to signal an abnormal state. These alarms serve as the critical input for the subsequent diagnosis phase (detailed in Section 4), laying the groundwork for any necessary remedial actions or parameter adjustments. In this section, we examine the application of AI methods to the two core components of quality monitoring: intelligent control charts and CPD.

3.1 Control charts

Control charts, pioneered by Shewhart (1930), are the primary tools for SPC. In practice, control charts are applied in two phases as shown in Fig. 4. Phase I uses in-control (IC) data to estimate process parameters and establish control limits; Phase II applies these limits for online monitoring and raises alarms when observations indicate an out-of-control (OOC) state (Montgomery, 2020; Qiu, 2013; Woodall and Montgomery, 1999). Beyond classical univariate (X-bar (Shewhart, 1930), EWMA (Hunter, 1986), CUSUM (Page, 1954), etc.) and multivariate (Hotelling's T^2 (Hotelling, 1992), Multivariate EWMA, Multivariate CUSUM, etc.) control charts, modern applications extend the framework to profile, image, and batch data. Comprehensive overviews of these charts can be found in recent surveys (Jalilibal et al., 2024; Liu et al., 2024a; Ottenstreuer et al., 2023; Ramos et al., 2021; Aykroyd et al., 2019).

Recent work enhances traditional control charts with AI. Qiu (2024) summarized three generic frameworks for integrating machine learning with control charts: one-class classification, artificial contrasts, and transparent sequential learning. These frameworks primarily address the common constraint that only IC data are available in most control chart applications. Additionally, a broader range of AI methods has been explored for SPC. In this survey, we group these methods into two families, as depicted in Fig. 4: classical machine learning (e.g., SVM, PCA, tree-based methods, and ensemble learning) and deep learning (e.g., AE/VAE, neural networks, probabilistic models, and hybrid models).

3.1.1 Classical machine learning methods for control charts

3.1.1.1 Support vector machines

SVMs are supervised classifiers trained via convex optimization that seek a separating hyperplane with maximum margin. Through the kernel trick, they implement nonlinear decision functions as linear separators in a high-dimensional feature space. In control charts, SVMs have been applied in several ways: (i) pattern recognition for identifying abnormal patterns in control charts, (ii) one-class SVM and support vector data description (SVDD) using IC data only, and (iii) SVR-based risk-adjusted EWMA control charts.

SVMs are widely used for control chart pattern recognition (CCPR), which treats Phase II monitoring as a supervised classification task that assigns control chart windows to in-control or abnormal patterns. Early studies showed that SVMs can robustly recognize trends, sudden shifts, mixtures, and cycles when variables move jointly, outperforming traditional discriminants (Cheng and Cheng, 2007). A prominent line of work treats CCPR explicitly as an imbalanced classification problem and uses weighted SVMs to improve sensitivity to rare abnormal pattern classes without compromising overall accuracy (Xanthopoulos and Razzaghi, 2014). Another practically important line targets autocorrelated processes: SVM-based online recognizers achieve consistent gains over competing baselines across multiple AR (autoregressive) structures and coefficient settings, indicating robustness when independence assumptions fail (Lin et al., 2011). Beyond single-pattern cases, concurrent patterns can be recognized by combining feature transforms (e.g., wavelets) with multiclass SVMs for real-time monitoring (Du et al., 2013). To improve robustness under distributional misspecification, hybrid schemes combine unsupervised grouping (e.g., spectral clustering) with SVM classifiers for CCPR under non-normal conditions (Lee et al., 2022). For broader, practice-oriented comparisons of SVM-based CCPR implementations, see Cuentas et al. (2017).

Another application of SVMs in control charts is one-class SVM and SVDD models, which are trained on Phase I IC data to learn the IC region and then, in Phase

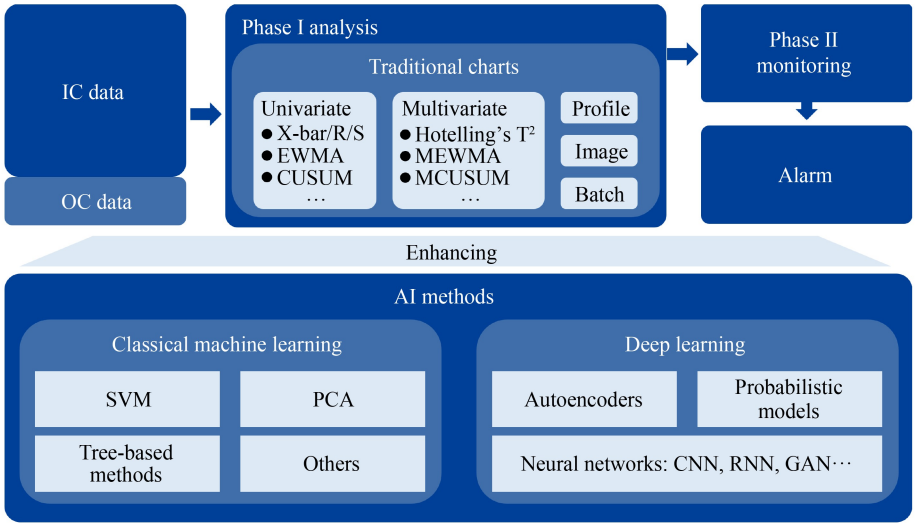

Fig. 4 AI-enhanced framework for control charts.

II, monitor a distance or score statistic to raise alarms. An early study by Sukchotrat et al. (2009) formalized this paradigm for multivariate control charts by introducing "one-class classification–based control charts" and showing how Phase I modeling of the IC data set yields practical Phase II decision rules. A closely related line of research used SVDD to define a kernel distance as the charting statistic—often called the K-chart—providing a nonparametric alternative to Hotelling's T^2 for multivariate monitoring; see Sun and Tsung (2003) for the original kernel-distance formulation and subsequent developments. Subsequent studies addressed design choices and operating characteristics. For parameter and limit selection, Weese et al. (2017) analyzed bandwidth selection for the K-chart and reported its impact on Phase I and Phase II performance. To improve robustness and adaptivity, Lee and Kim (2018) developed a time-adaptive SVDD that tracks nonstationary IC regimes and nonnormal data, thereby improving detection stability under drift. For high-frequency streams, Kakde et al. (2017) introduced the KT-chart, an SVDD-based variant that addresses the original K-chart's shortcomings at high sampling rates. Control limit construction has also been extended beyond parametric thresholds. Ahsan et al. (2023) proposed SVDD–KDE (Kernel Density Estimation) limits that estimate thresholds nonparametrically based on the SVDD statistic; simulations demonstrated performance gains across several shift scenarios and in an application to network intrusion data. In parallel, error-rate control has been made explicit: Kim and Kim (2018) designed SVDD charts with explicit false-alarm control by calibrating limits to achieve desired in-control error properties, outperforming baseline SVDD charts in experiments. Recently, SVDD scores have been hybridized with cumulative schemes (e.g., Multivariate EWMA) to strengthen small-shift sensitivity in multivariate settings (Nguyen et al., 2025).

Finally, we review SVR-based risk-adjusted EWMA charts, which follow the classical risk-adjusted EWMA logic (Steiner and Jones, 2010). Noor-ul Amin et al. (2024) modeled patient risk via SVR and applied EWMA to SVR residuals, demonstrating better small-shift detection than conventional risk-adjusted EWMA on simulations and cardiac-surgery data. To improve responsiveness, Kazmi and Noor-ul-Amin (2024) introduced an SVR-based adaptive EWMA that updates smoothing or decision parameters online as a function of predicted risk or recent residual behavior, yielding gains in average run length (ARL) performance for small location shifts; the article also discusses kernel choice and Phase I tuning.

3.1.1.2 Principal component analysis

Control charts based on PCA learn an IC subspace from historical data and then monitor principal component scores with a Hotelling's T^2 control chart, and the residual subspace with a Q (Squared Prediction Error) control chart. This two-chart paradigm and its approximate control limits are well established in the MSPC literature and industrial tutorials (e.g., MacGregor and Kourti (1995)).

For batch or trajectory data, multiway PCA (MPCA) constructs control charts directly on batch-wise time trajectories; the standard reference is Nomikos and MacGregor (1995), which defines MPCA-based Hotelling's T^2 and Q control charts with demonstrated online monitoring of new batches. To handle temporal dependence and decrease false alarms from autocorrelation, dynamic PCA (DPCA) incorporates time-lagged variables within the PCA modeling step while retaining Hotelling's T^2 and Q control charts layer. Chen and Liu (2002) presented online batch monitoring using dynamic

PCA, reporting simple charting rules and improved tracking of batch progress and upsets relative to static MPCA. Follow-on work included adaptive variants and time-slice dynamic schemes for modern high-rate batch settings (Li et al., 2024a; Du et al., 2022; Zhang and Edgar, 2007).

When quality characteristics deviate from normality or include mixed types (continuous and categorical), the usual chi-square approximations for PCA-chart limits may be unreliable. Ahsan et al. (2024) developed and evaluated PCA-mix Hotelling's T^2 control charts with kernel-density control limits, documenting improved error control and shift detection under simulated non-normal and real mixed-type scenarios. For nonlinear relations among variables, Kernel PCA (KPCA)-based control charts have been proposed and compared against linear PCA-mix; KPCA consistently improves detection for nonlinear/mixed characteristics while maintaining competitive behavior for large mean shifts in balanced categorical proportions (Ahsan et al., 2022).

3.1.1.3 Tree-based methods

Tree-based methods, including classification and regression trees (CART), random forests, and gradient-boosted trees, are commonly applied in statistical process control. They function both as classifiers for CCPR and as generators of monitoring statistics or residuals, which can then be monitored using EWMA control charts.

Decision trees have been widely used for CCPR. A representative early study applied CART to detect six canonical patterns, reporting competitive classification performance while providing transparent decision rules, thus establishing tree-based methods as practical tools for CCPR (Wang et al., 2008). More recent CCPR work addressed multiple patterns co-occurring. An online approach combined singular spectrum analysis (SSA) for feature extraction with a random forest classifier to recognize concurrent patterns in real time, showing improved recognition relative to baselines (Chiu and Tsai, 2021). A survey dedicated to concurrent CCPR synthesizes design choices and evaluation protocols across this subarea and documents the growing role of ensemble trees (García et al., 2022).

Tree ensembles have also been used to define the charting statistic itself when only IC data are available. A recent multivariate control chart built on Isolation Forest (iForest): Phase I learns the IC region; Phase II monitors the iForest anomaly score with calibrated limits, outperforming Hotelling's T^2 under complex, non-Gaussian conditions (Wang and Liu, 2024). Choi and Jung (2025) showed how data distribution and bootstrap settings affect iForest performance in SPC contexts, providing guidance for parameterization before chart construction. In practice, iForest and other one-class methods are often used to compress multivariate data into a single anomaly score (Gbashi et al., 2025). This score is subsequently charted with a conventional EWMA control chart.

Another common approach fits tree regressors in Phase I to model covariate and temporal effects; during Phase II, residuals are monitored with EWMA to identify deviations from the baseline. Alfasanah et al. (2025) fit XGBoost and monitored residuals with EWMA control charts, reporting superior detection with minimized false alarms relative to direct monitoring of the raw series; although demonstrated on environmental data, the architecture is readily transferable to industrial streams. Related studies compared tree-residual EWMA against alternative learners (e.g., SVR) within the same two-stage scheme for control charts (Rahim and Ahsan, 2025).

3.1.1.4 Others

Beyond SVMs, PCA, and tree models, several other established machine learning techniques have been incorporated into the methodology and operational workflows of control charts.

K-NN mechanisms have been used to learn the in-control region from Phase I data and to define a monitoring statistic in Phase II. Li et al. (2021) built nonparametric CUSUM schemes in which a k-NN learning step produces an empirical score that feeds an empirical cumulative sum (ECUSUM) chart. This k-NN-ECUSUM design improves sensitivity without parametric distributional assumptions and has been validated on multivariate data. Another approach is to use k-NN as a novelty/density model and construct a chart directly on k-NN-based distances or density weights. Liu et al. (2020) showed that a density-sensitive k-NN control chart can efficiently monitor multivariate processes under complex distributions, aligning it with the broader class of one-class, nonparametric charts.

Recent studies combine multiple base learners into ensembles to strengthen CCPR, with particular benefits for small shifts and for complex or concurrent pattern detection. For example, a feature-enhanced stacking ensemble integrated several classifiers for dimensional-accuracy monitoring, while a multi-scale weighted ordinal-pattern ensemble was tailored to small shifts; both reported gains over single-model baselines (Chu et al., 2024; Li et al., 2024d). Similarly, for small-variation patterns on X-bar control charts, a heterogeneous ensemble (e.g., decision tree, SVM, k-NN, ANN) achieved substantially higher recognition accuracy than individual learners and can serve directly as the Phase II decision rule (Alwan et al., 2023).

The Real-Time Contrasts (RTC) framework casts monitoring explicitly as a real-time classification problem, with the classifier's output serving as the charting statistic and decision rule. A random-forest implementation using weighted voting improved detection delay and

robustness compared with original distance-based contrast approaches (Jang et al., 2017). Later studies augmented RTC with variable-importance and novelty-detection components, and investigated adaptive break-points together with symbolic aggregation to support streaming scenarios (Lee et al., 2020; Shin et al., 2019).

3.1.2 Deep learning methods for control charts

3.1.2.1 Autoencoders

AEs learn a low-dimensional representation of IC data and provide two types of Phase II monitoring statistics: latent-space scores and reconstruction residuals. Two implementation patterns dominate. First, latent-space monitoring applies multivariate control charts to encoder outputs. Lee et al. (2019) developed a VAE-based T^2 scheme for high-dimensional, nonlinear processes and reported simultaneous reductions in false alarms and missed detections relative to latent-variable baselines. This approach can be regarded as a "deep analogue" of PCA-based monitoring, in which limit calibration is performed in the learned latent space (Lee et al., 2019; Ahmed et al., 2025).

Second, residual-driven monitoring treats reconstruction errors or latent-distance measures as univariate or scalarized statistics and monitors them with EWMA- or CUSUM-type procedures to improve sensitivity to small shifts. This workflow is particularly useful for count data or other non-Gaussian signals. For example, a recent design for Poisson multistage processes trained stacked AEs within a state-space representation and recommends EWMA on AE errors in Phase II, showing performance gains in simulation studies (Yeganeh et al., 2025). Industrial implementations similarly operationalized AE residuals or latent distances as monitoring statistics in multivariate in-process control (Biegel et al., 2022). Quality management work demonstrated that deep AEs effectively capture nonlinear variation in high-dimensional profiles, providing a practical modeling basis for either latent- or residual-based monitoring in Phase II (Howard et al., 2018).

3.1.2.2 Probabilistic models

Probabilistic formulations integrate stochastic modeling directly into monitoring by three complementary strategies: using likelihood- or predictive-distribution–based statistics (often with risk adjustment); modeling latent regimes (for example, hidden Markov or state-space models) and monitoring residuals or log-likelihood ratios; and placing a probabilistic prior over profiles or trajectories (e.g., Gaussian processes) and assessing deviations from posterior predictions.

A common probabilistic route constructs the monitoring statistic from a likelihood or posterior predictive distribution, frequently incorporating risk adjustment to account for case mix. A landmark example is the risk-adjusted CUSUM (RA-CUSUM) of Steiner et al. (2000), which scored each case by its log-likelihood contribution conditional on patient covariates. This approach reduced spurious alarms due to changing case mix and remains the standard RA chart in surgical monitoring, with extensive follow-up work on calibration and estimation error (Steiner et al., 2000; Jones and Steiner, 2012). More recent Bayesian designs exploited priors or power priors to obtain predictive control limits and favorable head-start properties when historical data are scarce (Bourazas et al., 2022). Bayesian EWMA variants—including adaptive or variable-sample-size forms that update posterior beliefs online and set limits from the posterior predictive distribution—have also been proposed to improve small-shift detection and to support dynamic sampling strategies (Khan et al., 2024). For multivariate profiles, Bayesian translations of profile CUSUM yield regression-based scores that propagate uncertainty coherently through the monitoring statistic (Ahmadi Yazdi et al., 2024).

When regime switches or serial dependence dominate the data, latent-state models are commonly paired with monitoring based on residuals or log-likelihood ratios. One approach is to fit a hidden Markov model (HMM) to Phase I IC dynamics and then monitor log-likelihood ratios or model residuals in Phase II. For Poisson HMMs, log-LR CUSUMs have been shown to provide clearer detection advantages over Shewhart or ordinary CUSUM procedures when latent states drive the counts (Ottenstreuer et al., 2021). Related implementations fit state-space or HMM predictors and chart the one-step-ahead innovations with Individuals/EWMA methods, which improves false-alarm control under strong autocorrelation (Li et al., 2022b, 2018).

For monitoring functional profiles or trajectories with within-profile correlation, GP regression supplies a flexible probabilistic prior over functions; Phase II monitoring then assesses GP predictive discrepancies or residuals. GP-based schemes naturally accommodate time-varying covariates and autocorrelated profiles, and they integrate straightforwardly with CUSUM/EWMA decision rules by using predictive means and variances to form standardized scores. Recent designs and transfer-learning variants demonstrated the practicality of GP monitoring for complex profiles and showed improved detection and UQ in applied evaluations (Ding et al., 2024; Fallahdizcheh and Wang, 2025).

3.1.2.3 Neural networks

Neural networks enhance control charts in three principal ways: as the Phase II decision rule for CCPR, as Phase I forecasters whose residuals or adjusted scores are

monitored in Phase II, and as direct statistic generators for image or spatial data. Across these roles, studies have employed a range of architectures, including artificial neural networks (ANNs), CNNs, RNNs and LSTM networks, attention-based models such as transformers, generative adversarial networks (GANs), and hybrid models (e.g., CNN–RNN or CNN–transformer). In what follows, we focus on representative uses rather than enumerating model variants.

A substantial literature has investigated neural networks as Phase II decision rules for CCPR. Recent contributions replaced hand-crafted features with 1D-CNNs and attention mechanisms, reporting higher recognition rates and better handling of concurrent or variable-length patterns (Xu et al., 2019; Hong et al., 2019; Zan et al., 2025). In these studies, the output is the Phase II decision rule, so the monitoring statistic is the network's score or class probability.

