

Residual-Based Monitoring with Incremental Model Update for CNC Milling Processes

Robin Ströbel^{1,†}, Samuel Deucker¹, Manuel Raimann¹, Hafez Kader², Alexander Puchta¹, Benjamin Noack², and Jürgen Fleischer¹

¹wbk Institute of Production Science, Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany

²AMS - Autonomous Multisensor Systems, Otto von Guericke University Magdeburg, 39106 Magdeburg, Germany

[†]Corresponding author: Robin Ströbel (e-mail: robin.stroebel@kit.edu)

October 16, 2026

Abstract

Flexible CNC milling requires process monitoring systems that remain reliable under changing workpieces, materials, and machine conditions. This paper proposes a monitoring framework that combines signal-level process prediction, residual-based anomaly detection, and data-centric incremental model maintenance. A hybrid predictive model generates axis- and spindle-specific current references from kinematic and physically informed process features, while interpretable statistical detectors operate on the resulting residuals to identify process anomalies. Two experiments are conducted to first define the monitoring model and second to quantify the integration of an incremental learning with a density-based data selection approach, active-learning-based sampling, and physics-informed dataset optimization. First, established statistical monitoring methods are compared for deployment within the residual domain, where Statistical Process Control achieves the best overall trade-off between detection quality and stability. Second, the full monitoring pipeline is validated under progressive domain shift in an incremental setting across two machines. The results show that all adaptive variants significantly outperform the static baseline, confirming the necessity of maintaining models in flexible production. The adaptive variants differ more in their temporal progression, stability, and data efficiency than in their aggregated performance. Overall, the study demonstrates that residual-based monitoring combined with data-centric adaptation provides an interpretable and deployment-oriented solution for robust CNC process monitoring under evolving production conditions.

Keywords: CNC, milling, incremental model update, process monitoring, anomaly detection

1 Introduction

The rapid evolution of modern manufacturing systems is driven by two converging trends: an increasing shortage of skilled labor [1, 2] and a simultaneous rise in product variety and demand volatility [3]. While production environments are expected to deliver highly customized components with short lead times, they must do so under stringent cost and quality constraints [4]. This structural tension places growing pressure on manufacturing systems to operate reliably, autonomously, and with minimal human intervention. In this context, CNC milling is central in small-batch and high-mix manufacturing. Process disturbances such as tool wear, chattering, or material defects can lead to reduced quality, rejects, or machine downtime. Consequently, effective process monitoring is essential not only for quality assurance but also for maintaining productivity and resource efficiency [5].

Current process monitoring systems in milling are mainly rule-based or rely on static thresholds applied directly to raw sensor signals [5]. While these approaches can be effective under stable operating conditions, they struggle in highly flexible environments [5]. As a result, many existing solutions lack the adaptability required for modern, variant-rich production settings. Be-

yond rule-based approaches, traditional machine learning methods such as isolation forests or long short-term memories (LSTM) have been applied to process monitoring using hand-crafted features extracted from sensor signals (e.g., time-, frequency-, or time-frequency-domain descriptors) [6]. To further enhance the reliability, physics-informed learning has been introduced into intelligent process monitoring [7]. While these approaches improve sensitivity compared to fixed thresholds, they strongly depend on feature engineering and representative training data, and lack transparency. Consequently, their performance often degrades under domain shifts caused by changes in tools, materials, machines, or process parameters, requiring frequent re-training and limiting scalability in highly variant production environments [8]. Additionally, as transparency is key in manufacturing, old rule-based or static thresholds are still commonly applied in current monitoring systems [9].

To address these limitations while retaining the transparency required in manufacturing, a model-based residual monitoring strategy can be adopted. Predictive process models have gained increasing attention and offer the opportunity to estimate a specific sensor signal based on process parameters, yielding a physically and operationally grounded reference signal [10, 11]. Con-

sequently, the process model can be interpreted as a compressor that creates a representation of the input signals. Deviations from nominal behavior are then captured as residuals between predicted and measured signals, which suppress expected variability and highlight fault-induced effects. Crucially, this residual representation enables the use of well-established, interpretable process monitoring techniques, such as statistical control and threshold-based detectors [12], on a signal that is substantially more robust to changing tools, materials, machines, and process settings, thereby combining flexibility with traceability [9].

However, the effectiveness of such a residual-based monitoring system depends directly on the quality and representativeness of the underlying predictive model. In flexible CNC production, training data is collected incrementally, affected by distribution shifts, and often imbalanced or locally sparse in operating regions relevant for reliable monitoring [13]. A robust condition-monitoring approach therefore requires not only an interpretable residual-based detector, but also a data-centric strategy for continuously maintaining and improving the reference model through representative data selection, targeted adaptation, and physics-consistent dataset enrichment.