A second stream fits a neural network in Phase I to model covariates and/or dynamics; Phase II then monitors residuals or adjusted scores with EWMA to improve small-shift sensitivity and to manage autocorrelation. Representative designs reported gains for autocorrelated multivariate processes and clinical/industrial risk-adjusted settings (Ahmadini et al., 2025; Lee and Liao, 2023). Methodologically, these approaches mirror SVM/SVR residual charts but substitute neural network regressors when strong nonlinear effects would otherwise inflate false alarms.

For products and processes observed as images or spatial fields (e.g., textures, speckle patterns), CNNs are used to compute a monitoring statistic (feature-map score or learned distance) that is smoothed by EWMA or compared with calibrated limits. Recent work proposed CNN–EWMA procedures and benchmarked them against Hotelling's T^2 and generalized likelihood ratio alternatives, showing improved detection of localized and subtle changes (Okhrin et al., 2025; Sabahno and Khodadad, 2025). Here, the network furnishes a scalar (or low-dimensional) statistic per frame or region, preserving standard Phase I/Phase II practice while leveraging learned spatial features.

3.1.3 Discussion

In practice, different data regimes call for different AI integrations. When only in-control (IC) data are available, one-class and other nonparametric statistics (e.g., SVDD, Isolation Forest) provide natural Phase II monitoring scores (Sun and Tsung, 2003; Wang and Liu, 2024). Label-scarce settings motivate contrast-style formulations that turn window-level predictions into decision rules. For correlated or dynamic processes, sequential learning and residual-based monitoring help separate assignable causes from ordinary dependence.

A central cross-cutting issue is nonstationarity and serial dependence. Modern streams—ranging from networked sensors to batch trajectories and image/video frames—often violate i.i.d. assumptions underlying classical limits. Two complementary strategies are well-established. First, time-series or latent-state predictors (ARMA, state-space, HMM, etc.) are fit in Phase I, and Phase II monitoring is performed on one-step-ahead innovations, typically with Individuals/EWMA charts; this centers the decision on unexpected changes while accounting for serial correlation. Second, sequential learning methods make dependence explicit within the learning objective and pass stabilized scores to the chart. Both approaches reduce spurious alarms and yield operating characteristics that can be calibrated under correlation, extending the residual-monitoring ideas popularized in the SPC literature for autocorrelated data.

A second theme is calibration and evaluation when monitoring statistics are model-derived. Classical criteria—ARL, expected detection delay (EDD), and false-alarm probability—continue to govern design, but learned scores (latent distances, anomaly indices, posterior or predictive log-likelihoods) require principled limit setting and accounting for estimation error. Longstanding SPC guidance emphasizes that issues such as correlated observations, variable sampling schemes, and economic design materially affect performance and should be made explicit in study protocols; these points remain pertinent for AI-assisted monitoring. In multiunit deployments—many features, windows, or parallel streams—multiplicity becomes system-level: controlling the false discovery rate (FDR) offers a practical way to manage aggregate false alarms while preserving sensitivity, and recent work examines both batch and online FDR control for surveillance-type monitoring (Javanmard and Montanari, 2018).

A third, unavoidable challenge is concept drift (Gama et al., 2014). Tool wear, recipe changes, supply variation, and operator effects induce gradual evolution that may or may not be quality-relevant. While several reviewed methods include online adaptation (e.g., smoothed residuals or time-adaptive boundaries), the broader drift literature suggests complementary mechanisms—explicit tests for distributional change, adaptive/semisupervised updates, and validation protocols that distinguish benign evolution from shifts warranting intervention. Codifying such drift-aware procedures for AI-driven monitoring remains an open opportunity.

A fourth emerging theme is the interpretability and operationalization of monitoring outcomes. While traditional AI models excel at detecting signals, they often function as "black boxes," leaving operators to decipher the context of an alarm. The integration of LLMs offers a transformative solution by converting numerical anomalies into semantic insights. For instance, multi-modal models such as Google's Gemini (Team et al., 2023) show promise in directly interpreting visual control chart

patterns. Furthermore, domestic models like Alibaba's Qwen (Bai et al., 2023), DeepSeek (Guo et al., 2025), and ByteDance's Doubao (Qin et al., 2025) are being explored to automatically generate daily quality monitoring reports from heterogeneous data streams, enhancing the interpretability of SPC systems.

Finally, practice and reproducibility deserve emphasis. Many case studies report gains, but comparisons are often based on bespoke data sets with heterogeneous preprocessing and tuning. Shared benchmarks spanning univariate/multivariate, autocorrelated, image, and profile settings—coupled with transparent limit-calibration recipes—would make evaluations more comparable and accelerate adoption. Within risk-adjusted applications, for example, building the monitoring statistic from a likelihood or predictive distribution (as in risk-adjusted CUSUM) remains influential precisely because the scoring rule, covariate adjustment, and limit calibration are specified end-to-end.

3.2 Change-point detection

CPD identifies time indices at which the distribution of a data stream changes (Basseville and Nikiforov, 1993; Page, 1954; Truong et al., 2020). It is applied to signals, images, videos, and other multi-modal data from industrial sensors. Classical methods include offline segmentation via Pruned Exact Linear Time (PELT), sequential schemes such as CUSUM, and Bayesian online change detection. In-depth discussions on these methods and their applications can be found in recent comprehensive reviews on CPD (Truong et al., 2020; Aminikhanghahi and Cook, 2017; Niu et al., 2016).

AI methods enhance this pipeline under two data regimes, as shown in Fig. 5. When labels are available, each segment of the training data are assigned a class label. Supervised methods can cast CPD as a window-level classification task, learning a mapping from input segments to labels. When labels are unavailable, unsupervised methods learn the underlying structure from the data and detect distributional shifts to localize potential change points. This section reviews how AI methods implement CPD and what improvements they bring in terms of detection accuracy and robustness.

3.2.1 Unsupervised methods for CPD

3.2.1.1 Kernel-based methods

Kernel-based CPD compares distributions across candidate segments via RKHS embeddings, so that any shift in law—not just mean or variance—can be expressed as a change in a kernelized discrepancy. Early work formalized this viewpoint for single and multiple changes and connected it to dynamic-programming segmentation (Harchaoui et al., 2008). Building on that foundation, Arlot et al. (2019) introduced the KCP procedure with a nonasymptotic model-selection penalty that chooses the number of changes directly in a kernel empirical-risk framework. The same line of work established consistency and localization guarantees under characteristic kernels and mild regularity (Garreau and Arlot, 2018). In parallel, kernel two-sample testing ideas were adapted to scanning and sequential CPD: Li et al. (2015) developed the M-statistic with a tail characterization that yields calibrated thresholds (controlling significance offline and average run length online), and Li et al. (2019) proposed the Scan-B statistic that stays computationally tractable when large background data are available. A more recent theme is to learn the kernel for CPD. Chang et al. (2019) optimized a lower bound on test power using an auxiliary deep generator, effectively endowing MMD-style tests with a data-driven deep kernel while retaining the interpretability and calibration benefits of two-sample testing. These works show how kernel CPD scales from principled offline segmentation to fast, thresholded detectors, and

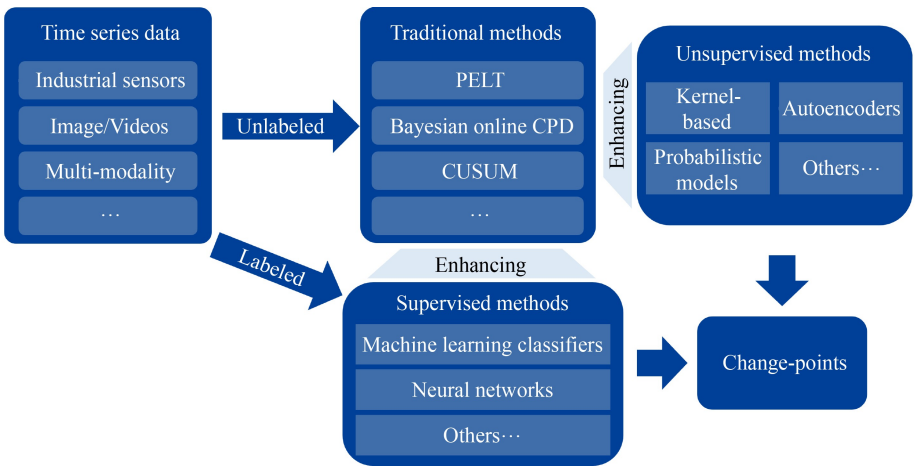

Fig. 5 AI-enhanced framework for CPD.

how kernel choice—either fixed and characteristic, or learned—controls sensitivity to high-dimensional, structured changes.

3.2.1.2 Probabilistic models

Bayesian online CPD (BOCPD) frames streaming CPD as inference over the run length, the time since the last change, so that posterior updates combine a hazard prior with predictive likelihoods under a within-segment model (Adams and MacKay, 2007). This modular construction makes it straightforward to swap in appropriate segment models while retaining a coherent online posterior over change times. Recent advances kept the spirit but pushed practicality. Knoblauch and Damoulas (2018) extended BOCPD to online model selection across a family of spatio-temporal vector autoregressions, performing prediction and Maximum A Posteriori (MAP) segmentation jointly with constant memory and linear time. Han et al. (2019) added confirmatory tests that guard against spurious alarms by validating changes in the covariance structure of a local GP surrogate. Sellier and Dellaportas (2023) replaced the GP with a reduced-rank Student-t process and dependent t noise to gain robustness to heavy tails while preserving fast approximate updates. These contributions illustrated the BOCPD template's key advantage for industrial monitoring: one can tailor the within-segment dynamics, regularize the hazard, and still reason probabilistically about both alarms and forecast performance in real time.

A complementary Bayesian route explains regime changes via latent discrete states whose transitions induce segment boundaries. In finite HMMs and switching linear dynamical systems (SLDS), change points appear when the discrete state switches, while the continuous latent dynamics account for within-segment evolution. To avoid pathological rapid switching, Fox et al. (2011) introduced the sticky HDP-HMM, a Bayesian nonparametric HMM with state persistence that has become standard for time-series segmentation with an unknown number of regimes. Modern variants tied transition probabilities to the continuous latent trajectory, yielding recurrent SLDS that segment complex nonstationary signals while revealing why a switch occurs (Linderman et al., 2017). Subsequent multi-population extensions showed scalability and stable inference in high-dimensional sensors (Glaser et al., 2020). On the sequential side, quickest detection for HMMs has seen Bayesian optimality results that quantify delay-false-alarm trade-offs when the change pertains to the hidden state process itself (Ford et al., 2023). For quality monitoring, these models are appealing when physical insight suggests few operating regimes with regime-specific dynamics (e.g., load levels or tool conditions), and when joint segmentation-and-diagnosis is valued.

GPs serve as expressive within-segment priors, with change points realized either by kernel compositions or by explicit segment boundaries. A seminal construction coupled GPs to BOCPD so that each run length indexes a locally stationary GP expert, enabling nonparametric forecasting and online detection in one pass (Saatçi et al., 2010). Recent work enhanced scalability and robustness. Caldarelli et al. (2022) proposed ADAGA, adapting kernels on the fly to track nonstationarity and detect changes with competitive accuracy. Han et al. (2019) provided hypothesis tests that confirm changes in the GP covariance, tightening thresholds for BOCPD-style alarms. In practice, GP-based CPD is attractive when one needs calibrated uncertainty for both segmentation and forecasting, and when domain knowledge can be injected as kernels (e.g., periodic or spectral mixture components) while still allowing abrupt regime switches.

3.2.1.3 Autoencoders

A popular representation-first route is to learn a compact description of nominal dynamics and convert reconstruction (or predictive) discrepancies into change evidence. In a carefully engineered study, De Ryck et al. (2021) showed that training autoencoders with a time-invariant loss tailored to CPD can sensitively capture not only mean and variance changes but also spectral and autocorrelation shifts, with a simple matched-filter postprocessing to stabilize false alarms. For real-time settings with multivariate streams, Gupta et al. (2022) used an autoencoder-based deep pipeline that outputs per-time change scores and calibrates them for streaming thresholds, reporting competitive localization on industrial-style benchmarks. Beyond plain AEs, there is growing work that brings variational structure to the latent space in order to separate regime content from nuisance variability. Large-sample supervised detectors that learn statistics from labeled examples complement these unsupervised AEs. Li et al. (2024b) showed that a small network can emulate likelihood-ratio-type test statistics and automatically tailor an offline detector to the user's change model, bridging classical tests and learned representations.

3.2.1.4 Graph models

When observations are naturally relational due to sensor correlations, line–station interactions, or user–item events, CPD becomes a question about structural shifts of a graph sequence. A classical starting point is to posit a generative network model and ask whether its parameters have changed. In evolving social networks, Peel and Clauset (2015) formalized this idea via an online Bayesian test around a hierarchical random graph: changes are declared when posterior evidence favors a new structural "norm," and the detected times align with known exogenous shocks. For correlation networks, where edges summarize time-varying dependence among

many channels, Barnett and Onnela (2016) designed a test on the Frobenius distance between correlation matrices before or after a candidate time, requiring minimal distributional assumptions and working well on fMRI and stock data. More recently, the field has moved beyond fixed, hand-chosen graph distances toward learned similarities tailored to the data domain: Sulem et al. (2024) developed a metric-learning approach that adapts graph comparisons to the detection task and showed that the choice of distance can dominate performance. At the same time, theory and inference have caught up with practice: high-quality work studied CPD in partially observed dynamic networks (Enikeeva and Klopp, 2025), multilayer random-dot-product graphs with online tests (Wang et al., 2026), and separable temporal ERGMs with group-fused penalties that localize multiple parameter shifts (Kei et al., 2025a).

When graphs carry rich node/edge attributes and structure evolves at multiple scales, graph neural networks (GNNs) can serve as task-specific representation learners whose outputs feed a CPD statistic. A recent example by Penaloza and Stevens (2024) estimated modularity via a GNN and monitors its trajectory across snapshots, flagging change points when community structure, jointly explained by topology and high-dimensional attributes shifts. The method detected a real-world event in the Iranian Twitter reply network while remaining lightweight enough for streaming. Closely related lines coupled decoder-only latent graph models with empirical-Bayes priors to expose segment boundaries in time-series of graphs, and proposed simple generative baselines for dynamic graphs that nevertheless recover clear change times (Kei et al., 2025b). These approaches are appealing in industrial monitoring where edges encode process interdependence (e.g., cross-tool correlations, line-cell flows) and where one seeks a single boundary score that already aggregates heterogeneous signals; the learned graph embedding stabilizes nuisances, while the CPD head (a calibrated discrepancy or Bayesian run-length posterior) takes care of timing.

3.2.1.5 Self-supervised learning

Self-supervision has become a natural fit for CPD, as labels for boundary times are often scarce. Deldari et al. (2021) introduced TS-CP^2, a contrastive predictive coding scheme that trains by pulling together adjacent windows and pushing apart windows that straddle a change, yielding robust embeddings and strong gains over classical unsupervised baselines. In a different but highly influential line, ClaSP framed segmentation as a classifier-based signature profile: at each candidate split, a lightweight time-series classifier is trained to discriminate the two halves; the most discriminable split is selected, and a recursive strategy yields multiple changes (Ermshaus et al., 2023). ClaSP's practical impact was reinforced by its open-source implementations and benchmarking in sktime, making it a reproducible workhorse for unsupervised segmentation (Ermshaus et al., 2023). For online CPD, contrastive ideas have been adapted to sequential scoring without labels; Puchkin and Shcherbakova (2023) cast run-time detection as maximizing a learned pre or post discrepancy with provable control of false alarms under mild conditions. Together, these works illustrated two complementary philosophies—contrastive pretraining to learn stable features for generic detectors, and classifier-at-split profiles that directly operationalize "separability" as a boundary cue—both well aligned with label-scarce industrial environments.

3.2.2 Supervised methods for CPD

3.2.2.1 Classical machine learning classifiers

Several studies cast CPD explicitly as a supervised classification or regression problem and employ classical machine learning models. A representative line of work is random-forest–based segmentation, where change points are inferred from differences in conditional distributions estimated by ensembles; for example, the ChangeForest algorithm uses random forests to detect distributional shifts in multivariate, possibly nonlinear regression settings and reports strong performance on synthetic and real benchmarks (Londschien et al., 2023). In a different direction, least-squares support vector machines (LS-SVM) were used to learn discriminative boundaries between pre- and post-change regimes, with change scores derived from the LS-SVM decision function to flag structural breaks in time series (Cheng, 2013).

Classical classifiers have also been benchmarked against statistical baselines in application domains. In nonlinear biomedical signals, comparative studies evaluated SVMs, k-NN, and tree-based ensembles as change detectors alongside CUSUM-type and time–frequency methods, highlighting settings where supervised ML can improve sensitivity or robustness to noise (Azizi, 2024). These works showed that, when labels or reliable pseudo-labels are available, classical ML classifiers offer a flexible supervised alternative that bridges classical CPD test statistics and modern representation-learning approaches.

3.2.2.2 Convolutional neural networks

CNN offers a non-recurrent route to long context with stable gradients, and has therefore been adopted in CPD when latency and parallelism matter. Rather than hand-crafting a discrepancy, one trains a 1D-CNN or TCN to produce boundary scores over time and then post-processes peaks as candidate change points. A recent study by Li et al. (2024b) provided a statistical perspective, showing that many classical test statistics can be

embedded in small neural network architectures. Building on this insight, they trained CNNs to learn the various features of multiple change types, constructing offline change detectors that match the accuracy of hand-crafted procedures while retaining transparent decision rules. In representation-first pipelines, a TCN encoder can be combined with a differentiable boundary module to handle both abrupt and gradual transitions. For example, RECURVE employed a TCN backbone and optimized a curvature-based change functional that improves robustness to slowly evolving regimes (Shin et al., 2024). These designs delivered two practical advantages in production monitoring: efficient batch or stream inference (via convolutions) and flexible receptive fields (via dilation), both of which are helpful when change scales vary across sensors.