In this paper, we first investigate different statistical methods to perform residual-based anomaly detection in a real world scenario. The finding is then integrated into the proposed holistic framework for condition monitoring in CNC milling to combine residual-based anomaly detection with data-centric incremental model maintenance (see figure 1). Consequently, the framework integrates a predictive monitoring model

with data-centric incremental learning (DcIL) through density-based dataset balancing [13], active-learning-driven adaptation, and physics-informed dataset optimization (PiDO) [14]. The core contribution of this paper is the unified investigation of these elements within one monitoring architecture, with particular emphasis on the monitoring model as the central basis for robust residual generation under changing manufacturing conditions.

The remainder of this paper is organized as follows. Section 2 reviews the state of the art in adaptive and residual-based process monitoring. Section 3 presents the proposed monitoring framework and its methodological components. Section 4 describes the experimental setup, including the identification of the monitoring model and the end-to-end validation under domain shift. Section 5 reports the experimental results. Section 6 discusses the findings and their implications for flexible CNC process monitoring. Finally, Section 7 concludes the paper and outlines directions for future work.

2 State of the Art

The state of machining processes is determined by multiple interacting physical phenomena. Consequently, a wide range of process anomalies can occur during milling, including chatter, tool wear, thermal effects, and material-related instabilities. Each anomaly exhibits distinct signal characteristics [12, 15]. Changes in tools, materials, and cutting parameters result in altered behavior of the nominal machine signals, further complicating reliable monitoring [16]. While process state can be inferred indirectly from sensor signals, such as

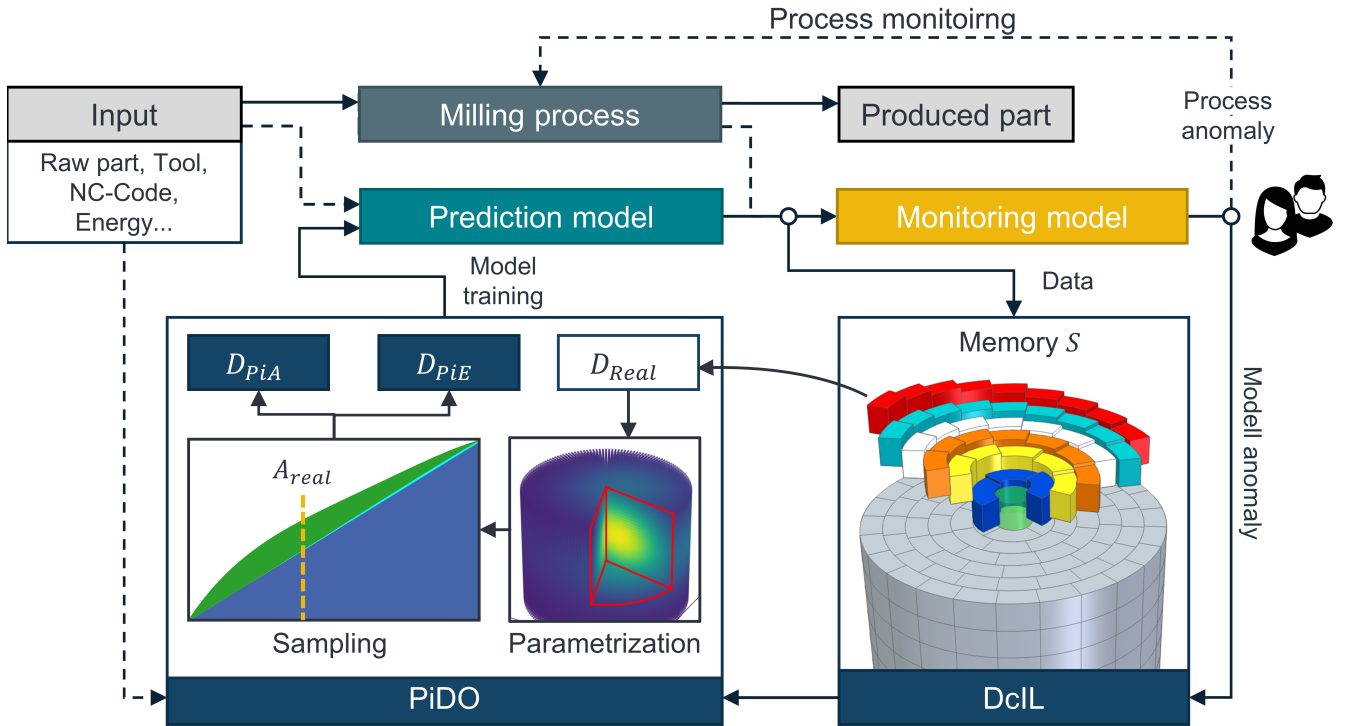


Figure 1: Overview and interaction between the process, the prediction model, the monitoring model, and the dataset management via DcIL and PiDO

spindle and axis currents, the relationship between the measured signals and the actual process conditions is neither explicit nor trivial [10].

Existing research addresses this challenge by using physics-based, data-driven, and hybrid approaches, often relying on specialized sensor setups or fault-specific algorithms, which limits transferability and scalability in flexible production environments [10]. The following section reviews current residual-based anomaly detection approaches and analyzes the current evolution from static to adaptive monitoring systems.