3.2.2.3 Recurrent neural networks

Recurrent architectures remain a practical choice when change points are entangled with long temporal dependencies. A typical recipe learns short-horizon normal dynamics and then converts deviations in prediction or reconstruction to change evidence. For online monitoring, Atashgahi et al. (2023) proposed a memory-free LSTM autoencoder that continuously adapts to incoming samples without storing past windows; its run-length-agnostic update gives near-linear throughput and competitive localization across public CPD benchmarks. In a design aimed at real-time alarms on streaming data, Gupta et al. (2022) developed a deep neural CPD framework that produced a change score at each time point and triggers alarms by comparing these scores against calibrated thresholds, enabling low-latency monitoring of multivariate industrial streams. Beyond improving raw detection accuracy, recent work has begun to formalize the reliability of such neural detectors. Selective-inference procedures can be applied on top of recurrent models to test whether detected boundaries are statistically significant, thereby attaching principled uncertainty measures to the reported change points (Shiraishi et al., 2024). Together, these developments showed how RNN models trained with unsupervised or weakly supervised objectives can deliver fast, adaptive CPD while retaining interpretable decision statistics (e.g., residuals or posterior scores) for downstream root cause analysis.

3.2.2.4 Transformers

Self-attention has recently entered CPD in two ways. First, dedicated architectures inject a change-point head into a forecasting transformer so that the model jointly explains nonstationary and flags structural breaks. Wan et al. (2024) exemplified this pattern: TCDformer couples a trend/seasonality forecaster with an explicit CPD module and reports gains on nonstationary series where distributional shifts degrade pure forecasters. Second, transformers are trained end-to-end to output boundary probabilities in domain settings with many correlated channels; Kozlov et al. (2023) presented a self-supervised transformer that detects operational changes in cellular networks from multivariate key performance indicator streams, showing strong segmentation without labeled change times. These studies indicated that attention can sharpen boundary localization under long-range dependence and heterogeneous periodicities, while retaining a clean post-processing interface (peak picking or Bayesian smoothing) that is compatible with classical CPD evaluation.

3.2.3 Discussion

A pragmatic division of labor has emerged across various methods. For retrospective problems with piecewise-constant structure and moderate dependence, penalized dynamic programming remains a dependable default. Optimal partitioning or PELT selects the number of changes transparently and scales well in practice, while multiscale procedures supply uncertainty for estimated boundaries. When changes live in higher-order structure—spectral content, cross-dependence, or marginal shape—kernel segmentation and density-ratio criteria expand sensitivity beyond low-order moments yet still admit calibrated thresholds and principled search over segmentations (Arlot et al., 2019; Li et al., 2015, 2019; Kawahara and Sugiyama, 2009; Kanamori et al., 2009; Liu et al., 2013). For streaming quality control, Bayesian templates such as BOCPD, HMM, SLDS, and GP (including robust t-process surrogates) are attractive because they couple alarm timing with run-length posteriors, explicit hazards, and forecast calibration (Adams and MacKay, 2007; Fox et al., 2011; Saatçi et al., 2010; Sellier and Dellaportas, 2023). Neural sequence models—RNNs, TCNs, and transformers—add representational headroom for long-range dependence; in practice, their scores or residuals are best routed through classical smoothing and thresholding to retain interpretability and control false alarms (Atashgahi et al., 2023; Li et al., 2024b; Wan et al., 2024). In inherently relational settings, correlation-network tests and learned graph similarities expose structural breaks that remain invisible at the marginal level (Barnett and Onnela, 2016; Peel and Clauset, 2015; Sulem et al., 2024).

Several caveats recur in deployment. Calibration has outsized influence: unlike univariate CUSUM with classical ARL guarantees, data-adaptive or learned detectors often depend on resampling, carefully chosen reference segments, and fair tolerance windows; reported gains can collapse if hyperparameter budgets are unequal (Truong et al., 2020; van den Burg and Williams, 2020). Closely spaced or gradual changes are still difficult—multiscale

ideas help, but power deteriorates as spacing shrinks or drift stretches over long horizons (Fryzlewicz, 2014; Baranowski et al., 2019). In high dimensions, dependence modeling becomes the rate-limiting step: kernel scans can suffer variance inflation without structure-aware kernels or windowing, and correlation-network tests may lose power near boundaries (Barnett and Onnela, 2016; Li et al., 2015).

Looking ahead, the most promising path is not to chase a single model but to close the loop between representation, calibration, and uncertainty. Two concrete moves stand out. First, bring post-selection inference and multi-scale uncertainty to kernel and neural detectors, so boundary confidence becomes model-agnostic rather than tied to a specific estimator (Frick et al., 2014). Second, design structure-aware models for high-dimensional dependence to stabilize variance and recover detection power near boundaries (Barnett and Onnela, 2016; Peel and Clauset, 2015). Finally, time-series foundation models are likely to matter, but mostly as priors and scoring engines rather than end-to-end CPD solvers. Their probabilistic outputs should be converted to residuals or likelihood-ratio surrogates, calibrated, and then evaluated with standard change tests under shift (Das et al., 2024; Ansari et al., 2024; Rasul et al., 2023). In short, statistical calibratability, structural priors, and high-capacity sequence modeling are complementary pieces; effective CPD systems will compose them rather than pick among them.

4 Quality diagnosis

Following the detection of process anomalies in Section 3, the focus naturally moves from identifying when an abnormality occurred to understanding why it happened. The alarms generated by quality monitoring serve as the critical trigger for quality diagnosis, which is essential for translating these detection signals into process adjustments. Quality diagnosis serves as the analytical core of quality management. Its primary objective is to identify the root cause of the anomaly and support informed decision-making. Modern quality diagnosis ensures that process recovery is precise and efficient, often by triggering the parameter re-optimization methods discussed in Section 2 to restore the system to its optimal state.

Traditional diagnostic approaches often struggle with the high dimensionality and nonlinearity of modern industrial data. AI-driven methodologies have emerged to address these complexities. In this framework, AI-enabled quality diagnosis operates through three complementary methodological pillars. First, variation propagation leverages the topological structure of the process to localize root sources of variation. Second, causal inference uncovers the mechanistic drivers of faults beyond simple correlation. Third, XAI ensures that model predictions are transparent and trustworthy for human engineers. In this section, we review these advanced paradigms, highlighting how they enable precise, causal, and interpretable diagnosis.

4.1 Data-driven methods

Before exploring causal and structured diagnostic frameworks, it is essential to acknowledge the foundational role of data-driven classification in quality diagnosis. Over the past decade, supervised machine learning and deep learning models have been extensively applied to map high-dimensional process features directly to specific fault labels or defect types.

In early implementations, latent-variable and regression-based methods were widely adopted in industries. PCA and Partial Least Squares (PLS) provide low-dimensional representations that separate systematic variation from noise; contribution plots and loading patterns then link excursions in latent scores back to specific process variables. Comparative studies of data-science workflows for quality diagnosis in bioprocessing reported that PLS-based models, sometimes combined with variable-selection schemes, can localize faults more effectively than univariate SPC when many correlated covariates shift simultaneously (Borchert et al., 2019; Choi and Lee, 2005). Tree-based models are also particularly attractive because their split structures and importance measures give an initial indication of which variables are most influential. Recent overviews of automatic diagnosis in manufacturing highlighted such tree-based models, SVMs, and other classical classifiers as core components in data-driven diagnostic pipelines, often combined with feature ranking to propose candidate faults to engineers (e Oliveira et al., 2023; Gómez-Andrades et al., 2016).

Probabilistic graphical models provided another natural template for post-alarm diagnosis. In a Bayesian network (BN), nodes represent process variables and quality outcomes, and directed edges encode conditional dependencies. Once evidence from sensors is entered, belief updating ranks candidate faults by their posterior probabilities. Early case studies showed that BN-based diagnosis can integrate heterogeneous measurements for process variation analysis (Dey and Stori, 2005), and later work extended this idea by fusing expert-specified substructures with data-driven learning to handle missing sensors (Ademujimi and Prabhu, 2021). Hierarchical and object-oriented BN formulations further supported large assets by assembling reusable subgraphs (Li et al., 2022a), while dynamic Bayesian networks introduced temporal coupling, enabling joint monitoring and diagnostic reasoning in time-evolving processes (Mori and Yu, 2013; Amin and Khan, 2022).

With the proliferation of high-frequency sensor data, deep learning has driven a paradigm shift in signal-domain diagnosis. One-dimensional CNNs, temporal convolutional networks (TCNs), and transformer

encoders are now used to map vibration, acoustic, or current traces to explicit fault labels. Hou et al. (2023) introduced Diagnosisformer, an efficient transformer that fuses time-and frequency-domain cues through attention, reporting consistent gains over strong CNN baselines on bearing data sets. Subsequent supervised variants emphasized channel-attentive mechanisms to sustain diagnostic accuracy under operating-condition shifts (Liu et al., 2024d).

However, these fully supervised approaches depend on extensive labeled fault data, which are often impractical to obtain. Motivated by this limitation, contrastive self-supervision has matured into a practical route. Cui et al. (2024) optimized a self-attention encoder with a contrastive objective on unlabeled runs, improving fault attribution under variable speeds after light fine-tuning. Wang et al. (2023c) further introduced a nearest-neighbor matching strategy in the contrastive loss to mitigate negative transfer across operating conditions. When the goal is few-shot attribution to specific defect types, metric learning and prototypical networks have been reported to be effective. Zhan et al. (2022) showed that learned prototypes deliver robust fabric-defect attribution under severe class imbalance, while other cross-domain variants combine attention-based embeddings with dynamic feature selection to maintain diagnostic stability across different factories as the label budget shrinks (Jiang et al., 2023; Su et al., 2024; Xu et al., 2024).

To make these data-driven models more persuasive, recent literature has increasingly integrated physics-informed architectures and Digital Twins (DTs). Physics-informed neural networks can reduce extrapolation error by binding predictions to governing structures, such as embedding kinematic/elastic relations inside a recurrent encoder to diagnose robot joint faults (Bradley et al., 2022; Rudolph et al., 2024; Kasilingam et al., 2024; Li et al., 2023b, 2024c; Gao et al., 2023). Similarly, DTs project physics into a simulable asset whose states can be synchronized to the production line. By fusing virtual states from a gearbox or reducer twin with physical measurements, researchers have framed the twin as a source of mechanistic priors that make posterior diagnostic attribution more robust, especially in regimes where real faults are rare or hazardous to elicit (Xia et al., 2023; Liu et al., 2024c).

Despite their strong predictive performance, these predominantly data-driven classifiers face inherent limitations when scaled to complex, multi-stage manufacturing systems. They primarily map inputs to fault labels by exploiting statistical correlations, yet they often function as opaque mechanisms that may not explicitly encode the topological propagation of variations across a manufacturing line, distinguish genuine causation from spurious correlation, or provide transparent, human-auditable evidence for their predictions. Consequently, to fully realize adaptive process control and reliable parameter re-optimization, the frontier of quality diagnosis is rapidly shifting from mere pattern recognition to structured, causal, and interpretable reasoning. These advanced paradigms are detailed in the subsequent subsections.

4.2 Variation propagation

Graph structure can make post-alarm reasoning explicit. Parts, fixtures, sensors, and process stages form nodes; geometric or flow relations define edges; and variation propagates along this topology. In multistage manufacturing, the Stream-of-Variation (SoV) methodology (Shi, 2006) casts propagation in state–space form and links stage states to downstream measurements, so that a single nonconformity can be traced back through the line. Building on early state–space formulations for fixture-fault diagnosis (Ding et al., 2002) and formal diagnosability analyses that characterize when distinct upstream faults are distinguishable at the measurement layer (Zhou et al., 2003), the SoV review by Ceglarek et al. (2004) codifies path-wise reasoning and contribution analysis as first-class diagnostic tools. Recent extensions continue to calibrate SoV models from inspection data and engineering priors, improving back-tracing under plant variability and partial measurements (Moliner-Heredia et al., 2023).

From a probabilistic perspective, the same structure can be expressed as a message-passing model. Given a directed or factor-graph representation of the line, evidence from sensors and gauges is entered at observed nodes, and beliefs are updated along edges; posterior responsibility on upstream nodes induces a ranking of likely sources and a decomposition of contributions along candidate paths. This perspective is practical when topologies are incomplete or measurements are asynchronous, because unobserved nodes can be treated as latent and marginalized rather than imputed. In multistage settings, belief updates play the role of "propagation accounting," while Kalman smoothing in the SoV state–space instantiation can provide time-consistent source localization when lot mixing and transport delays blur stage boundaries (Ding et al., 2002; Ceglarek et al., 2004).

Learning-based approaches exploit the same topology with heterogeneous graph neural networks (GNNs). Instead of fixing parametric propagation, GNNs can learn how evidence aggregates over part–process–sensor graphs and assigns responsibility via attention or path scores. On aircraft final assembly lines, Chen et al. (2025a) constructed an explainable multi-layer heterogeneous graph attention network that can predict process completion time while exposing stage- and path-level importances. Although the focus is on temporal performance rather than defects, the mechanism—attention over a typed assembly graph—translates to variation attribution and station responsibility. In semiconductor

manufacturing, a recent case study demonstrates GNN-based anomaly interpretation with diagnostic visualization on fab graphs, illustrating how message passing localizes likely sources and renders human-auditable subgraphs for engineers (Ha et al., 2025). More generally, causal or intervention-aware GNNs designed for industrial diagnostics show that injecting causal priors and counterfactual training improves robustness of source localization under confounding, a recurring issue when parallel stages and shared resources induce spurious correlations (Liu et al., 2024b). These studies motivate heterogeneous node/edge types (part–process–sensor), multi-hop attention to capture long-range propagation, and path scoring to report not only where the source is but also how it influences the measured nonconformity.

4.3 Causal inference

Causal inference reframes diagnosis in terms of cause and effect rather than mere correlation. It aims to determine which process variables, stations, or settings actually cause the observed nonconformity, and how the outcome would change if those factors were adjusted (Pearl, 2009). Structural causal models (SCMs) formalize this shift by pairing a directed acyclic graph with structural equations and "do"-interventions. In practice, this identification step anchors quality diagnosis to transparent assumptions about confounding and mediators, which can then be stress-tested through sensitivity analyses.

Two complementary strands of work have made these ideas concrete for industrial diagnosis. The first focuses on discovering, from data and partial domain knowledge, graph structures that support diagnostic reasoning. Score-based continuous optimization methods such as NOTEARS recast DAG learning as a smooth constrained program and spawned more efficient variants (e.g., GOLEM), enabling scalable structure discovery in high dimensions (Zheng et al., 2018). At the same time, critiques caution against uncritical use—scale invariance and other issues can induce misleading graphs—so expert constraints and stability checks are essential when moving to the shop floor (Kaiser and Sipos, 2022). For temporally coupled processes, conditional-independence approaches tailored to autocorrelated time series (PCMCI/PCMCI$^+$) improve recall and control false positives by optimizing conditioning sets and jointly estimating lagged and contemporaneous links; these methods have become a reliable backbone for causal discovery in time-indexed industrial data (Runge et al., 2019; Runge, 2020). Recent manufacturing case studies build SCMs over process variables and quality outcomes to conduct data-driven diagnosis: Mbogu and Nicholson (2024) formulate diagnostic workflows within an SCM with time-to-event simulation to quantify causal associations, and Zhang et al. (2024a) infer causal networks for process faults, reporting improved diagnostic accuracy and robustness over correlation-based baselines.

The second strand targets the effects of actions, serving as the bridge to process adjustment. Once an admissible graph is available, modern estimators deliver average and heterogeneous treatment effects under explicitly stated assumptions. Toolkits such as DoWhy guide analysts through four stages—model, identify, estimate, and refute—making assumptions first-class citizens and offering refutation tests (placebo, bootstrap, hidden-confounder probes) that are directly useful in audit-heavy industrial settings (Sharma and Kiciman, 2020). When effect heterogeneity matters across products, tools, or operating regimes, ML-based estimators (e.g., doubly robust learners, orthogonal forests) implemented in EconML provide CATEs while mitigating regularization bias through orthogonalization and cross-fitting (Oprescu et al., 2019). A complementary perspective exploits invariance across environments (lots, recipes, machines): invariant causal prediction uses stability of conditional distributions under environment changes to identify causal parents and to construct confidence intervals, guarding against spurious correlations that vary across lines (Peters et al., 2016; Heinze-Deml et al., 2018).

Temporal propagation, which is central to manufacturing, ties these threads together. Graphical discovery on time series supplies candidate causal paths; SCM identification clarifies which adjustments are valid under confounding; and effect estimation quantifies the likely benefit of changing a setpoint or replacing a tool. When interventions are scarce, counterfactual validation can be approximated by off-policy evaluation or by leveraging naturally occurring environment shifts (e.g., recipe changes) as quasi-experiments; PCMCI-style links and invariant prediction strengthen these designs by separating stable causal parents from nuisance associations (Runge et al., 2019; Peters et al., 2016).

Evaluation mirrors the causal questions. Beyond standard classification metrics, quality diagnosis should report top-k posterior or causal responsibility (does the true cause appear in the highest-ranked set?), but also counterfactual validity: do back-door or front-door estimates agree with outcomes in held-out interventions or quasi-experiments? Ultimately, post-intervention recovery—quality restored after applying the recommended parameter re-optimization—provides the most persuasive evidence. In this spirit, Mbogu and Nicholson (2024) combine SCM-based attribution with simulated interventions to quantify expected gains, while Zhang et al. (2024a) compare causal-graph-guided diagnoses against operator-confirmed fixes. Practical concerns remain: latent confounders, sparse interventions, and evolving equipment can compromise identifiability. Here, expert constraints during discovery, refutation tests in DoWhy, and invariance-based checks across environments constitute a pragmatic guardrail that keeps causal diagnosis both interpretable and action-worthy.

4.4 Explainable AI

Over the past decade, classical machine learning and deep neural networks (DNNs) have fundamentally transformed defect recognition and sensor fusion. Architectures such as CNNs and RNNs have achieved unprecedented accuracy in modeling complex, high-dimensional industrial data. While these models excel at minimizing prediction error, they inherently lack the physical and logical transparency required for root cause reasoning. This interpretability crisis is a critical bottleneck in quality diagnosis, where engineers and operators cannot confidently execute equipment adjustments based on an unexplained alarm. Consequently, XAI has emerged as an indispensable bridge between complex neural representations and human-auditable engineering decisions.

Turning an alarm into an actionable diagnosis requires more than a label: engineers ask why the model points to a station, variable, or component, and how confident it is in that attribution. Post-hoc attribution for deep models has matured into a small toolbox. Gradient–path methods such as Integrated Gradients provide axiomatic guarantees—sensitivity and implementation invariance—while yielding input-level attributions that can be aggregated to variables, regions, or stations (Sundararajan et al., 2017). Class-activation approaches localize evidence on images without retraining; Selvaraju et al. (2017) introduced Grad-CAM to produce class-conditional heatmaps for arbitrary CNN-based architectures, and Score-CAM removes the reliance on gradients by weighting activation maps with their forward scores, often sharpening localization on industrial imagery (Wang et al., 2020). For models that must reason with parts, prototype/concept architectures make the evidence itself a first-class object: Chen et al. (2019a) designed ProtoPNet to classify by comparing input patches to learned, class-specific prototypes ("this looks like that"), and follow-on work demonstrates deformable prototypes that better match real-world variability (Donnelly et al., 2022). Concept-based testing complements these localizers: Kim et al. (2018) quantified the global influence of human-aligned concepts (e.g., solder bridges, burrs, scratches) by directional derivatives in network representation space, providing a way to check that a model's evidence aligns with engineering vocabulary.

The practical value of these tools is clearest when they are tethered to diagnostic workflows. In PCB and tire inspection, Grad-CAM overlays have been used to corroborate that detections concentrate on the physical defect zone rather than on spurious background cues, improving operator acceptance for inline deployment (Kim et al., 2021; Saleh et al., 2024). In process monitoring, CAM-based verification can gate decisions by rejecting suspicious explanations: Oh and Jeong (2020) learn the manifold of "reliable" class-activation maps with a variational autoencoder and flag out-of-manifold saliency as untrustworthy before surfacing a diagnosis, a pattern that directly addresses black-box concerns on the shop floor.

Faithfulness and stability are essential. A growing body of evidence shows that some saliency maps can be insensitive to the model or data, or that small, prediction-preserving perturbations can flip an explanation—issues that, if unchecked, risk sending engineers on fruitless hunts. Randomization tests and adversarial probes expose these failure modes: Adebayo et al. (2018) show that several visual saliency methods fail basic model/data sanity checks. Beyond method choice, data–model pathologies matter: shortcut learning can yield high apparent accuracy while explanations may rely on incidental correlates (e.g., fixture marks or lighting seams) rather than the causal signal (Geirhos et al., 2020). In industrial diagnosis, these cautions translate into practice as routine sanity checks: run attribution on permuted labels, ablate suspected nuisance features, and require that salient regions align with controllable phenomena (stations, features, specs) before authorizing corrective action.

Uncertainty quantification closes the loop from "what the model saw" to "how much we should trust it." Calibration metrics assess whether probabilities track empirical frequencies; reliability diagrams and expected calibration error (ECE) offer compact summaries, and simple post-hoc temperature scaling is often effective (Guo et al., 2017). When decisions must be accompanied by guaranteed coverage—for example, listing a small set of plausible defect types or stations to investigate—conformal prediction provides distribution-free sets with user-specified coverage, independent of the classifier internals (Angelopoulos and Bates, 2023). In quality diagnosis, this means trading set size for assurance: at a given miscoverage level α, the returned set of candidate causes contains the truth with probability at least $1-\alpha$, which is often preferable to overconfident single-point predictions.

Finally, explanations should map to interventions. Local feature weights (e.g., SHAP) can prioritize process variables or recipe settings for a trial fix, while concept-level tests verify that a model's reliance on domain concepts is aligned with engineering understanding (Ribeiro et al., 2016; Lundberg and Lee, 2017). Where a small set of actionable recommendations is needed, counterfactual explanations specify minimally sufficient changes to flip the model's decision; Wachter et al. (2018) formalize this idea without opening the black box, and diverse, feasibility-aware variants improve practicality by proposing multiple, non-redundant options (Mothilal et al., 2020). In manufacturing diagnosis, this connects explanations to the standard corrective action workflow: highlight the implicated station or variable with saliency or prototypes; quantify confidence through calibration or conformal coverage; and, when safe and feasible, propose counterfactual adjustments that trigger the re-optimization

of the process back into conformance.

4.5 Discussion

Across the literature reviewed above, three themes emerge that shape how quality diagnosis should be practiced and evaluated in modern manufacturing. First, when propagation is intrinsic in multistage lines and assemblies, models that encode the process topology and its physics provide sharper and more auditable diagnoses than flat classifiers. SoV formulations and their state–space instantiations turn downstream symptoms into upstream responsibility through estimation and smoothing over the line; identifiability analyses clarify when distinct upstream faults are distinguishable at the measurement layer (Ding et al., 2002; Ceglarek et al., 2004; Zhou et al., 2003). In time-indexed environments, dynamic Bayesian formulations serve a similar role, with posterior belief updates that remain defined even under missing or asynchronous measurements. These "structured" approaches remain the most defensible backbone for diagnostic processes whenever measurement layout, fixture geometry, and stage couplings are known.

Second, evidence and uncertainty must travel together. Supervised deep models, particularly for imagery and vibration, have made attribution fast and accurate, but without calibrated probabilities and error bars, they risk overconfident recommendations. Practical deployments that combine discriminative backbones with calibrated posteriors (reliability diagrams, ECE) and with saliency or prototype evidence see higher operator acceptance and better triage, provided that explanation quality is sanity-checked to avoid method-induced artifacts (Guo et al., 2017; Adebayo et al., 2018). Probabilistic programs for diagnosis generalize this idea: Bayesian networks encode prior engineering knowledge, fuse heterogeneous signals, and return posterior responsibility with traceable provenance (Dey and Stori, 2005; Ademujimi and Prabhu, 2021). As plants move toward hybrid data regimes—rich imagery but sparse interventions—these probabilistic layers offer a principled place to combine data with priors and to propagate uncertainty into maintenance decisions.

Third, interventions—not correlations—close the loop. Causal identification and effect estimation provide the language for actionability: which adjustment will most likely restore conformance, and with what uncertainty. Score-based and constraint-based discovery brings scale to graph construction from data, but experience on the shop floor shows that unconstrained discovery can be brittle; expert constraints and stability checks are essential (Zheng et al., 2018; Runge et al., 2019). Once an admissible graph is in hand, estimators of heterogeneous treatment effects place a quantitative value on candidate actions and support counterfactual "what-if" analyses, with refutation tests to keep conclusions auditable in regulated settings (Sharma and Kiciman, 2020; Oprescu et al., 2019). The methodological message is not to replace discriminative diagnosis with causal modeling, but to layer them: a fast, well-calibrated classifier narrows the search to candidate stations or variables; a structured propagation model (SoV/BN/DBN) apportion responsibility along paths; and a causal module quantifies which intervention at which node is most likely to succeed in triggering the necessary parameter optimization.

A recurring source of error across these methods is a mismatch between the formal model and the evolving plant. Incomplete or drifting topologies, lot mixing, and asynchronous measurements can erode both graph discovery and state–space back-tracing. Here, message passing on factor graphs and Kalman smoothing play the role of propagation accounting under uncertainty, while causal discovery tailored to autocorrelated time series helps separate stable causal parents from nuisance associations that vary across regimes (Runge et al., 2019). Where virtual assets are available, digital twins mitigate data scarcity and support "in silico" counterfactual validation, but the twin–to–real gap must be monitored and corrected; otherwise, diagnostic pipelines risk learning simulator idiosyncrasies rather than plant realities.

Two pragmatic patterns appear especially durable. The first is a hybrid stack that binds learning to structure: discriminative encoders for high-dimensional sensors feed into SoV/BN layers that enforce topology and propagate uncertainty; causal estimators translate responsibility into expected improvement under concrete actions. The second is a governed workflow: explanations undergo sanity checks; probabilities are calibrated; assumptions are explicit and refuted where possible; and every recommendation is traceable to a path, a prior, or an identified effect. These patterns align methodological rigor with the realities of production lines, where accountability and the translation of diagnostic insights into adaptive process control matter as much as raw accuracy.

5 Open challenges and future directions

In this section, we synthesize the major challenges in applying AI to quality management and outline promising directions for future research.

5.1 Imbalanced and scarce data

As more advanced AI methods are applied in quality management, their performance is increasingly constrained by the availability and quality of training data. When the training data are highly imbalanced or scarce, AI models typically perform poorly on the rare but important cases, and standard evaluation metrics can give a misleading overall high performance (He and Garcia, 2009; Krawczyk, 2016; Chen et al., 2024a).

However, this challenge may be somewhat paradoxical

in quality management, because the core objective is to keep quality stable and failures rare. As quality management matures, industrial data become dominated by long stretches of nominal operation, and faults or abrupt degradations occur rarely or are mitigated before they fully manifest. Especially in high-reliability domains such as nuclear power and aerospace industries, even single failures can be unacceptable. At the same time, deliberately generating defective products or inducing failures in live production to "enrich" the data are economically, ethically, and sometimes legally infeasible. These features make imbalanced and scarce data an especially persistent challenge for AI in quality management.

Several strategies have emerged to mitigate scarcity in industrial settings. Data-centric strategies include realistic augmentation or synthesizing additional minority-class samples to enrich fault representations or design spaces, including the use of GAN-based augmentation (Tian et al., 2024). Such approaches are promising, yet they rely critically on the fidelity of generated data and can suffer from mode collapse. Another strategy is to use transfer learning. Instead of relying solely on scarce labeled data from a single line or machine, these methods reuse knowledge learned from related assets, operating conditions, or even simulated environments to compensate for data imbalance and scarcity in the target domain (Zhao et al., 2021; Li et al., 2022c).

When labeled data are scarce or delayed, weak-supervision and semi-supervision frameworks become attractive. For example, positive-unlabeled (PU) learning treats a large stream of "unlabeled/mostly-normal" data alongside a small set of verified faults, aligning well with industrial realities (Song et al., 2025; Takahashi et al., 2024; Saunders and Freitas, 2022). However, this paradigm brings its own difficulties: estimating the true fault-class prior is challenging, and the unlabeled data set may be contaminated by undetected faults. Few- or zero-shot ideas likewise aim to generalize to unseen fault modes, but their success depends heavily on relevant auxiliary information (such as device-type metadata or physical relationships) (Shyalika et al., 2024).

Although notable progress has been made, key gaps persist: public data sets rarely mirror the true rarity and variability of industrial faults or defects, many augmentation or weak-supervision methods lack transparent evaluation or uncertainty quantification, and end-to-end studies seldom translate data imbalance into economic or safety metrics such as false-alarm cost. Further research is needed to bridge these gaps and enable robust deployment in real-world production environments.

5.2 Uncertainty quantification in multi-modal models

In industrial quality management, UQ is central to decision-making across the entire lifecycle. Traditional statistical approaches, such as hypothesis tests, Bayesian calibration, and prediction intervals, offer well-established methods for expressing confidence and risk (Kennedy and O'Hagan, 2001). These methods are most reliable when applied to single-modal data with clearly defined statistical assumptions.

However, modern systems increasingly fuse images, time series, and text/logs at scale. Classical UQ tools face challenges because the modalities involved are distributionally heterogeneous and often incomplete or misaligned. At the same time, deep learning and recent multi-modal large language models (MLLMs) enable such fusion. Yet, UQ becomes more difficult: common fusion frameworks—such as feature fusion, Mixture of Experts, and MLLM—typically lack embedded probabilistic mechanisms for propagating and combining aleatoric and epistemic uncertainty across these modalities.

Abdar et al. (2021); He et al. (2025) have systematized UQ techniques for deep learning, covering Bayesian approximations and ensembles, with a rapidly growing body of literature across vision, NLP, and other applications. Domain-specific applications are now beginning to explicitly target multi-modal fusion uncertainty. For example, recent work in autonomous perception quantified feature-level epistemic uncertainty and adaptively reweights modalities to reduce uncertainty propagation during fusion, achieving improvements under sensor degradations while maintaining competitive compute budgets (Chen et al., 2025b). GPs provide a solid framework for uncertainty quantification. Deep GPs stack multiple GP layers to propagate uncertainty through latent mappings, offering a principled, model-embedded Bayesian approach to UQ. The main challenges remain training cost and mixed robustness under distribution shifts (Damianou, 2015; Salimbeni et al., 2019). Conformal prediction has also been extended to multi-modal settings, offering distribution-free coverage that complements model-embedded UQ (Vishwakarma and Rezaei, 2024).

Despite recent advances, there is still no widely accepted, model-embedded probabilistic fusion framework for multi-modal UQ, nor standardized industrial benchmarks and decision-centric metrics. Future research should focus on propagating modality-specific uncertainty during fusion, ensuring validity under shifts and modality dropout, and evaluating methods on multi-modal quality management data sets with coverage and calibration criteria, alongside cost-aware utilities.

5.3 Model transferability and generalization

Models that perform well on one product, machine, or line often degrade when moved to another. Differences in operating regimes, sensor layouts, materials or recipes, maintenance practices, and equipment age induce data set shift. Surrogate models for quality design may extrapolate poorly to new design spaces or machines; SPC detectors

may miss shifts under altered noise spectra.

A growing body of work has explored transfer learning, domain adaptation, and domain generalization in industrial settings. For example, Chen et al. (2019b) proposed a domain-adaptation extreme learning machine on real-world image and text data sets. Zhao et al. (2021) provided a comprehensive survey of how transfer methods are used in diagnostics and prognostics tasks, documenting persistent gaps in generalization across machines and operating contexts. More broadly, surveys of transfer learning highlighted that while many algorithms align shared domains, they struggle with unknown classes in the target domain and large domain shifts (Zhuang et al., 2021).

Generalization at scale is increasingly linked to pretraining on heterogeneous data. Recent "foundation" approaches for time series (large pretrained transformers) reported strong zero- or few-shot transfer across domains, yet open challenges remain around compute cost, calibration, and domain relevance after fine-tuning (Liang et al., 2024). In parallel, physics-guided learning offers a complementary route to transfer by anchoring representations to site-invariant structure. Reviews and analyses of PINNs and physics-informed ML discussed both benefits and known failure modes, underscoring the need for careful numerical treatment in industrial settings (Karniadakis et al., 2021; Krishnapriyan et al., 2021).

Future work may focus on two priorities: (i) practical, reusable frameworks for model transfer that fit industrial quality management data, and (ii) generally pre-trained models that can be vertically tuned to specific machines, products, or duty cycles. Concretely, this points to standard protocols and toolchains for data set-shift monitoring and adaptation, in conjunction with pretraining or fine-tuning regimes to keep models calibrated and reliable at deployment.

5.4 Model interpretability

Increasingly complex AI systems (e.g. deep neural networks and large ensemble models) have delivered substantial gains in prediction accuracy and fault detection performance across quality management applications. Yet, when these "black-box" models drive decisions about product quality, maintenance timing or process control, the lack of transparency becomes a critical barrier to adoption. For quality management deployments, interpretability and explainability are now widely regarded as prerequisites for trust, governance, and effective integration of AI outputs.

Foundational techniques now widely used include local, model-agnostic explanations such as LIME and SHAP (Ribeiro et al., 2016; Lundberg and Lee, 2017). For deep networks, gradient-based attribution methods such as Integrated Gradients and class-discriminative saliency methods such as Grad-CAM assign contribution scores to individual input features (Sundararajan et al., 2017; Selvaraju et al., 2017). In parallel, concept-based methods link internal representations to human-interpretable concepts that match domain terminology, making explanations easier to operationalize (Kim et al., 2018). In industrial settings, case studies reported SHAP-guided process tuning and action prioritization in semiconductor fabrication and explainable predictive maintenance in rotating machinery—each showing that attribution summaries can be translated into actionable rules (Senoner et al., 2022; Gawde et al., 2024).

Despite the progress, interpretability research faces systemic challenges in real-world settings (Lipton, 2018; Slack et al., 2020). First, many explanation methods are developed in a laboratory environment and emphasize local or post-hoc interpretability (e.g., feature importance) without verifying how well they generalize to the full model behavior. Second, explanations that are technically valid may nonetheless fail to align with domain-expert reasoning or operational workflows. Third, evaluation of interpretability itself remains underdeveloped. Many studies report anecdotal examples rather than systematic metrics of explanation quality, consistency, or impact on decision outcomes.

Beyond quality management, the interpretability of AI itself remains a fundamentally hard problem. There is no universally accepted definition, measurement protocol, or technique that yields theoretically complete and incontrovertible explanations across tasks and models. Consequently, the engineering use of AI in quality management should be anchored in governance and risk-management practice rather than a promise of perfect interpretability. Practitioners should define explicit scope and limitations, assign accountability, and maintain traceable documentation.

5.5 AI-empowered decision making

While most existing AI for quality management studies emphasize data analytics (e.g., building predictive or diagnostic models), a central objective of quality management is decision-making under uncertainty. Despite advances in machine learning and uncertainty quantification, a persistent gap remains between AI outputs and action (Soori et al., 2026; Kovari, 2024). A key open challenge is to develop principled decision-making frameworks that leverage AI-driven predictive and diagnostic models to support decision-making for quality improvement of engineering systems.

Beyond this, most existing AI for quality management methods are still developed and deployed in silos: surrogate modeling and Bayesian optimization are used for offline quality design, SPC and change detection are applied to online process monitoring, and diagnostic models drive downstream root cause analysis. In practice, however, these stages often operate on the same physical assets and

share common data streams, constraints, and risk preferences. This fragmentation is particularly problematic in complex, multi-stage manufacturing systems, where deviations in upstream stages can propagate and amplify downstream, rendering passive monitoring insufficient. To achieve true adaptive process control, monitoring and diagnosis must be tightly coupled with parameter re-optimization. Many quality-critical decisions are inherently sequential—for example, dynamic process adjustments and compensatory control—where current actions affect future system states, information, and costs. By formulating these adjustments as a Markov decision process or a stochastic control problem rather than a one-shot optimization, advanced AI agents can leverage real-time monitoring data to dynamically adjust downstream process parameters. This closed-loop approach shifts the paradigm from merely signaling alarms to actively compensating for propagated deviations in real-time.

A promising direction is to move toward unified, decision-centric frameworks, such as digital twins, that consistently integrate high-fidelity physics models, data-driven surrogates, real-time sensing, and uncertainty-aware decision-making modules (Zhong et al., 2023; Mayer et al., 2025). Realizing such unified, decision-centric digital twins raises several open challenges. First, digital twin architectures for quality management must support multi-scale and multi-fidelity modeling, allowing models in different quality management stages to interact through shared state representations and uncertainty quantification (Ouedraogo et al., 2025). Second, the framework should embed decision-making components—for example, BO for design updates and sequential decision rules for process adjustment and inspection—so that learning and acting are co-designed rather than separated. Third, lifecycle digital twins must remain trustworthy and maintainable over time: models need to be updated under distribution shift, validated against physical constraints, and made interpretable enough for engineers to audit and override recommended actions. Developing such integrated, AI-enabled digital twins for quality management, together with decision-centric benchmarks and evaluation metrics, is a key opportunity for future research to turn predictive intelligence into tangible, system-level quality gains.