2.1 Current approaches in residual-based anomaly detection

Residual-based anomaly detection models the expected system response and flags anomalies based on the deviation between predicted and measured signals. By operating on reconstruction or prediction errors rather than raw sensor streams, residual-based methods suppress nominal process variability while amplifying abnormal behavior, which is particularly beneficial for complex, non-stationary time series [17].

In technical applications, residuals have repeatedly proven to be an effective monitoring signal. In aerospace, Hundman et al. show that prediction residuals from LSTM models provide a compact representation of nominal spacecraft dynamics and enable robust detection when paired with adaptive, nonparametric thresholding [18]. For generic time-series monitoring, Munir et al. introduce DeepAnT, where a convolutional predictor forecasts future values and anomalies are identified via prediction errors, demonstrating strong unsupervised performance in data-scarce regimes [19]. Hybrid approaches further combine classical and deep predictors to separate linear and nonlinear dynamics before evaluating residuals, improving robustness under drifting operating conditions [20].

Within manufacturing and robotics, residual-based monitoring is already adopted, but frequently implemented with complex learning architectures. In robotic execution monitoring, Park et al. use an LSTM-based variational autoencoder to reconstruct multimodal sensory streams and detect anomalous executions through reconstruction residuals, remaining sensitive to execution failures while tolerating benign task variability [21]. For real-time deployment, Kang et al. propose a sparse latent-space autoencoder enabling millisecond-level inference on embedded robotic hardware [20]. From a control-oriented viewpoint, Alamir and Clavel derive residual generators from implicit sensor relationships using sparse piecewise polynomials, achieving tightly bounded residuals at low computational cost [22]. In machine tool monitoring, Pittino et al. combine residual-like deviations with adaptive control-chart decision rules to enable robust in-production detection without frequent recalibration [23]. For CNC systems, Demetgul et al. apply transfer-learned autoencoders on motor-current signals, where reconstruction residuals indicate axis misalignment without additional sensing [24]. Overall, these works confirm residuals as a unify-

ing monitoring signal for technical systems, while also motivating lightweight and interpretable residual-based pipelines for flexible CNC milling.

2.2 Evolution from static to adaptive monitoring systems

Monitoring in machining has evolved from static, rule-based limit checks toward adaptive reference modeling that better reflects highly variable operating conditions [5, 25, 9]. Classical statistical process control and fixed thresholds remain attractive due to their transparency and ease of validation, but they implicitly assume quasi-stationary behavior and therefore degrade when tools, materials, and parameter sets change frequently [9, 5]. A key research trend is therefore to increase the fidelity of the reference signal: from a single reference per part, implemented via classical SPC, to segment- and subsegment-specific references, and ultimately to reference signals resolved at the sampling-time level [5, 25]. This shift is motivated by the fact that many anomalies are temporally local and tied to specific machining operations, so segment-wise or signal-level references reduce confounding variability while preserving traceability [5, 25]. As reference generation shifts from static limits to signal-level predictive models, the monitoring problem increasingly depends on the validity and adaptability of the underlying model itself, since changes in geometry, cutting conditions, or machine behavior can otherwise be misinterpreted as process anomalies.

In parallel, monitoring increasingly relies on machine-internal signals that are available at scale, particularly current and power signals, because they correlate with process forces, torque, and tool wear and can be acquired without additional sensors [26, 27, 10]. However, direct thresholding on these signals can exhibit limited sensitivity under small deviations and strong dependence on the operating point [26]. To address this, recent approaches combine domain knowledge with data-driven models, including physics-informed learning, to obtain more robust references while maintaining interpretability [26, 28].

Residual-based monitoring provides the bridging paradigm in which a predictive model generates an expected signal, and established transparent statistical detectors operate on residuals to suppress nominal variability and highlight abnormal behavior [12, 9]. Although the trend toward signal-level references in machining monitoring is evident, important challenges remain unresolved. Adaptive reference modeling is increasingly recognized as necessary in flexible production environments. Yet, robust concepts for maintaining the validity of predictive reference signals under changing process conditions are still limited [5, 25]. This issue is particularly relevant for machine-internal signals such as axis currents whose informative value for monitoring is well established but whose direct interpretation strongly depends on the operating point and the surrounding process context [26]. The present work addresses this challenge by transferring established monitoring principles into the residual domain through signal-level refer-

ence prediction while explicitly considering the reliability of the underlying monitoring model.

2.3 Research Gap

As described, machining process monitoring is moving beyond static thresholding toward references that adapt to changing tools, materials, and process parameters, because fixed limits and part-level baselines degrade in high-mix environments [5, 25, 9]. Classical statistical methods remain attractive in this context due to their transparency and industrial acceptance, and recent work has increased flexibility by forming more context-dependent references [5, 25]. However, the transition from static limits to predictive and signal-level references introduces a new challenge, since monitoring performance increasingly depends on the validity and adaptability of the underlying reference model under changing operating conditions [29, 30]. At the same time, CNC monitoring increasingly relies on machine-internal signals such as current and power, because these signals correlate with relevant process phenomena and can be acquired without additional sensor integration [26]. Yet, their direct interpretation remains strongly dependent on the operating point and process context, which limits the robustness of direct thresholding and fixed-rule approaches [26].