In practice, engineers must balance multiple, often conflicting objectives—such as performance, reliability, cost, and safety—within dynamic, uncertain environments. These decisions are further constrained by engineering limits (e.g., capacity, regulatory requirements, maintenance windows) and soft preferences (e.g., risk attitudes, cost–performance trade-offs, tolerances for false alarms versus missed detections), turning many quality management problems into intrinsically multi-objective and constrained optimization tasks (Shan et al., 2025; Shahtaheri et al., 2019; Wang et al., 2023b). Moreover, in most industrial settings, engineers retain ultimate responsibility for actions, so decision-support systems must not only recommend policies but also provide interpretable rationales and transparent trade-offs, enabling human experts to calibrate, override, or adapt AI-generated decisions.

6 Conclusions

This review provides a holistic synthesis of how AI methods are reshaping quality management, covering topics ranging from quality optimization, monitoring, and diagnosis. These advances have allowed for more accurate, efficient, and adaptable solutions in modern industrial quality management.

Across these domains, a common shift is evident: quality management practice is moving from assumption-based statistical methods toward adaptive data-driven AI modeling. At the same time, new challenges emerge as industrial systems and AI models grow more complex. On the data side, practitioners must contend with high-dimensional, multi-modal, heterogeneous, and highly imbalanced data. On the modeling side, they must ensure generalization, quantify uncertainty, and provide explanations that support engineering trust and decision-making in practice. These issues cannot be solved by simply tuning a single model because they arise at the system level of modern quality management, encompassing data collection, modeling, and deployment. Therefore, new frameworks are required to integrate statistical rigor, physical insight, and AI techniques.

Looking ahead, how to construct operationally deployable AI-integrated systems that span the full industrial lifecycle will likely define the next decade of quality management research. The key opportunity lies in developing unified, uncertainty-aware frameworks that are scientifically interpretable. Real progress will depend on closer collaboration among statisticians, engineers, and AI researchers, supported by open industrial data sets, benchmarks, and reproducible platforms.

Ultimately, the goal is not to simply replace classical quality management methods with AI but to augment them by embedding data-driven intelligence and adaptability within these time-tested principles. We hope this review serves as both a conceptual road map and a call to action for developing scalable, intelligent, and reliable quality management for the next generation of industrial engineering.

Acknowledgment The authors sincerely thank the editors and the reviewers for their constructive comments and insightful suggestions, which significantly improved the quality of this paper.

Competing Interests The authors declare that they have no competing interests.

Open Access This article is licensed under a Creative Commons Attribution

4.0 International License, which permits use, sharing, adaptation, distributionand reproduction in any medium or format, as long as you give appropriatecredit to the original author(s) and the source, provide a link to the CreativeCommons licence, and indicate if changes were made.The images or other third party material in this article are included in thearticle's Creative Commons licence, unless indicated otherwise in a creditline to the material. If material is not included in the article's CreativeCommons licence and your intended use is not permitted by statutory regu-lation or exceeds the permitted use, you will need to obtain permissiondirectly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, Fieguth P, Cao X, Khosravi A, Acharya U R, Makarenkov V, Nahavandi S (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, 76: 243–297

Achituve I, Chechik G, Fetaya E (2023). Guided deep kernel learning. In: Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence, PMLR, 216: 11–21

Adams R P, MacKay D J C (2007). Bayesian online changepoint detection. Preprint at arXiv. arXiv:0710.3742

Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B (2018). Sanity checks for saliency maps. In: Proceedings of Advances in neural information processing systems, Curran Associates, Inc., 31: 1–11

Ademujimi T, Prabhu V (2021). Fusion-Learning of Bayesian network models for fault diagnostics. Sensors (Basel), 21(22): 7633

Ahmadi Yazdi A, Shafiee Kamalabad M, Oberski D L, Grzegorczyk M (2024). Bayesian multivariate control charts for multivariate profiles monitoring. Quality Technology & Quantitative Management, 21(3): 386–421

Ahmadini A A, Abbas T, AlQadi H (2025). Adaptive EWMA control chart by adjusting the risk factors through artificial neural network. Quality and Reliability Engineering International, 41(3): 992–1001

Ahmed F, Mahmood T, Riaz M, Abbas N (2025). Comprehensive review of high-dimensional monitoring methods: trends, insights, and interconnections. Quality Technology & Quantitative Management, 22(4): 727–751

Ahsan M, Khusna H, Wibawati, Lee M H (2023). Support vector data description with kernel density estimation (SVDD-KDE) control chart for network intrusion monitoring. Scientific Reports, 13(1): 19149

Ahsan M, Mashuri M, Khusna H, Wibawati (2022). Kernel principal component analysis (PCA) control chart for monitoring mixed nonlinear variable and attribute quality characteristics. Heliyon, 8(6): e09590

Ahsan M, Mashuri M, Prastyo D D, Lee M H (2024). Performance of T2-based PCA mix control chart with KDE control limit for monitoring variable and attribute characteristics. Scientific Reports, 14(1): 7372

Al-Adly A I, Kripakaran P (2024). Physics-informed neural networks for structural health monitoring: a case study for Kirchhoff–Love plates. Data-Centric Engineering, 5: e6

Alfasanah Z, Niam M Z H, Wardiani S, Ahsan M, Lee M H (2025). Monitoring air quality index with EWMA and individual charts using XGBoost and SVR residuals. MethodsX, 14: 103107

Alwan W, Ngadiman N H A, Hassan A, Saufi S R, Mahmood S (2023). Ensemble classifier for recognition of small variation in x-bar control chart patterns. Machines, 11(1): 115

Amin M T, Khan F (2022). Dynamic process safety assessment using adaptive Bayesian network with loss function. Industrial & Engineering Chemistry Research, 61(45): 16799–16814

Aminikhanghahi S, Cook D J (2017). A survey of methods for time series change point detection. Knowledge and Information Systems, 51(2): 339–367

Angelopoulos A N, Bates S (2023). Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning, 16(4): 494–591

Anil R, Dai A M, Firat O, Johnson M, Lepikhin D, Passos A, Shakeri S, Taropa E, Bailey P, Chen Z, et al (2023). PaLM 2 Technical Report. Preprint at arXiv. arXiv:2305.10403

Anjanapura Venkatesh A K, Shilton A, Rana S, Gupta S, Venkatesh S (2021). Kernel functional optimisation. In: Proceedings of Advances in Neural Information Processing Systems, Curran Associates, Inc., 34: 4725–4737

Ansari A F, Stella L, Turkmen A C, Zhang X, Mercado P, Shen H, Shchur O, Rangapuram S S, Arango S P, Kapoor S, Zschiegner J, Maddix D C, Wang H, Mahoney M W, Torkkola K, Wilson A G, Bohlke-Schneider M, Wang B (2024). Chronos: learning the language of time series. Transactions on Machine Learning Research, 2024: 1–42

Arlot S, Celisse A, Harchaoui Z (2019). A kernel multiple change-point algorithm via model selection. Journal of Machine Learning Research, 20: 1–56

Ashman M, So J, Tebbutt W, Fortuin V, Pearce M, Turner R E (2020). Sparse gaussian process variational autoencoders. Preprint at arXiv. arXiv:2010.10177

Astudillo R, Frazier P (2021). Bayesian optimization of function networks. In: Proceedings of Advances in Neural Information Processing Systems, Curran Associates, Inc., 34: 14463–14475

Atashgahi Z, Mocanu D C, Veldhuis R, Pechenizkiy M (2023). Memory-free online change-point detection: a novel neural network approach. Preprint at arXiv. arXiv:2207.03932

Awasthi S S, Imran M I I S, Arrigoni S, Braghin F (2025). Bayesian optimization applied for accelerated virtual validation of the autonomous driving function. Preprint at arXiv. arXiv:2507.22769

Aykroyd R G, Leiva V, Ruggeri F (2019). Recent developments of control charts, identification of big data sources and future trends of current research. Technological Forecasting and Social Change, 144: 221–232

Azizi T (2024). Comparative analysis of statistical, time-frequency, and SVM techniques for change detection in nonlinear biomedical signals. Signals, 5(4): 736–755

Bachoc F, Béthune L, Gonzalez-Sanz A, Loubes J M (2023). Gaussian processes on distributions based on regularized optimal transport. In: Proceedings of The 26th International Conference on Artificial Intelligence and Statistics, PMLR, 206: 4986–5010

Bachoc F, Gamboa F, Loubes J M, Venet N (2018). A Gaussian

process regression model for distribution inputs. IEEE Transactions on Information Theory, 64(10): 6620–6637

Bai J, Bai S, Chu Y, Cui Z, Dang K, Deng X, Fan Y, Ge W, Han Y, Huang F, et al (2023). Qwen technical report. Preprint at arXiv. arXiv:2309.16609

Baranowski R, Chen Y, Fryzlewicz P (2019). Narrowest-over-threshold detection of multiple change points and change-point-like features. Journal of the Royal Statistical Society Series B: Statistical Methodology, 81(3): 649–672

Bardou A, Thiran P, Ranieri G (2024). This too shall pass: removing stale observations in dynamic Bayesian optimization. In: Proceedings of Advances in Neural Information Processing Systems, Curran Associates, Inc., 37: 42696–42737

Barnett I, Onnela J P (2016). Change point detection in correlation networks. Scientific Reports, 6(1): 18893

Basseville M, Nikiforov I V (1993). Detection of Abrupt Changes: Theory and Application. Englewood Cliffs: Prentice hall

Beckers T, Wu Q, Pappas G J (2023). Physics-enhanced Gaussian process variational autoencoder. In: Proceedings of The 5th Annual Learning for Dynamics and Control Conference, PMLR, 211: 521–533

Betancourt J, Bachoc F, Klein T, Idier D, Pedreros R, Rohmer J (2020). Gaussian process metamodeling of functional-input code for coastal flood hazard assessment. Reliability Engineering & System Safety, 198: 106870

Biegel T, Jourdan N, Hernandez C, Cviko A, Metternich J (2022). Deep learning for multivariate statistical in-process control in discrete manufacturing: A case study in a sheet metal forming process. Procedia CIRP, 107: 422–427

Binois M, Wycoff N (2022). A survey on high-dimensional Gaussian process modeling with application to Bayesian optimization. ACM Transactions on Evolutionary Learning and Optimization, 2(2): 1–26

Borchert D, Suarez-Zuluaga D A, Sagmeister P, Thomassen Y E, Herwig C (2019). Comparison of data science workflows for root cause analysis of bioprocesses. Bioprocess and Biosystems Engineering, 42(2): 245–256

Bourazas K, Kiagias D, Tsiamyrtzis P (2022). Predictive control charts (PCC): A Bayesian approach in online monitoring of short runs. Journal of Quality Technology, 54(4): 367–391

Bradley W, Kim J, Kilwein Z, Blakely L, Eydenberg M, Jalvin J, Laird C, Boukouvala F (2022). Perspectives on the integration between first-principles and data-driven modeling. Computers & Chemical Engineering, 166: 107898

Brunel L, Balesdent M, Brevault L, Le Riche R, Sudret B (2025). A survey on MultiFidelity surrogates for simulators with functional outputs: Unified framework and benchmark. Computer Methods in Applied Mechanics and Engineering, 435: 117577

Brunzema P, Jordahn M, Willes J, Trimpe S, Snoek J, Harrison J (2024). Variational last layers for Bayesian optimization. In: Proceedings of NeurIPS 2024 Workshop on Bayesian Decision-making and Uncertainty: 1–15

Caldarelli E, Wenk P, Bauer S, Krause A (2022). Adaptive Gaussian process change point detection. In: Proceedings of the International Conference on Machine Learning, PMLR, 162: 2542–2571

Carbery C M, Woods R, Marshall A H (2018). A Bayesian network based learning system for modelling faults in large-scale manufacturing. In: 2018 IEEE International Conference on Industrial Technology (ICIT): 1357–1362

Carroll J, McDonald A, McMillan D (2016). Failure rate, repair time and unscheduled O&M cost analysis of offshore wind turbines. Wind Energy (Chichester, England), 19(6): 1107–1119

Ceglarek D, Huang W, Zhou S, Ding Y, Kumar R, Zhou Y (2004). Time-based competition in multistage manufacturing: Stream-of-variation analysis (SOVA) methodology. International Journal of Flexible Manufacturing Systems, 16(1): 11–44

Chang C Y, Azvar M, Okwudire C, Kontar R A (2025). LLINBO: Trustworthy LLM-in-the-loop Bayesian optimization. Preprint at arXiv. arXiv:2505.14756

Chang W C, Li C L, Yang Y, Póczos B (2019). Kernel change-point detection with auxiliary deep generative models. In: Proceedings of the International Conference on Learning Representations, ICLR: 1–14

Charisi N D, Hopman H, Kana A A (2025). Multi-fidelity design framework integrating compositional kernels to facilitate early-stage design exploration of complex systems. Journal of Mechanical Design, 147(1): 011701

Chen B, Zhang J, Xiong J, Tang W, Jiang S (2025a). An explainable multi-layer graph attention network for product completion time prediction in aircraft final assembly lines. Journal of Manufacturing Systems, 80: 1053–1071

Chen C, Li O, Tao D, Barnett A, Rudin C, Su J K (2019a). This looks like that: Deep learning for interpretable image recognition, In: Proceedings of Advances in neural information processing systems, Curran Associates, Inc., 32: 1–12

Chen J, Liu K C (2002). On-line batch process monitoring using dynamic PCA and dynamic PLS models. Chemical Engineering Science, 57(1): 63–75

Chen L, Wang J, Mortlock T, Khargonekar P, Al Faruque M A (2025b). Hyperdimensional uncertainty quantification for multi-modal uncertainty fusion in autonomous vehicles perception. In: Proceedings of the Computer Vision and Pattern Recognition Conference: 22306–22316

Chen L, Wang Q, Yang Z, Qiu H, Gao L (2025c). Optimization of expensive black-box problems with penalized expected improvement. Computer Methods in Applied Mechanics and Engineering, 433: 117521

Chen W, Yang K, Yu Z, Shi Y, Chen C P (2024a). A survey on imbalanced learning: latest research, applications and future directions. Artificial Intelligence Review, 57(6): 137

Chen X, Ma M, Zhao Z, Zhai Z, Mao Z (2022). Physics-informed deep neural network for bearing prognosis with multisensory signals. Journal of Dynamics, Monitoring and Diagnostics, 1(4): 200–207

Chen Y, Song S, Li S, Yang L, Wu C (2019b). Domain space transfer extreme learning machine for domain adaptation. IEEE Transactions on Cybernetics, 49(5): 1909–1922

Chen Z, Mak S, Wu C F J (2024b). A hierarchical expected improvement method for Bayesian optimization. Journal of the American Statistical Association, 119(546): 1619–1632

Cheng H P, Cheng C S (2007). A support vector machine for recognizing

control chart patterns in multivariate processes. In: Proceedings of the 5th Asian quality congress: 17–18

Cheng N, Papenmeier L, Becker S, Nardi L (2025). A unified framework for entropy search and expected improvement in Bayesian optimization. In: Proceedings of the 42nd International Conference on Machine Learning, PMLR, 267: 1–15

Cheng Z (2013). An intelligent method of change-point detection based on LS-SVM algorithm. HKIE Transactions, 20(3): 141–147

Chiu J E, Tsai C H (2021). On-line concurrent control chart pattern recognition using singular spectrum analysis and random forest. Computers & Industrial Engineering, 159: 107538

Choi H, Jung K (2025). Impact of data distribution and bootstrap setting on anomaly detection using isolation forest in process quality control. Entropy (Basel, Switzerland), 27(7): 761

Choi S W, Lee I B (2005). Multiblock PLS-based localized process diagnosis. Journal of Process Control, 15(3): 295–306

Chu H, Dong Y, Cheng Q, Yan J, Zhao Y, Cao J, Zhang C, Chen X (2024). Pattern recognition of control charts based on data feature enhancement and ensemble learning of classifiers for dimensional accuracy of products. International Journal of Production Research

Cross E J, Rogers T J, Pitchforth D J, Gibson S J, Zhang S, Jones M R (2024). A spectrum of physics-informed Gaussian processes for regression in engineering. Data-Centric Engineering, 5: e8

Cuentas S, Peñabaena-Niebles R, Garcia E (2017). Support vector machine in statistical process monitoring: a methodological and analytical review. International Journal of Advanced Manufacturing Technology, 91(1-4): 485–500

Cui L, Tian X, Wei Q, Liu Y (2024). A self-attention based contrastive learning method for bearing fault diagnosis. Expert Systems with Applications, 238: 121645

Da Veiga S (2025). Distributional encoding for Gaussian process regression with qualitative inputs. Preprint at arXiv. arXiv:2506.04813

Damiano L, Johnson M, Teixeira J, Morris M D, Niemi J (2022). Automatic dynamic relevance determination for Gaussian process regression with high-dimensional functional inputs. Preprint at arXiv. arXiv:2209.00044

Damianou A (2015). Deep Gaussian processes and variational propagation of uncertainty. Dissertation for the Doctoral Degree. Sheffield: University of Sheffield

Das A, Kong W, Sen R, Zhou Y (2024). A decoder-only foundation model for time-series forecasting. In: Proceedings of the Forty-first International Conference on Machine Learning, PMLR, 235: 10148–10167

De Ryck T, De Vos M, Bertrand A (2021). Change point detection in time series data using autoencoders with a time-invariant representation. IEEE Transactions on Signal Processing, 69: 3513–3524

Deldari S, Smith D V, Xue H, Salim F D (2021). Time series change point detection with self-supervised contrastive predictive coding. In: Proceedings of the web conference 2021: 3124–3135

Dey D, Datta A, Banerjee S (2022). Graphical Gaussian Process Models for Highly Multivariate Spatial Data. Biometrika, 109(4): 993–1014

Dey S, Stori J (2005). A Bayesian network approach to root cause diagnosis of process variations. International Journal of Machine Tools & Manufacture, 45(1): 75–91

Ding L, Mak S, Wu C F J (2025). The BdryMatérn GP: Reliable incorporation of boundary information on irregular domains for Gaussian process modeling. Preprint at arXiv. arXiv:2507.09178

Ding N, He Z, He S (2024). Pointwise profile monitoring considering covariates based on Gaussian process. Computers & Industrial Engineering, 194: 110348

Ding Y, Ceglarek D, Shi J (2002). Fault diagnosis of multistage manufacturing processes by using state space approach. Journal of Manufacturing Science and Engineering, 124(2): 313–322

Donnelly J, Barnett A J, Chen C (2022). Deformable protopnet: An interpretable image classifier using deformable prototypes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition: 10265–10275

Du J, Tao C, Cao X, Tsung F (2025). 3D vision-based anomaly detection in manufacturing: A survey. Frontiers of Engineering Management, 12(2): 343–360

Du L, Jin W, Wang Y, Jiang Q (2022). Dynamic batch process monitoring based on time-slice latent variable correlation analysis. ACS Omega, 7(45): 41069–41081

Du S, Huang D, Lv J (2013). Recognition of concurrent control chart patterns using wavelet transform decomposition and multiclass support vector machines. Computers & Industrial Engineering, 66(4): 683–695

Dunlop M M, Girolami M A, Stuart A M, Teckentrup A L (2018). How deep are deep Gaussian processes? Journal of Machine Learning Research, 19: 1–46

Eltoukhy A E, Wang Z, Chan F T, Chung S H, Ma H L, Wang X (2020). Robust aircraft maintenance routing problem using a turnaround time reduction approach. IEEE Transactions on Systems, Man, and Cybernetics. Systems, 50(12): 4919–4932

Enikeeva F, Klopp O (2025). Change-point detection in dynamic networks with missing links. Operations Research, 73(5): 2417–2429

Ermshaus A, Schäfer P, Leser U (2023). ClaSP: parameter-free time series segmentation. Data Mining and Knowledge Discovery, 37(3): 1262–1300

Fallahdizcheh A, Wang C (2025). Variational inference-based transfer learning for profile monitoring with incomplete data. IISE Transactions, 57(4): 351–366

Fan Z, Wang W, Ng S H, Hu Q (2024). Minimizing UCB: A better local search strategy in local Bayesian optimization. In: Proceedings of Advances in Neural Information Processing Systems, Curran Associates, Inc., 37: 130602–130634

Feng Z, Ye X, Huang W, Zhai C (2025). Gaussian process-based robust optimization with symmetry modeling and variable selection. Symmetry, 17(1): 113

Fernández-Godino M G (2023). Review of multi-fidelity models. Advances in Computational Science and Engineering, 1(4): 351–400

Folch J P, Odgers J A C, Zhang S, Lee R M, Shafei B, Walz D, Tsay C, van der Wilk M, Misener R (2023). Practical path-based Bayesian optimization. In: Proceedings of NeurIPS 2023 Workshop on Adaptive Experimental Design and Active Learning in the Real World: 1–12

Folch J P, Tsay C, Lee R M, Shafei B, Ormaniec W, Krause A, van der Wilk M, Misener R, Mutný M (2024). Transition constrained Bayesian optimization via Markov decision processes. In: Proceedings of Advances in Neural Information Processing Systems, Curran Associates, Inc., 37: 88194–88235

Folch J P, Zhang S, Lee R, Shafei B, Walz D, Tsay C, van der Wilk M, Misener R (2022). SnAKe: Bayesian optimization with pathwise exploration. In: Proceedings of Advances in Neural Information Processing Systems, Curran Associates, Inc., 35: 35226–35239

Ford J J, James J, Molloy T L (2023). Exactly optimal Bayesian quickest change detection for hidden Markov models. Automatica, 157: 111232

Fortuin V, Baranchuk D, Rätsch G, Mandt S (2020). GP-VAE: Deep probabilistic time series imputation. In: Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR: 1651–1661

Fox E B, Sudderth E B, Jordan M I, Willsky A S (2011). A sticky HDP-HMM with application to speaker diarization. Annals of Applied Statistics, 5(2A): 1020–1056

Frazier P I (2018). A tutorial on Bayesian optimization. Preprint at arXiv. arXiv:1807.02811

Frick K, Munk A, Sieling H (2014). Multiscale change point inference. Journal of the Royal Statistical Society Series B: Statistical Methodology, 76(3): 495–580

Fryzlewicz P (2014). Wild binary segmentation for multiple change-point detection. Annals of Statistics, 42(6): 2243–2281

Gama J, Žliobaite I, Bifet A, Pechenizkiy M, Bouchachia A (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4): 1–37

Gao S, Liu J, Zhang Z, Feng C, He B, Zio E (2023). Physics-guided generative adversarial networks for fault detection of underwater thruster. Ocean Engineering, 286: 115585

García E, Penabaena-Niebles R, Jubiz-Diaz M, Perez-Tafur A (2022). Concurrent control chart pattern recognition: A systematic review. Mathematics, 10(6): 934

Garnett R (2023). Bayesian optimization. Cambridge, New York: Cambridge University Press

Garreau D, Arlot S (2018). Consistent change-point detection with kernels. Electronic Journal of Statistics, 12(2): 4440–4486

Gawde S, Patil S, Kumar S, Kamat P, Kotecha K, Alfarhood S (2024). Explainable predictive maintenance of rotating machines using LIME, SHAP, PDP, ICE. IEEE Access : Practical Innovations, Open Solutions, 12: 29345–29361

Gbashi S M, Olatunji O O, Adedeji P A, Madushele N (2025). Control chart-integrated machine learning models for incipient fault detection in wind turbine main bearing. Discover Artificial Intelligence, 5(1): 149

Geirhos R, Jacobsen J H, Michaelis C, Zemel R, Brendel W, Bethge M, Wichmann F A (2020). Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11): 665–673

Ghosh M, Li Y, Zeng L, Zhang Z, Zhou Q (2021). Modeling multivariate profiles using Gaussian process-controlled B-splines. IISE Transactions, 53(7): 787–798

Glaser J, Whiteway M, Cunningham J P, Paninski L, Linderman S (2020). Recurrent switching dynamical systems models for multiple interacting neural populations. In: Proceedings of Advances in Neural Information Processing Systems, Curran Associates, Inc., 33: 14867–14878

Gómez-Andrades A, Munoz P, Serrano I, Barco R (2016). Automatic root cause analysis for LTE networks based on unsupervised techniques. IEEE Transactions on Vehicular Technology, 65(4): 2369–2386

Gondur R, Sikandar U B, Schaffer E, Aoi M C, Keeley S L (2024). Multi-modal Gaussian process variational autoencoders for neural and behavioral data. In: Proceedings of the International Conference on Learning Representations, ICLR: 14257–14281

Le Gratiet L, Garnier J (2014). Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. International Journal for Uncertainty Quantification, 4(5): 365–386

Guo C, Pleiss G, Sun Y, Weinberger K Q (2017). On calibration of modern neural networks. In: Proceedings of the International Conference on Machine Learning, PMLR: 1321–1330

Guo D, Yang D, Zhang H et al. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature, 645(8081): 633–638

Gupta M, Wadhvani R, Rasool A (2022). Real-time change-point detection: A deep neural network-based adaptive approach for detecting changes in multivariate time series data. Expert Systems with Applications, 209: 118260

Guth S, Mojahed A, Sapsis T P (2024). Quality measures for the evaluation of machine learning architectures on the quantification of epistemic and aleatoric uncertainties in complex dynamical systems. Computer Methods in Applied Mechanics and Engineering, 420: 116760

Ha J, Lee S, Kim D, Choi J (2025). A case study of graph neural network-based anomaly detection and root cause visualization for quality improvement in semiconductor manufacturing. Journal of Korean Society for Quality Management, 53(2): 237–248

Han J, Lee K, Tong A, Choi J (2019). Confirmatory Bayesian online change point detection in the covariance structure of gaussian processes. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization: 2449–2455

Hansen M H, Nair V N, Friedman D J (1997). Monitoring wafer map data from integrated circuit fabrication processes for spatially clustered defects. Technometrics, 39(3): 241–253

Harchaoui Z, Moulines E, Bach F (2008). Kernel change-point analysis. In: Proceedings of Advances in neural information processing systems, Curran Associates, Inc., 21: 1–8

He H, Garcia E A (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9): 1263–1284

He W, Jiang Z, Xiao T, Xu Z, Li Y (2025). A survey on uncertainty quantification methods for deep learning. ACM Computing Surveys, 58(7): 179

Heinze-Deml C, Peters J, Meinshausen N (2018). Invariant causal prediction for nonlinear models. Journal of Causal Inference, 6(2): 20170016

Hong Z, Li Y, Zeng Z (2019). Convolutional neural network for control chart patterns recognition. In: Proceedings of the 3rd International Conference on Computer Science and Application Engi-

neering: 1–9

Hotelling H (1992). The Generalization of Student's Ratio. New York: Springer New York: 54–65

Hou Y, Wang J, Chen Z, Ma J, Li T (2023). Diagnosisformer: An efficient rolling bearing fault diagnosis method based on improved Transformer. Engineering Applications of Artificial Intelligence, 124: 106507

Howard P, Apley D W, Runger G (2018). Identifying nonlinear variation patterns with deep autoencoders. IISE Transactions, 50(12): 1089–1103

Huang W, Lu H, Zhang H (2023). Hierarchical kernels in deep kernel learning. Journal of Machine Learning Research, 24: 1–30

Hunter J S (1986). The exponentially weighted moving average. Journal of Quality Technology, 18(4): 203–210

Jalilibal Z, Karavigh M H A, Maleki M R, Amiri A (2024). Control charting methods for monitoring high dimensional data streams: A conceptual classification scheme. Computers & Industrial Engineering, 191: 110141

Jang S, Park S H, Baek J G (2017). Real-time contrasts control chart using random forests with weighted voting. Expert Systems with Applications, 71: 358–369

Jardine A K, Lin D, Banjevic D (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mechanical Systems and Signal Processing, 20(7): 1483–1510

Javanmard A, Montanari A (2018). Online rules for control of false discovery rate and false discovery exceedance. Annals of Statistics, 46(2): 526–554

Jazbec M, Ashman M, Fortuin V, Pearce M, Mandt S, Rätsch G (2021). Scalable Gaussian process variational autoencoders. In: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, PMLR: 3511–3519

Jiang C, Chen H, Xu Q, Wang X (2023). Few-shot fault diagnosis of rotating machinery with two-branch prototypical networks. Journal of Intelligent Manufacturing, 34(4): 1667–1681

Jiang X, Georgaka S, Rattray M, Álvarez M A (2025). Scalable multi-output Gaussian processes with stochastic variational inference. Transactions on Machine Learning Research, 2025: 1–26

Jin T (2023). Bridging reliability and operations management for superior system availability: challenges and opportunities. Frontiers of Engineering Management, 10(3): 391–405

Jones D R, Schonlau M, Welch W J (1998). Efficient global optimization of expensive black-box functions. Journal of Global Optimization, 13(4): 455–492

Jones M A, Steiner S H (2012). Assessing the effect of estimation error on risk-adjusted CUSUM chart performance. International Journal for Quality in Health Care, 24(2): 176–181

Kaiser M, Sipos M (2022). Unsuitability of notears for causal graph discovery when dealing with dimensional quantities. Neural Processing Letters, 54(3): 1587–1595

Kakde D, Peredriy S, Chaudhuri A (2017). A non-parametric control chart for high frequency multivariate data. In: Proceedings of the 2017 Annual Reliability and Maintainability Symposium (RAMS), IEEE: 1–6

Kanamori T, Hido S, Sugiyama M (2009). A least-squares approach to direct importance estimation. Journal of Machine Learning Research, 10: 1391–1445

Karniadakis G E, Kevrekidis I G, Lu L, Perdikaris P, Wang S, Yang L (2021). Physics-informed machine learning. Nature Reviews. Physics, 3(6): 422–440

Kasilingam S, Yang R, Singh S K, Farahani M A, Rai R, Wuest T (2024). Physicsbased and data-driven hybrid modeling in manufacturing: a review. Production & Manufacturing Research, 12(1): 2305358

Kawahara Y, Sugiyama M (2009). Change-point detection in time-series data by direct density-ratio estimation. In: Proceedings of the 2009 SIAM international conference on data mining, SIAM: 389–400

Kazmi M W, Noor-ul-Amin M (2024). Adaptive EWMA control chart by using support vector regression. Quality and Reliability Engineering International, 40(7): 3831–3843

Kei Y L, Li H, Chen Y, Padilla O H M (2025a). Change point detection on a separable model for dynamic networks. Transactions on Machine Learning Research, 2025: 1–41

Kei Y L, Li J, Li H, Chen Y, Padilla O H M (2025b). Change point detection in dynamic graphs with decoder-only latent space model. Transactions on Machine Learning Research, 2025: 1–32

Kennedy M C, O'Hagan A (2001). Bayesian calibration of computer models. Journal of the Royal Statistical Society Series B: Statistical Methodology, 63(3): 425–464

Khan I, Alamri A M, Almarashi A M, Elhag A A, Aripov M, Hussain S (2024). Bayesian control chart using variable sample size with engineering applications. Scientific Reports, 14(1): 24683

Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: Proceedings of the International Conference on Machine Learning, PMLR: 2668–2677

Kim T H, Kim H R, Cho Y J (2021). Product inspection methodology via deep learning: An overview. Sensors (Basel), 21(15): 5039

Kim Y, Kim S B (2018). Optimal false alarm controlled support vector data description for multivariate process monitoring. Journal of Process Control, 65: 1–14

Knoblauch J, Damoulas T (2018). Spatio-temporal Bayesian on-line changepoint detection with model selection. In: Proceedings of the International Conference on Machine Learning, PMLR: 2718–2727

Kovari A (2024). AI for decision support: Balancing accuracy, transparency, and trust across sectors. Information (Basel), 15(11): 725

Kozlov I, Rivkin D, Chang W D, Wu D, Liu X, Dudek G (2023). Self-supervised transformer architecture for change detection in radio access networks. In: Proceedings of the 2023-IEEE International Conference on Communications, IEEE: 2227–2232

Krawczyk B (2016). Learning from imbalanced data: open challenges and future directions. Progress in artificial intelligence, 5: 221–232

Krishnapriyan A, Gholami A, Zhe S, Kirby R, Mahoney M W (2021). Characterizing possible failure modes in physics-informed neural networks. In: Proceedings of Advances in neural information processing systems, Curran Associates, Inc., 34: 26548–26560

Lakshminarayanan B, Pritzel A, Blundell C (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In: Proceedings of Advances in neural information processing systems,

Curran Associates, Inc., 30: 1–12

Laugksch K, Rousseau P, Laubscher R (2023). A PINN surrogate modeling methodology for steady-state integrated thermofluid systems modeling. Mathematical & Computational Applications, 28(2): 52

Lee I, Park S H, Baek J G (2020). Random-forest-based real-time contrasts control chart using adaptive breakpoints with symbolic aggregate approximation. Expert Systems with Applications, 158: 113407

Lee P H, Liao S L (2023). Residual control chart based on a convolutional neural network and support vector regression for type-i censored data with the Weibull model. Mathematics, 12(1): 74

Lee P H, Torng C C, Lin C H, Chou C Y (2022). Control chart pattern recognition using spectral clustering technique and support vector machine under gamma distribution. Computers & Industrial Engineering, 171: 108437

Lee S, Kim S B (2018). Time-adaptive support vector data description for nonstationary process monitoring. Engineering Applications of Artificial Intelligence, 68: 18–31

Lee S, Kwak M, Tsui K L, Kim S B (2019). Process monitoring using variational autoencoder for high-dimensional nonlinear processes. Engineering Applications of Artificial Intelligence, 83: 13–27

Lei B, Kirk T Q, Bhattacharya A, Pati D, Qian X, Arroyave R, Mallick B K (2021). Bayesian optimization with adaptive surrogate models for automated experimental design. npj Computational Materials, 7(1): 194

Li C, Cui X, Xiong S (2023a). Design and analysis of computer experiments with both numeral and distributional inputs. Technometrics, 65(3): 406–417

Li D, Wang K (2025a). A partial domain generalization method for modeling multiple multistage manufacturing processes. IISE Transactions, 57(9): 1105–1120

Li G, Wang Y, Kar S, Jin X (2026). Bayesian optimization with active constraint learning for advanced manufacturing process design. IISE Transactions, 58(3): 257–271

Li H, Gou L, Li H, Liu Z (2023b). Physics-guided neural network model for aeroengine control system sensor fault diagnosis under dynamic conditions. Aerospace (Basel, Switzerland), 10(7): 644

Li H, Jia M, Mao Z (2024a). Time-slice dynamic prediction and multiway serial PCA for batch industrial process monitoring. Computers & Chemical Engineering, 182: 108580

Li H, Jiao J, Liu Z, Lin J, Zhang T, Liu H (2025a). Trustworthy Bayesian deep learning framework for uncertainty quantification and confidence calibration: Application in machinery fault diagnosis. Reliability Engineering & System Safety, 255: 110657

Li J, Fearnhead P, Fryzlewicz P, Wang T (2024b). Automatic change-point detection in time series via deep learning. Journal of the Royal Statistical Society Series B: Statistical Methodology, 86(2): 273–285

Li J, Wang H (2025b). Gaussian process regression for uncertainty quantification: An introductory tutorial. Preprint at arXiv. arXiv:2502.03090

Li M, Kontar R (2022). On negative transfer and structure of latent functions in multioutput Gaussian processes. SIAM/ASA Journal on Uncertainty Quantification, 10(4): 1714–1732

Li R, Xia T, Luo F, Jiang Y, Chen Z, Xi L (2024c). Hybrid physics-embedded recurrent neural networks for fault diagnosis under time-varying conditions based on multivariate proprioceptive signals. Advanced Engineering Informatics, 62: 102851

Li S, Xie Y, Dai H, Song L (2015). M-statistic for kernel change-point detection. In: Proceedings of Advances in neural information processing systems, Curran Associates, Inc., 28: 1–9

Li S, Xie Y, Dai H, Song L (2019). Scan B-statistic for kernel change-point detection. Sequential Analysis, 38(4): 503–544