In parallel, residual-based monitoring replaces static references with signal-level expectations from predictive models. It performs detection on residuals, yielding dynamic references that can suppress nominal variability while preserving interpretability through established statistical detectors [12, 9]. Although residual-based charting for autocorrelated and multivariate processes has been studied [31, 17], and recent work has highlighted the potential of physics-informed and hybrid modeling strategies for condition monitoring [28, 32], existing approaches still address predictive modeling, residual-based detection, and model adaptation largely in isolation. In particular, there is still a lack of holistic condition-monitoring approaches for CNC milling that systematically combine signal-level predictive references, interpretable residual-based anomaly detection, and explicit mechanisms to maintain model validity under domain shifts, such as unseen workpiece geometries, changing process conditions, and evolving training-data regimes.

This gap motivates the contribution of this paper, which proposes a holistic condition-monitoring framework for CNC milling and, at the same time, examines the individual contributions of its main components in a structured manner. The proposed approach jointly considers predictive reference generation, interpretable residual-based monitoring, and adaptive data-centric model maintenance to identify which extensions provide measurable benefit under domain shift and which do not. In doing so, the study also makes explicit the strong interaction between the predictive and monitoring models, since the reliability of anomaly detection depends directly on the quality, validity, and adaptability of the generated signal-level reference. Concretely, the

paper investigates a residual-based monitoring architecture centered on a predictive current-signal model and evaluates how adaptive data-centric mechanisms support robust anomaly detection under changing production conditions. In this way, the work extends established statistical monitoring logic into a more flexible and comprehensive framework that remains transparent and compatible with industrial validation requirements.

3 Methodology

3.1 Derivation of monitoring model

The core idea is to couple the physical process with a hybrid predictive model that predicts the current for each axis [10]. The monitoring model then compares measured signals with the reference and evaluates the resulting residuals using established statistical detectors, yielding a transparent decision on whether a relevant process deviation is present.

In detail, the process model presented in [10] forms the basis of the monitoring system, accurately predicting the expected current signals of the machine axes based on the current process parameters. It combines kinematic machine data, such as axis position, velocity, and acceleration, with soft sensors to calculate process-relevant physical variables. These include cutting forces, as defined by the Kienzle model, and the material removal rate (MRR), which is determined by voxel-based simulation. This derives signals that would otherwise be difficult or expensive to measure. These physically informed features extend the purely data-driven approach and ensure improved generalization [10].

In the case of a three-axis milling machine, the hybrid model structure is given by axis- and spindle-specific ML models

$$Out_n = P_n(\overline{In}) = \tilde{I}_n, \text{ where } n \in \{x, y, z, sp\} \quad (1)$$

which maps the input vector

$$\overline{In} = [\dot{\mathbf{x}}^T \quad \ddot{\mathbf{x}}^T \quad \dot{\varphi} \quad \ddot{\varphi} \quad \mathbf{F}^T \quad M_{sp} \quad \text{MRR}]^T \quad (2)$$

to the respective target $\tilde{I}_x$, $\tilde{I}_y$, $\tilde{I}_z$, and $\tilde{I}_{sp}$.

The model has demonstrated sufficient prediction accuracy across different machining scenarios and robustness against natural process fluctuations [10]. This is an essential requirement, since the quality of the downstream monitoring model depends directly on the quality of the calculated residuals. Insufficient prediction accuracy would increase false-positive rates in anomaly detection and thereby compromise the overall system.

The monitoring model M analyses the residuals r_i generated by the process model P_i in order to identify process anomalies, as shown in figure 2. A central requirement for M is the multivariate analysis of residuals across the machine’s X-, Y-, and Z-axes, as well as the main spindle. This is because process anomalies occur as correlated patterns rather than isolated deviations

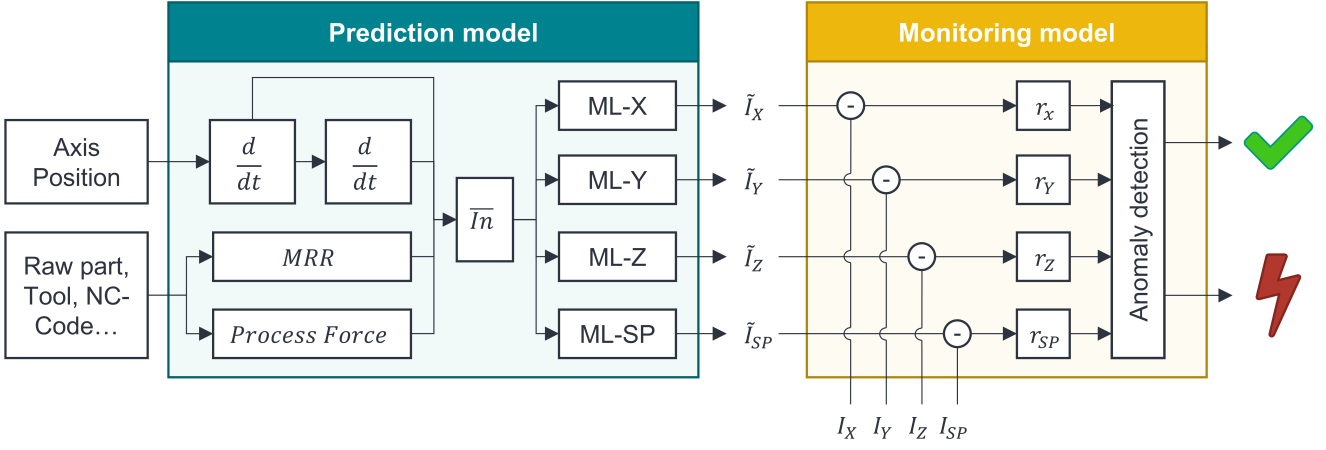


Figure 2: Interaction between the Prediction model [10] and the Monitoring model

in individual signals. The simultaneous evaluation of all axes thus enables the detection of complex anomaly types [33]. The monitoring model must achieve high sensitivity to reliably detect critical process events while maintaining a controllable false-positive rate, as excessive alarms reduce operator trust and negatively impact productivity [34].

3.2 Residual-based monitoring

For each time point t and each axis $i \in \{X, Y, Z, \text{spindle}\}$ the residuals are calculated as

$$r^{(i)}(t) = I_{\text{real}}^{(i)}(t) - I_{\text{pred}}^{(i)}(t), \quad (3)$$

where $I_{\text{real}}^{(i)}(t)$ denotes the measured current and $I_{\text{pred}}^{(i)}(t)$ denotes the prediction of the process model M_n . The residuals $r^{(i)}(t)$ serve as input data for the monitoring model (see figure 2). After point-by-point anomaly scoring, the monitoring model updates anomaly windows online while the process is running. At each sampling instant t_i , the binary decision $s_i \in \{0, 1\}$ is incorporated into a small state machine that either opens a new window when an anomaly starts, keeps the current window active, or closes it once the anomaly has ended. To improve robustness against brief measurement drops or classifier flicker, short non-anomalous interruptions within an active window are tolerated up to a patience of p_{max} consecutive 0 labels. A window is reported as a valid anomaly only when its duration exceeds the minimum duration Δt_{min} . Algorithm 1 formalizes this streaming window update, enabling anomaly windows to be emitted in real time without requiring access to the full signal history.

Fusion across all axes is performed at each time step by a majority decision. A process deviation is flagged if at least n_{min} axes report $s_i = 1$. The threshold n_{min} is a hyperparameter of the monitoring model and is calibrated during training.

Algorithm 1 Anomaly window identification

Require: $A = \{(t_i, s_i)\}_{i=1}^n, s_i \in \{0, 1\}$
Require: Δt_{min} (min. duration), p_{max} (max. consecutive zeros)

```

1: function IDENTIFYANOMALYWINDOWS( $A, \Delta t_{\text{min}}, p_{\text{max}}$ )
2:    $W \leftarrow \emptyset; t_s \leftarrow \perp; p \leftarrow 0$ 
3:   for  $i = 1$  to  $n$  do
4:     if  $s_i = 1$  and  $t_s = \perp$  then
5:        $t_s \leftarrow t_i$ 
6:     end if
7:     if  $t_s \neq \perp$  then
8:        $p \leftarrow (s_i = 1) ? 0 : p + 1$ 
9:       if  $p \geq p_{\text{max}}$  then
10:        if  $t_i - t_s \geq \Delta t_{\text{min}}$  then
11:           $W \leftarrow W \cup \{(t_s, t_i)\}$ 
12:        end if
13:         $t_s \leftarrow \perp; p \leftarrow 0$ 
14:      end if
15:    end if
16:  end for
17:  if  $t_s \neq \perp$  and  $t_n - t_s \geq \Delta t_{\text{min}}$  then
18:     $W \leftarrow W \cup \{(t_s, t_n)\}$ 
19:  end if
20:  return  $W$ 
21: end function

```

3.3 Established Models for Anomaly Detection

The residual signal $r^{(i)}(t)$ transforms the complex axis-specific sensor signal into a simplified one-dimensional representation relative to the predicted reference signal. The monitoring model operates on $r^{(i)}(t)$ and aims to identify deviations from the expected process state. In this work, the detector layer is not implemented as a set of fixed-limit charts. Instead, all statistical methods are used as adaptive residual-based variants that estimate their reference statistics online from the residual stream. This design reflects the non-stationary nature of the monitoring task and allows the detectors to adapt to changing local residual characteristics while remain-

ing interpretable and computationally lightweight.