Li S, Xing W, Kirby R, Zhe S (2020). Multi-fidelity Bayesian optimization via deep neural networks. In: Proceedings of Advances in Neural Information Processing Systems, Curran Associates, Inc., 33: 8521–8531

Li S, Yuan Z, Du T, Hao R, Zhao H, Wang K (2025b). Surrogate model for lunar sampling based on Bayesian neural network and active learning to enhance coring efficiency. Aerospace (Basel, Switzerland), 12(2): 128

Li T, Zhou Y, Zhao Y, Zhang C, Zhang X (2022a). A hierarchical object oriented Bayesian network-based fault diagnosis method for building energy systems. Applied Energy, 306: 118088

Li W, Zhang C, Tsung F, Mei Y (2021). Nonparametric monitoring of multivariate data via KNN learning. International Journal of Production Research, 59(20): 6311–6326

Li Y, Dai W, He Y (2024d). Control chart pattern recognition under small shifts based on multi-scale weighted ordinal pattern and ensemble classifier. Computers & Industrial Engineering, 189: 109940

Li Y, Huang M, Pan E (2018). Residual chart with hidden Markov model to monitoring the auto-correlated processes. Journal of Shanghai Jiaotong University (Science), 23(S1): 103–108

Li Y, Li H, Chen Z, Zhu Y (2022b). An improved hidden Markov model for monitoring the process with autocorrelated observations. Energies, 15(5): 1685

Li Y, Zhou Q, Jiang W, Tsui K L (2024e). Optimal composite likelihood estimation and prediction for distributed Gaussian process modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(2): 1134–1147

Li Y L, Rudner T G J, Wilson A G (2024f). A study of Bayesian neural network surrogates for Bayesian optimization. In: Proceedings of the Twelfth International Conference on Learning Representations, ICLR: 1–39

Li Z, Kristoffersen E, Li J (2022c). Deep transfer learning for failure prediction across failure types. Computers & Industrial Engineering, 172: 108521

Liang Y, Wen H, Nie Y, Jiang Y, Jin M, Song D, Pan S, Wen Q (2024). Foundation models for time series analysis: A tutorial and survey. In: Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining: 6555–6565

Lin H, Yang L, Bao H, Zhang F, Zhao F, Lu C (2025a). The reliability analysis of a turbine rotor structure based on the Kriging surrogate model. Machines, 13(7): 625

Lin J A, Ament S, Balandat M, Eriksson D, Hernández-Lobato J M, Bakshy E (2025b). Scalable Gaussian processes with latent Kronecker structure. In: Proceedings of the 42nd International Conference on Machine Learning, PMLR, 267: 1–15

Lin Q, Hu J, Zhou Q, Cheng Y, Hu Z, Couckuyt I, Dhaene T (2021). Multi-output Gaussian process prediction for computationally expensive problems with multiple levels of fidelity. Knowledge-Based Systems, 227: 107151

Lin Q, Hu J, Zhou Q, Shu L, Zhang A (2024a). A multi-fidelity Bayesian optimization approach for constrained multi-objective optimization problems. Journal of Mechanical Design, 146(7): 071702

Lin S Y, Guh R S, Shiue Y R (2011). Effective recognition of control chart patterns in autocorrelated data using a support vector machine based approach. Computers & Industrial Engineering, 61(4): 1123–1134

Lin W A, Sung C L, Chen R B (2024b). Category tree Gaussian process for computer experiments with many-category qualitative factors and application to cooling system design. Journal of Quality Technology, 56(5): 391–408

Linardatos P, Papastefanopoulos V, Kotsiantis S (2021). Explainable AI: A review of machine learning interpretability methods. Entropy (Basel, Switzerland), 23(1): 18

Linderman S, Johnson M, Miller A, Adams R, Blei D, Paninski L (2017). Bayesian learning and inference in recurrent switching linear dynamical systems. In: Proceedings of 20th International Conference on Artificial Intelligence and Statistics, PMLR: 914–922

Lipton Z C (2018). The mythos of model interpretability. Communications of the ACM, 61(10): 36–43

Liu H C, Liu R, Gu X, Yang M (2023). From total quality management to quality 4.0: A systematic literature review and future research agenda. Frontiers of Engineering Management, 10(2): 191–205

Liu P, Xu H, Zhang C (2024a). A comprehensive survey of recent research on profile data analysis. Journal of Quality Technology, 56(5): 428–454

Liu R, Zhang Q, Lin D, Zhang W, Ding S X (2024b). Causal intervention graph neural network for fault diagnosis of complex industrial processes. Reliability Engineering & System Safety, 251: 110328

Liu S, Yamada M, Collier N, Sugiyama M (2013). Change-point detection in time-series data by relative density-ratio estimation. Neural Networks, 43: 72–83

Liu W, Han B, Zheng A, Zheng Z (2024c). Fault diagnosis for reducers based on a digital twin. Sensors (Basel), 24(8): 2575

Liu Y, Liu Y, Jung U (2020). Nonparametric multivariate control chart based on density-sensitive novelty weight for non-normal processes. Quality Technology & Quantitative Management, 17(2): 203–215

Liu Z, Li Y, Yue X, Pan E (2025). Latent functional Gaussian process incorporating output spatial correlations. IISE Transactions, 57(12): 1436–1449

Liu Z, Zhang P, Yu Y, Li M, Zeng Z (2024d). A novel fault diagnosis model of rolling bearing under variable working conditions based on attention mechanism and domain adversarial neural network. Journal of Mechanical Science and Technology, 38(3): 1101–1111

Londschien M, Bühlmann P, Kovács S (2023). Random forests for change point detection. Journal of Machine Learning Research, 24: 1–45

Lu Q, Polyzos K D, Li B, Giannakis G B (2023a). Surrogate modeling for Bayesian optimization beyond a single Gaussian process. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(9): 11283–11296

Lu W, Wang Y, Zhang M, Gu J (2024). Physics guided neural network: Remaining useful life prediction of rolling bearings using long short-term memory network through dynamic weighting of degradation process. Engineering Applications of Artificial Intelligence, 127: 107350

Lu Z, Guo C, Liu M, Shi R (2023b). Remaining useful lifetime estimation for discrete power electronic devices using physics-informed neural network. Scientific Reports, 13(1): 10167

Lübsen J, Hespe C, Eichler A (2024). Towards Safe Multi-Task Bayesian Optimization. In: Proceedings of the 6th Annual Learning for Dynamics & Control Conference, PMLR: 839–851

Lundberg S M, Lee S I (2017). A unified approach to interpreting model predictions. In: Proceedings of Advances in neural information processing systems, Curran Associates, Inc., 30: 1–10

Ma C, Álvarez M A (2023). Large scale multi-output multi-class classification using Gaussian processes. Machine Learning, 112(4): 1077–1106

Ma P, Mondal A, Konomi B A, Hobbs J, Song J J, Kang E L (2022). Computer model emulation with high-dimensional functional output in large-scale observing system uncertainty experiments. Technometrics, 64(1): 65–79

MacGregor J F, Kourti T (1995). Statistical process control of multivariate processes. Control Engineering Practice, 3(3): 403–414

Marque-Pucheu S, Perrin G, Garnier J (2020). An efficient dimension reduction for the Gaussian process emulation of two nested codes with functional outputs. Computational Statistics, 35(3): 1059–1099

Mason R L, Tracy N D, Young J C (1995). Decomposition of T2 for multivariate control chart interpretation. Journal of Quality Technology, 27(2): 99–108

Mayer J, Engels M, Kaufmann T, Niemietz P, Bergs T (2025). A decision-making methodology for selecting digital twin applications in the product service phase considering value and effort. Production Engineering, 19(6): 1075–1092

Mbogu H M, Nicholson C D (2024). Data-driven root cause analysis via causal discovery using time-to-event data. Computers & Industrial Engineering, 190: 109974

Miettinen K (1999). Nonlinear Multiobjective Optimization. Berlin: Springer Science & Business Media

Moliner-Heredia R, Peñarrocha-Alós I, Abellán-Nebot J V (2023). A methodology for data-driven adjustment of variation propagation models in multistage manufacturing processes. Journal of Manufacturing Systems, 67: 281–295

Montgomery D C (2020). Introduction to Statistical Quality Control. Hoboken: John wiley & sons

Moreno-Muñoz P, Artés A, Álvarez M (2018). Heterogeneous multi-output Gaussian process prediction. In: Proceedings of Advances in Neural Information Processing Systems, Curran Associates, Inc., 31: 1–10

Mori J, Yu J (2013). Dynamic Bayesian network based networked process monitoring for fault propagation identification and root cause diagnosis of complex dynamic processes. IFAC Proceedings

Volumes, 46: 678–683

Moss J, England J, Lio P (2024). Deep kernel learning of nonlinear latent force models. Transactions on Machine Learning Research, 2024: 1–15

Mothilal R K, Sharma A, Tan C (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency: 607–617

Mucsányi B, Kirchhof M, Oh S J (2024). Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. In: Proceedings of Advances in Neural Information Processing Systems, Curran Associates, Inc., 37: 50972–51038

Murphy K P (2022). Probabilistic Machine Learning: An Introduction. Cambridge: The MIT Press, Adaptive Computation and Machine Learning Series

Nguyen T T V, Heuchenne C, Tran K D, Tartare G, Tran K P (2025). SVDD control charts based on MEWMA technique for monitoring compositional data. Computers & Industrial Engineering, 201: 110865

Niu Y S, Hao N, Zhang H (2016). Multiple change-point detection: a selective overview. Statistical Science, 31(4): 611–623

Nomikos P, MacGregor J F (1995). Multivariate SPC charts for monitoring batch processes. Technometrics, 37(1): 41–59

Noor-ul-Amin M, Khan I, Alzahrani A R R, Ayari-Akkari A, Ahmad B (2024). Risk adjusted EWMA control chart based on support vector machine with application to cardiac surgery data. Scientific Reports, 14(1): 9633

Oh C, Jeong J (2020). VODCA: Verification of diagnosis using cam-based approach for explainable process monitoring. Sensors (Basel), 20(23): 6858

Okhrin Y, Schmid W, Semeniuk I (2025). A control chart for monitoring image processes based on convolutional neural networks. Statistica Neerlandica, 79(1): e12366

e Oliveira E, Miguéis V L, Borges J L (2023). Automatic root cause analysis in manufacturing: an overview & conceptualization. Journal of Intelligent Manufacturing, 34(5): 2061–2078

Oprescu M, Syrgkanis V, Battocchi K, Hei M, Lewis G (2019). EconML: A machine learning library for estimating heterogeneous treatment effects. In: Proceedings of the 33rd Conference on Neural Information Processing Systems: 6: 1–6

Ottenstreuer S, Weiß C H, Knoth S (2021). Control charts for monitoring a Poisson hidden Markov process. Quality and Reliability Engineering International, 37(2): 484–501

Ottenstreuer S, Weiß C H, Testik M C (2023). A review and comparison of control charts for ordinal samples. Journal of Quality Technology, 55(4): 422–441

Ouedraogo E B, Hawbani A, Wang X, Liu Z, Zhao L, Al-qaness M A, Alsamhi S H (2025). Digital twin data management: A comprehensive review. IEEE Transactions on Big Data, 11(5): 2224–2243

Oune N, Bostanabad R (2021). Latent map Gaussian processes for mixed variable metamodeling. Computer Methods in Applied Mechanics and Engineering, 387: 114128

Page E S (1954). Continuous inspection schemes. Biometrika, 41(1-2): 100–115

Pan S, Vermetten D, López-Ibáñez M, Bäck T, Wang H (2025). Transfer learning of surrogate models: Integrating domain warping and affine transformations. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion: 2481–2490

Paulson J A, Lu C (2022). Cobalt: Constrained Bayesian optimization of computationally expensive grey-box models exploiting derivative information. Computers & Chemical Engineering, 160: 107700

Paulson J A, Tsay C (2025). Bayesian optimization as a flexible and efficient design framework for sustainable process systems. Current Opinion in Green and Sustainable Chemistry, 51: 100983

Pearl J (2009). Causality. Cambridge: Cambridge University Press

Peel L, Clauset A (2015). Detecting change points in the large-scale structure of evolving networks. Proceedings of the AAAI Conference on Artificial Intelligence, 29(1): 1–7

Penaloza E, Stevens N (2024). Changepoint detection in highly-attributed dynamic graphs. In: Proceedings of the 41st International Conference on Machine Learning, ICML, 235: 1–12

Pensoneault A, Yang X, Zhu X (2020). Nonnegativity-enforced Gaussian process regression. Theoretical & Applied Mechanics Letters, 10(3): 182–187

Peters J, Bühlmann P, Meinshausen N (2016). Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society Series B: Statistical Methodology, 78(5): 947–1012

Pires A V, Moustapha M, Marelli S, Sudret B (2025). Reliability analysis for data-driven noisy models using active learning. Structural Safety, 112: 102543

Puchkin N, Shcherbakova V (2023). A contrastive approach to online change point detection. In: Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR: 5686–5713

Qin Y, Ye Y, Fang J et al. (2025). UI-TARS: Pioneering Automated GUI Interaction with Native Agents. Preprint at arXiv. arXiv:2501.12326

Qiu P (2013). Introduction to statistical process control. Boca Raton: CRC press

Qiu P (2024). Machine Learning Approaches for Statistical Process Control. In: John Wiley & Sons, Ltd: 1–8

Rahaman R (2021). Uncertainty quantification and deep ensembles. In: Proceedings of Advances in Neural Information Processing Systems, Curran Associates, Inc., 34: 20063–20075

Rahim Y, Ahsan M (2025). Residual-based air quality monitoring: A hybrid xgboost-AEWMA control chart for detecting pm 2.5 anomalies. In: Proceedings of the 2025 International Conference on Data Science and Its Applications, IEEE: 1208–1215

Ramos M, Ascencio J, Hinojosa M V, Vera F, Ruiz O, Jimenez-Feijoó M I, Galindo P (2021). Multivariate statistical process control methods for batch production: A review focused on applications. Production & Manufacturing Research, 9(1): 33–55

Rasmussen C E, Williams C K I (2006). Gaussian Processes for Machine Learning. Cambridge: MIT Press, Adaptive Computation and Machine Learning

Rasul K, Ashok A, Williams A R, Khorasani A, Adamopoulos G, Bhagwatkar R, Biloš M, Ghonia H, Hassen N, Schneider A (2023). Lag-llama: Towards foundation models for time series forecasting. In: Proceedings of Robustness of Few-shot and Zero-shot Learning in Large Foundation Models, R0-FoMo: 1–13

Rathore P, Lei W, Frangella Z, Lu L, Udell M (2024). Challenges in training PINNs: a loss landscape perspective. In: Proceedings of the 41st International Conference on Machine Learning, PMLR, 235: 42159–42191

Ribeiro M T, Singh S, Guestrin C (2016). 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: 1135–1144

Rodemann J, Augustin T (2024). Imprecise Bayesian optimization. Knowledge-Based Systems, 300: 112186

Rudolph M, Kurz S, Rakitsch B (2024). Hybrid modeling design patterns. Journal of Mathematics in Industry, 14(1): 3

Runge J (2020). Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In: Proceedings of the conference on uncertainty in artificial intelligence, PMLR: 1388–1397

Runge J, Nowack P, Kretschmer M, Flaxman S, Sejdinovic D (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances, 5(11): eaau4996

Saatçi Y, Turner R D, Rasmussen C E (2010). Gaussian process change point models. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10): 927–934

Sabahno H, Khodadad D (2025). A convolutional neural network-based joint detection and localization spatiotemporal scheme for process control through speckle pattern imaging. Computers & Industrial Engineering, 210: 111538

Sacks J, Welch W J, Mitchell T J, Wynn H P (1989). Design and analysis of computer experiments. Statistical Science, 4: 409–423

Saleem R, Yuan B, Kurugollu F, Anjum A, Liu L (2022). Explaining deep neural networks: a survey on the global interpretation methods. Neurocomputing, 513: 165–180

Saleh R A, Al-Areqi F, Konyar M Z, Kaplan K, Öngir S, Ertunc H M (2024). Advancingtire safety: Explainable artificial intelligence-powered foreign object defect detection with xception networks and grad-cam interpretation. Applied Sciences (Basel, Switzerland), 14(10): 4267

Salimbeni H, Dutordoir V, Hensman J, Deisenroth M (2019). Deep Gaussian processes with importance-weighted variational inference. In: Proceedings of the International Conference on Machine Learning, PMLR: 5589–5598

Santner T J, Williams B J, Notz W I (2018). The Design and Analysis of Computer Experiments. New York: Springer New York, Springer Series in Statistics

Saranya A, Subhashini R (2023). A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. Decision Analytics Journal, 7: 100230

Sauer A, Cooper A, Gramacy R B (2023a). Vecchia-approximated deep Gaussian processes for computer experiments. Journal of Computational and Graphical Statistics, 32(3): 824–837

Sauer A, Gramacy R B, Higdon D (2023b). Active learning for deep Gaussian process surrogates. Technometrics, 65(1): 4–18

Saunders J D, Freitas A A (2022). Evaluating the predictive performance of positive unlabelled classifiers: a brief critical review and practical recommendations for improvement. SIGKDD Explorations, 24(2): 5–11

Sellier J, Dellaportas P (2023). Bayesian online change point detection with Hilbert space approximate Student-t process. In: International Conference on Machine Learning, PMLR: 30553–30569

Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017). Gradcam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision: 618–626

Semenova E, Sheinkman A, Hitge T J, Hall S M, Cockayne J (2025). SMRS: Advocating a unified reporting standard for surrogate models in the artificial intelligence era. In: Proceedings of the 39th Annual Conference on Neural Information Processing Systems Position Paper Track: 1–26

Senoner J, Netland T, Feuerriegel S (2022). Using explainable artificial intelligence to improve process quality: evidence from semiconductor manufacturing. Management Science, 68(8): 5704–5723

Shahriari B, Swersky K, Wang Z, Adams R P, De Freitas N (2016). Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE, 104(1): 148–175

Shahtaheri Y, Flint M M, de la Garza J M (2019). A multi-objective reliability-based decision support system for incorporating decision maker utilities in the design of infrastructure. Advanced Engineering Informatics, 42: 100939

Shan R, Jia X, Su X, Xu Q, Ning H, Zhang J (2025). AI-driven multi-objective optimization and decision-making for urban building energy retrofit: Advances, challenges, and systematic review. Applied Sciences (Basel, Switzerland), 15(16): 8944

Sharma A, Kiciman E (2020). DoWhy: An end-to-end library for causal inference. Preprint at arXiv. arXiv:2011.04216

Shen B, Gnanasambandam R, Wang R, Kong Z J (2023). Multi-task Gaussian process upper confidence bound for hyperparameter tuning and its application for simulation studies of additive manufacturing. IISE Transactions, 55(5): 496–508

Shewhart W A (1930). Economic quality control of manufactured product. Bell System Technical Journal, 9(2): 364–389

Shewhart W A (2022). Economic control of quality of manufactured product. Barakaldo: Barakaldo Books

Shi J (2006). Stream of variation modeling and analysis for multistage manufacturing processes. Boca Raton: CRC press

Shi J Q, Choi T (2011). Gaussian Process Regression Analysis for Functional Data. Boca Raton: Chapman and Hall/CRC

Shin K S, Lee I, Baek J G (2019). An improved real-time contrasts control chart using novelty detection and variable importance. Applied Sciences (Basel, Switzerland), 9(1): 173

Shin Y, Park J, Yoon S, Song H, Lee B S, Lee J G (2024). Exploiting representation curvature for boundary detection in time series. In: Proceedings of Advances in Neural Information Processing Systems, Curran Associates, Inc., 37: 5974–5995

Shiraishi T, Miwa D, Le Duy V N, Takeuchi I (2024). Selective inference for change point detection by recurrent neural network. Neural Computation, 37(1): 160–192

Shyalika C, Wickramarachchi R, Sheth A P (2024). A comprehensive survey on rare event prediction. ACM Computing Surveys, 57(3): 1–39

Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods.