The implemented Z-Score detector uses rolling standardization of the residuals rather than a fixed global reference. For each axis, a local mean and local standard deviation are estimated over a sliding window, and the absolute standardized residual magnitude is computed as

$$z_t^{(i)} = \left| \frac{r_t^{(i)} - \mu_t^{(i,w)}}{\sigma_t^{(i,w)}} \right|, \quad (4)$$

where $\mu_t^{(i,w)}$ and $\sigma_t^{(i,w)}$ denote the rolling mean and rolling standard deviation of the residual signal for axis i . The resulting axis-wise scores are combined into a multiaxis anomaly score

$$a_t = \prod_i z_t^{(i)}, \quad (5)$$

and an anomaly is flagged when a_t exceeds a tunable global threshold. The detector should therefore be interpreted as a rolling multiaxis Z-Score detector rather than as a classical fixed-threshold Z-Score chart. Overall, this approach remains computationally efficient and highly interpretable while providing adaptive scale normalization across axes and operating regions.

To improve sensitivity to gradual deviations, an Exponentially Weighted Moving Average-based detector is considered [35]. In the implemented variant, exponentially weighted moving statistics are used to construct an adaptive local reference for each residual signal. Specifically, an EWMA-based local mean and an EWMA-based local standard deviation are estimated online,

$$m_t^{(i)} = \text{EWMA}(r_t^{(i)}), \quad s_t^{(i)} = \text{EWMSD}(r_t^{(i)}), \quad (6)$$

and a point is flagged as anomalous when the residual exceeds either a variance-scaled band or an absolute-offset band around the adaptive centerline,

$$|r_t^{(i)} - m_t^{(i)}| > k s_t^{(i)} \quad \text{or} \quad |r_t^{(i)} - m_t^{(i)}| > \delta_{\text{abs}}, \quad (7)$$

where k is a tunable standard-deviation multiplier and δ_{abs} is a tunable absolute offset. The EWMA detector is therefore used here as an adaptive-band method rather than as a classical EWMA chart with analytically derived static control limits. This implementation preserves the core strength of EWMA, namely sensitivity to small but persistent residual shifts, while adapting the reference locally to changing residual behavior.

Statistical Process Control, SPC, comprises a family of established methods for process monitoring, with control charts as the central tool [36]. In the present work, SPC is implemented as an adaptive control-chart scheme on the residual signals rather than as a static Shewhart chart with fixed limits. For each axis, the centerline and dispersion are estimated online either by rolling statistics or by exponentially weighted statistics. Based on these local statistics, Western Electric Rules 1 to 4 are evaluated. The resulting axis-wise binary alarms are then fused by majority logic, requiring a minimum number of affected axes before a multiaxis process deviation

is reported. The SPC detector should therefore be interpreted as an adaptive SPC variant with moving control limits and rule-based multiaxis fusion.

In addition, a CUSUM-inspired detector is evaluated. In contrast to a canonical tabular CUSUM chart, the implemented variant first computes the cumulative sum of the residual signal and then derives rolling statistics on this cumulative trace. The current residual is subsequently compared with variance-scaled and absolute-offset limits defined relative to the cumulative residual reference. Accordingly, this detector is a cumulative-residual monitoring variant (CUMSUM) inspired by CUSUM principles.

As shown in figure 3 each method emphasizes different deviation patterns, such as isolated outliers, gradual drifts, or persistent shifts. Therefore, composite monitoring schemes are additionally considered to combine complementary detector principles. Overall, the statistical approaches used in this work provide deterministic behavior, low runtime complexity, and transparent threshold calibration. At the same time, they are limited in capturing strongly nonlinear or high-dimensional dependencies across residual channels. The detector layer is therefore deliberately restricted to methods that remain interpretable, parameterizable, and readily deployable on industrial CNC controllers, where validation and traceability are primary requirements. Table 1 summarizes the methods considered in this study.

Table 1: Overview of monitoring methods used for anomaly and condition detection.

Method	Category	Type	Key characteristic
Z-Score	Statistical	Threshold-based	Outlier detection
EWMA	Statistical	Threshold-based	Trend analysis
CUMSUM	Statistical	Threshold-based	Persistent deviations
SPC	Statistical	Rule-based	Western Electric rules
Composite	Hybrid	Combination	Method combination

3.4 Density-based dataset balancing for incremental learning (DcIL)

[13] proposes a data-centric memory management strategy for incremental learning in axis- and spindle-current prediction for CNC machining. The core idea is to map the original high-dimensional hybrid-model input space to a reduced and interpretable 3D feature space and discretize it into volumetric bins. Incoming samples are inserted into this discretized space, while an adaptive memory retains only a bounded number of samples per bin, controlled by the local sample density. This yields a compact yet representative training set that maintains coverage across operating conditions. When model-induced deviations indicate insufficient information in a region, the method increases the target density locally. It can trigger targeted oversampling, improving performance on newly emerging conditions while keeping storage and retraining costs tractable. Empirical validation across multiple machines, geometries, and process parameters shows that the approach substantially reduces training data volume without sacrificing prediction accuracy. [13]

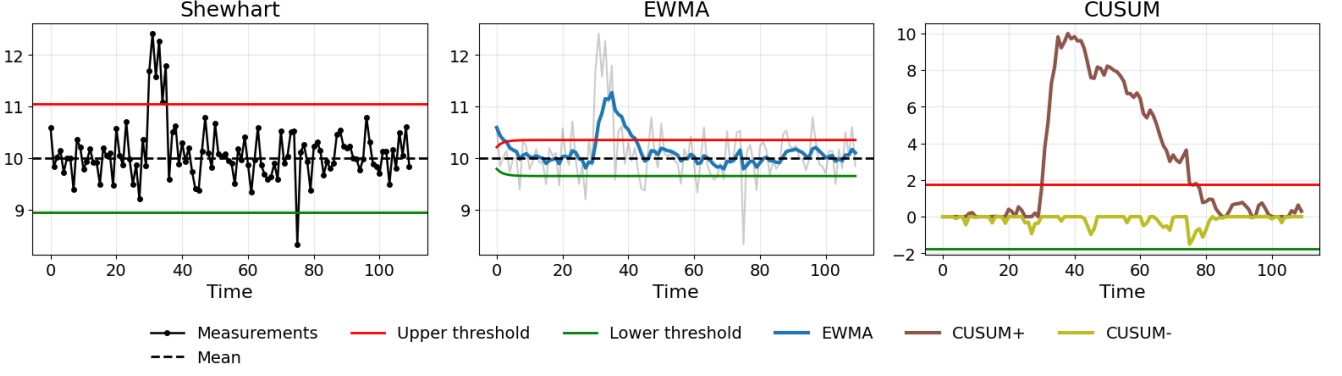


Figure 3: Illustration of the SPC-, EWMA-, and CUSUM-inspired detector principles used on the residual signals.

The reduced feature space is discretized into bins in cylindrical coordinates over feed rate v_c , in-plane direction angle ϕ , and tool intersection AR , i.e., $B_{ijk} = (v_c, \phi, AR)$. The volume of a bin is obtained by integrating over its boundaries,

$$dV = \int_{\zeta_k}^{\zeta_{k+1}} \int_{\xi_i}^{\xi_{i+1}} \int_{\eta_j}^{\eta_{j+1}} v_c d\phi dv_c dAR, \quad (8)$$

which yields the closed-form approximation

$$V_{\text{Bin}} = v_{c,\text{center}} \cdot \Delta\phi \cdot \Delta v_c \cdot \Delta AR. \quad (9)$$

The bin density is then defined as

$$\rho_{\text{Bin}} = \frac{N_{\text{Bin}}}{V_{\text{Bin}}}, \quad (10)$$

where N_{Bin} is the number of stored samples in the bin. By enforcing a target density per bin $\rho_{\text{max}} = 240 \text{ s}^2/\text{mm}^2$, the memory selectively discards redundant samples in high-density regions while preserving samples in sparse, performance-critical regions. Furthermore, an AL-Interface is given by the oversampling to $\rho_{\text{AL}} = 1.5\rho_{\text{max}}$. [13]

3.5 Physics-informed dataset optimization (PiDO)

Physics-informed Dataset Optimization (PiDO) [14] is a data-centric incremental learning approach designed for non-stationary and imbalanced streaming data. The method enriches the training memory with physics-consistent samples in operating regions relevant to upcoming parts. Therefore, a local physical model of the current dynamics is identified for each bin by means of distance-weighted least squares. Based on this model, two complementary modes are employed. Physics-informed augmentation (PiA) generates additional targets for measured inputs using the physical equation. In contrast, physics-informed extrapolation (PiE) samples new inputs from the local feature distribution and computes the corresponding targets through the same model. In this way, PiDO improves dataset coverage in sparse but relevant regions and supports higher predictive accuracy than purely data-driven incremental baselines. [14]

The physical model is formulated as a linear relation over the physics feature vector X_p

$$I_i = \sum_{p=1}^{17} c_{p,i} X_p, \quad i \in \{x, y, z, sp\}, \quad (11)$$

where $c_{p,i}$ denotes the bin-local model coefficients for the respective axis or spindle current. These coefficients are estimated by weighted least squares using a modified Gaussian neighborhood weighting

$$w_{b_{ijk}}^{(n)} = \mathcal{N}\left(b_{b_{ijk}}^{(n)} \mid \bar{b}_{ijk}, \sigma^2 \text{diag}(1, 1, \alpha^{-2})\right) + \beta. \quad (12)$$