In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society: 180–186

Snoek J, Rippel O, Swersky K, Kiros R, Satish N, Sundaram N, Patwary M, Prabhat M, Adams R (2015). Scalable Bayesian optimization using deep neural networks. In: Proceedings of the 32nd International Conference on Machine Learning, PMLR: 2171–2180

Song J, Cui Y, Wei P, Valdebenito M A, Zhang W (2024). Constrained Bayesian optimization algorithms for estimating design points in structural reliability analysis. Reliability Engineering & System Safety, 241: 109613

Song K, Liu C, Jiang D (2025). A positive-unlabeled learning approach for industrial anomaly detection based on self-adaptive training. Neurocomputing, 647: 130488

Soori M, Jough F K G, Dastres R, Arezoo B (2026). AI-based decision support systems in Industry 4.0, a review. Journal of Economy and Technology, 4: 206–225

Spitieris M, Ruocco M, Murad A, Nocente A (2025). PIGPVAE: Physics-informed Gaussian process variational autoencoders. Applied Intelligence, 55(12): 894

Steiner S H, Cook R J, Farewell V T, Treasure T (2000). Monitoring surgical performance using risk-adjusted cumulative sum charts. Biostatistics (Oxford, England), 1(4): 441–452

Steiner S H, Jones M (2010). Risk-adjusted survival time monitoring with an updating exponentially weighted moving average (EWMA) control chart. Statistics in Medicine, 29(4): 444–454

Su Y, Yan P, Lin J, Wen C, Fan Y (2024). Few-shot defect recognition for the multidomain industry via attention embedding and fine-grained feature enhancement. Knowledge-Based Systems, 284: 111265

Sukchotrat T, Kim S B, Tsung F (2009). One-class classification-based control charts for multivariate process monitoring. IIE Transactions, 42(2): 107–120

Sulem D, Kenlay H, Cucuringu M, Dong X (2024). Graph similarity learning for changepoint detection in dynamic networks. Machine Learning, 113(1): 1–44

Sun R, Tsung F (2003). A kernel-distance-based multivariate control chart using support vector methods. International Journal of Production Research, 41(13): 2975–2989

Sundararajan M, Taly A, Yan Q (2017). Axiomatic attribution for deep networks. In: Proceedings of the International Conference on Machine Learning, PMLR: 3319–3328

Sung C L, Wang W, Cakoni F, Harris I, Hung Y (2024). Functional-input Gaussian processes with applications to inverse scattering problems. Statistica Sinica, 34: 1–20

Takahashi H, Iwata T, Kumagai A, Yamanaka Y (2024). Deep positive-unlabeled anomaly detection for contaminated unlabeled data. Preprint at arXiv. arXiv:2405.18929

Tan M H (2018). Gaussian process modeling of a functional output with information from boundary and initial conditions and analytical approximations. Technometrics, 60(2): 209–221

Tan M H Y (2019). Gaussian process modeling of finite element models with functional inputs. SIAM/ASA Journal on Uncertainty Quantification, 7(4): 1133–1161

Tang J, Lin X, Zhao F, Chen X (2024). Process quality control through Bayesian optimization with adaptive local convergence. Chemical Engineering Science, 293: 120039

Team G, Anil R, Borgeaud S, Alayrac J B, Yu J, Soricut R, Schalkwyk J, Dai A M, Hauth A, Millican K, et al (2023). Gemini: A family of highly capable multimodal models. Preprint at arXiv. arXiv:2312.11805

Teufel F, Stahlhut C, Ferkinghoff-Borg J (2024). Batched energy-entropy acquisition for Bayesian optimization. In: Proceedings of Advances in Neural Information Processing Systems, Curran Associates, Inc., 37: 93118–93170

Thebelt A, Tsay C, Lee R, Sudermann-Merx N, Walz D, Shafei B, Misener R (2022a). Tree ensemble kernels for Bayesian optimization with known constraints over mixed-feature spaces. In: In: Proceedings of Advances in Neural Information Processing Systems, Curran Associates, Inc., 35: 37401–37415

Thebelt A, Tsay C, Lee R M, Sudermann-Merx N, Walz D, Tranter T, Misener R (2022b). Multi-objective constrained optimization for energy applications via tree ensembles. Applied Energy, 306: 118061

Thuy A, Benoit D F (2025). Fast and reliable uncertainty quantification with neural network ensembles for industrial image classification. Annals of Operations Research, 353(2): 517–543

Tian Y, Shen J, Wang A, Li Z, Huang X (2024). Data augmentation and fault diagnosis for imbalanced industrial process data based on residual Wasserstein generative adversarial network with gradient penalty. Journal of Chemometrics, 38(12): e3624

Toumba R N, Eboke A, Tsimi G O, Kombé T (2024). Uncertainty quantification in industrial systems using deep gaussian process for accurate degradation modeling. IEEE Access : Practical Innovations, Open Solutions, 12: 164576–164587

Truong C, Oudre L, Vayatis N (2020). Selective review of offline change point detection methods. Signal Processing, 167: 107299

van den Burg G J J, Williams C K I (2020). An evaluation of change point detection algorithms. Preprint at arXiv. arXiv:2003.06222

van Hoof J, Vanschoren J (2021). Hyperboost: Hyperparameter optimization by gradient boosting surrogate models. Preprint at arXiv. arXiv:2101.02289

Vellanki P, Rana S, Gupta S, Leal D R d C, Sutti A, Height M, Venkatesh S (2019). Bayesian functional optimisation with shape prior. In: Proceedings of the AAAI Conference on Artificial Intelligence, 33: 1617–1624

Vien N A, Zimmermann H, Toussaint M (2018). Bayesian functional optimization. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1): 1–8

Vincent M C, Maier M, Wegener K (2025). Optimizing process parameters in manufacturing to reduce carbon footprint with contextual Bayesian optimization. International Journal of Advanced Manufacturing Technology, 139(7-8): 3381–3390

Vishwakarma R, Rezaei A (2024). Uncertainty-aware hardware Trojan detection using multimodal deep learning. In: 2024 Design, Automation & Test in Europe Conference & Exhibition (DATE), IEEE: 1–6

Wachter S, Mittelstadt B, Russell C (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2): 841–887

Wan J, Xia N, Yin Y, Pan X, Hu J, Yi J (2024). TCDformer: A trans-

former framework for non-stationary time series forecasting based on trend and change-point detection. Neural Networks, 173: 106196

Wang C H, Guo R S, Chiang M H, Wong J Y (2008). Decision tree based control chart pattern recognition. International Journal of Production Research, 46(17): 4889–4901

Wang F, Li W, Madrid Padilla O H, Yu Y, Rinaldo A (2026). Multilayer random dot product graphs: estimation and online change point detection. Journal of the Royal Statistical Society Series B: Statistical Methodology, 88(1): 282–312

Wang F, Zhai Z, Zhao Z, Di Y, Chen X (2024). Physics-informed neural network for lithium-ion battery degradation stable modeling and prognosis. Nature Communications, 15(1): 4332

Wang H, Wang Z, Du M, Yang F, Zhang Z, Ding S, Mardziel P, Hu X (2020). ScoreCAM: Score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops: 24–25

Wang J, Liu L (2024). A new multivariate control chart based on the isolation forest algorithm. Quality Engineering, 36(2): 390–406

Wang L, Lin P, Li Y, Luo H, Zhang L (2025a). Physics-informed online deep learning for advanced control of shield tail clearance in tunnel construction. Frontiers of Engineering Management, 12(4): 828–853

Wang P, Qu H, Zhang Q, Xu X, Yang S (2023b). Production quality prediction of multistage manufacturing systems using multi-task joint deep learning. Journal of Manufacturing Systems, 70: 48–68

Wang R, Chen H, Guan C (2023c). A self-supervised contrastive learning framework with the nearest neighbors matching for the fault diagnosis of marine machinery. Ocean Engineering, 270: 113437

Wang S, Ou W, Liu Z, Du B, Wang R (2025b). Competitive multi-task Bayesian optimization with an application in hyperparameter tuning of additive manufacturing. Expert Systems with Applications, 262: 125618

Wang S, Teng Y, Perdikaris P (2021). Understanding and mitigating gradient flow pathologies in physics-informed neural networks. SIAM Journal on Scientific Computing, 43(5): A3055–A3081

Wang T, Wang Y, Zhou J, Peng B, Song X, Zhang C, Sun X, Niu Q, Liu J, Chen S, Chen K, Li M, Feng P, Bi Z, Liu M, Zhang Y, Fei C, Yin C H, Yan L K (2025c). From aleatoric to epistemic: exploring uncertainty quantification techniques in artificial intelligence. Preprint at arXiv. arXiv:2501.03282

Wang X, Jin Y, Schmitt S, Olhofer M (2023d). Recent advances in Bayesian optimization. ACM Computing Surveys, 55(13s): 1–36

Wang X, Li Y, Yue X, Wu J (2025d). Nonstationary and sparsely-correlated multioutput gaussian process with spike-and-slab prior. INFORMS Journal on Data Science, 4(2): 114–132

Wang X, Wang C, Song X, Kirby L, Wu J (2023e). Regularized multi-output Gaussian convolution process with domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(5): 6142–6156

Wang Z, Xing W, Kirby R, Zhe S (2022). Physics informed deep kernel learning. In: Proceedings of the 25th International Conference on Artificial Intelligence and Statistics, PMLR: 1206–1218

Weese M L, Martinez W G, Jones-Farmer L A (2017). On the selection of the bandwidth parameter for the k-chart. Quality and Reliability Engineering International, 33(7): 1527–1547

Weibull W (1951). A statistical distribution function of wide applicability. Journal of Applied Mechanics, 18(3): 293–297

Westermann P, Evins R (2021). Using Bayesian deep learning approaches for uncertainty-aware building energy surrogate models. Energy and AI, 3: 100039

Wiegand M, Prots A, Meyer M, Schmidt R, Voigt M, Mailach R (2025). Robust design optimization of a compressor rotor using recursive cokriging based multi-fidelity uncertainty quantification and multi-fidelity optimization. Journal of Turbomachinery, 147(6): 061009

Wilson A G, Hu Z, Salakhutdinov R, Xing E P (2016a). Deep kernel learning. In: Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, PMLR: 370–378

Wilson A G, Hu Z, Salakhutdinov R R, Xing E P (2016b). Stochastic variational deep kernel learning. In: Proceedings of Advances in Neural Information Processing Systems, Curran Associates, Inc., 29: 1–9

Woodall W H, Montgomery D C (1999). Research issues and ideas in statistical process control. Journal of Quality Technology, 31(4): 376–386

Wu J, Sun Y N, Song Y T, Liu L L, Gao Z G, Qin W (2025). Uncertainty-aware Bayesian neural network with SHAP interpretability for data-driven assembly quality prediction in complex manufacturing systems. Advanced Engineering Informatics, 68: 103730

Xanthopoulos P, Razzaghi T (2014). A weighted support vector machine method for control chart pattern recognition. Computers & Industrial Engineering, 70: 134–149

Xia J, Huang R, Chen Z, He G, Li W (2023). A novel digital twin-driven approach based on physical-virtual data fusion for gearbox fault diagnosis. Reliability Engineering & System Safety, 240: 109542

Xiao Q, Mandal A, Lin C D, Deng X (2021). EzGP: Easy-to-interpret Gaussian process models for computer experiments with both quantitative and qualitative factors. SIAM/ASA Journal on Uncertainty Quantification, 9(2): 333–353

Xu J, Lv H, Zhuang Z, Lu Z, Zou D, Qin W (2019). Control chart pattern recognition method based on improved one-dimensional convolutional neural network. IFAC-PapersOnLine, 52(13): 1537–1542

Xu R, Song Z, Wu J, Wang C, Zhou S (2025). Change-point detection with deep learning: A review. Frontiers of Engineering Management, 12(1): 154–176

Xu Y, Kohtz S, Boakye J, Gardoni P, Wang P (2023). Physics-informed machine learning for reliability and systems safety applications: State of the art and challenges. Reliability Engineering & System Safety, 230: 108900

Xu Y, Ma X, Wang X, Wang J, Tang G, Ji Z (2024). Unified feature learning network for few-shot fault diagnosis. Neurocomputing, 598: 128035

Xu Z, Saleh J H (2021). Machine learning for reliability engineering and safety applications: Review of current status and future opportunities. Reliability Engineering & System Safety, 211: 107530

Yang C, Wang X, Lu Y, Liu H, Le Q V, Zhou D, Chen X (2023). Large language models as optimizers. In: Proceedings of the 12th

International Conference on Learning Representations, ICLR: 1–41

Yang H, Kan C, Krall A, Finke D (2020). Network modeling and Internet of things for smart and connected health systems—a case study for smart heart health monitoring and management. IISE Transactions on Healthcare Systems Engineering, 10(3): 159–171

Yang S, Yee K (2024). Towards reliable uncertainty quantification via deep ensemble in multi-output regression task. Engineering Applications of Artificial Intelligence, 132: 107871

Yang Y, Ming D, Guillas S (2025). Distribution of deep Gaussian process gradients and sequential design for simulators with sharp variations. Preprint at arXiv. arXiv:2503.16027

Yazdi F, Bingham D, Williamson D (2024). Deep Gaussian process emulation and uncertainty quantification for large computer experiments. Preprint at arXiv. arXiv:2411.14690

Yeganeh A, Sogandi F, Shongwe S C (2025). Autoencoders for monitoring Poisson-dependent process steps based on state space representation. Computers & Industrial Engineering, 207: 111258

Yin Y, Wang Y, Xu B, Li P (2024). ADO-LLM: Analog design Bayesian optimization with in-context learning of large language models. In: Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design: 1–9

Zan T, Jia X, Guo X, Wang M, Gao X, Gao P (2025). Research on variable-length control chart pattern recognition based on sliding window method and SECNN-BiLSTM. Scientific Reports, 15(1): 5921

Zhan Z, Zhou J, Xu B (2022). Fabric defect classification using prototypical network of few-shot learning algorithm. Computers in Industry, 138: 103628

Zhang J, Rangaiah G P, Dong L, Samavedham L (2024a). A novel fault diagnosis framework for industrial production processes based on causal network inference. Industrial & Engineering Chemistry Research, 63(21): 9471–9488

Zhang W, Xiao G, Gen M, Geng H, Wang X, Deng M, Zhang G (2024b). Enhancing multi-objective evolutionary algorithms with machine learning for scheduling problems: recent advances and survey. Frontiers in Industrial Engineering, 2: 1337174

Zhang X, Zou Y, Li S (2022). Bayesian neural network with efficient priors for online quality prediction. Digital Chemical Engineering, 2: 100008

Zhang Y, Edgar T F (2007). On-line batch process monitoring using modified dynamic batch PCA. In: 2007 American Control Conference, IEEE: 2551–2556

Zhang Y, Tao S, Chen W, Apley D W (2020). A latent variable approach to gaussian process modeling with qualitative and quantitative factors. Technometrics, 62(3): 291–302

Zhang Y, Tino P, Leonardis A, Tang K (2021). A survey on neural network interpretability. IEEE Transactions on Emerging Topics in Computational Intelligence, 5(5): 726–742

Zhao Z, Zhang Q, Yu X, Sun C, Wang S, Yan R, Chen X (2021). Applications of unsupervised deep transfer learning to intelligent fault diagnosis: A survey and comparative study. IEEE Transactions on Instrumentation and Measurement, 70: 1–28

Zheng Q, Xia X, Zou X, Dong Y, Wang S, Xue Y, Shen L, Wang Z, Wang A, Li Y (2023). Codegeex: A pre-trained model for code generation with multilingual benchmarking on humaneval-x. In: Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining: 5673–5684

Zheng X, Aragam B, Ravikumar P K, Xing E P (2018). DAGs with no tears: Continuous optimization for structure learning. In: Proceedings of Advances in neural information processing systems, Curran Associates, Inc., 31: 1–12

Zhong D, Xia Z, Zhu Y, Duan J (2023). Overview of predictive maintenance based on digital twin technology. Heliyon, 9(4): e14534

Zhou S, Ding Y, Chen Y, Shi J (2003). Diagnosability study of multistage manufacturing processes based on linear mixed-effects models. Technometrics, 45(4): 312–325

Zhou X, Tan J, Yu J, Gu X, Jiang T (2024). Online robust parameter design using sequential support vector regression based Bayesian optimization. Journal of Mathematical Analysis and Applications, 540(2): 128649

Zhu H, Balsells-Rodas C, Li Y (2023). Markovian Gaussian process variational autoencoders. In: Proceedings of the 40th International Conference on Machine Learning, PMLR: 42938–42961

Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q (2021). A comprehensive survey on transfer learning. Proceedings of the IEEE, 109(1): 43–76