As an outcome of the PiDO approach, the proportions of PiA and PiE are determined as functions of the available real-data share x . For this purpose, a fourth-order polynomial is used as a compromise between functional flexibility and moderate model complexity

$$f_{\text{PiA/PiE}}(x) = a_4 x^4 + a_3 x^3 + a_2 x^2 + a_1 x + a_0, \quad (13)$$

where a_n , $n \in \{0, 1, 2, 3, 4\}$, are the polynomial coefficients and x denotes the proportion of real data. A least-squares fit yields

$$\begin{aligned} f_{\text{PiA}}(x) = & -5.908 \cdot 10^{-8} x^4 + 3.653 \cdot 10^{-6} x^3 \\ & + 4.098 \cdot 10^{-4} x^2 - 1.017 \cdot 10^{-2} x \\ & + 1.768 \cdot 10^{-2} \end{aligned} \quad (14)$$

for PiA and

$$\begin{aligned} f_{\text{PiE}}(x) = & -1.119 \cdot 10^{-6} x^4 + 3.314 \cdot 10^{-4} x^3 \\ & - 3.522 \cdot 10^{-2} x^2 + 1.335 x \\ & + 1.766 \end{aligned} \quad (15)$$

for PiE. [14]

3.6 Evaluation Metrics

To evaluate the monitoring model, a window-based evaluation metric is employed, specifically tailored to process monitoring tasks. This allows not only an assessment of whether a deviation has been detected, but also

of its quality in terms of temporal localization and false-alarm behavior. Instead of point-wise classification, predicted anomaly windows

$$P = \{(p_s^{(i)}, p_e^{(i)})\}_{i=1}^{N_p}$$

are compared against ground-truth anomaly windows

$$T = \{(t_s^{(j)}, t_e^{(j)})\}_{j=1}^{N_t}$$

In real machining processes, anomaly boundaries are rarely perfectly sharp and can exhibit small offsets due to sensor noise, control dynamics, filtering, and labeling conventions. To avoid over-penalizing minor temporal shifts that are operationally irrelevant, a temporal tolerance of ± 5 s is applied. Thus, a ground-truth window $(t_s, t_e) \in T$ is considered detected if at least one predicted window $(p_s, p_e) \in P$ overlaps within the tolerance:

$$(t_s - 5 \leq p_s \leq p_e \leq t_e + 5) \vee (p_s - 5 \leq t_s \leq t_e \leq p_e + 5). \quad (16)$$

The True Positive Rate (TPR) is defined as

$$\text{TPR} = \begin{cases} \frac{|\{\text{detected ground-truth windows}\}|}{N_t}, & N_t > 0, \\ 1, & N_t = 0, \end{cases} \quad (17)$$

and quantifies the fraction of actual anomalies that are successfully detected.

To explicitly assess false alarms, two complementary false positive rates are used. The area-based false positive rate measures the proportion of predicted anomaly time not overlapping with any ground-truth window,

$$\text{FPR}_{\text{area}} = \begin{cases} \frac{\sum_{(p_s, p_e) \in P_{\text{rem}}} (p_e - p_s)}{\sum_{(p_s, p_e) \in P} (p_e - p_s)}, & N_p > 0, \\ 0, & N_p = 0, \end{cases} \quad (18)$$

where P_{rem} denotes predicted windows without overlap. In addition, the count-based false positive rate is defined as

$$\text{FPR}_{\text{count}} = \begin{cases} \frac{|P_{\text{rem}}|}{N_p}, & N_p > 0, \\ 0, & N_p = 0. \end{cases} \quad (19)$$

The final evaluation metric combines detection performance and false-alarm behavior into a single Window-Score,

$$F_W = \frac{8 \cdot \text{TPR} + (1 - \text{FPR}_{\text{area}}) + (1 - \text{FPR}_{\text{count}})}{10}. \quad (20)$$

The strong weighting of the TPR reflects the process-monitoring requirement that missed anomalies are significantly more critical than false alarms, while still penalizing excessive or prolonged false detections.

4 Experimental setup

4.1 Machines and Datasets

This study is based on the *Gear*, *Notch*, and *Plate* datasets introduced in [37]. These datasets represent

complementary machining scenarios with different geometric characteristics. All datasets are produced using aluminum and steel under different technological parameters (see table 2). All configurations, including those with identical parameters, are produced three times in succession and used independently in the experiment. During the final iterations of the depth test series, chattering (oscillating anomalies) occurred in the steel series, and galling, with potential tool breakage, occurred in the aluminum series (a long-term anomaly). Furthermore, the Blowhole Series provides samples of short-term anomalies. The anomalous parts are shown in figure 4, and the effects on the signals are visualized in figure 5.

Table 2: Datasets used in Experiment 1, recorded on a DMC 60H.

Material	Type	v_c [mm/min]	v_{sp} [min ⁻¹]	a_p [mm]	a_e [mm]
AL 2007 T4	Normal	350	3800	6	10
AL 2007 T4	SF	267/350/442	2900	6	10
AL 2007 T4	Depth	350	3800	3/6/12	10
AL 2007 T4	Blowhole	350	3800	6	10
S235JR	Normal	576	3200	6	10
S235JR	SF	432/576/720	2400	6	10
S235JR	Depth	576	3200	3/6/12	10
S235JR	Blowhole	576	3200	6	10

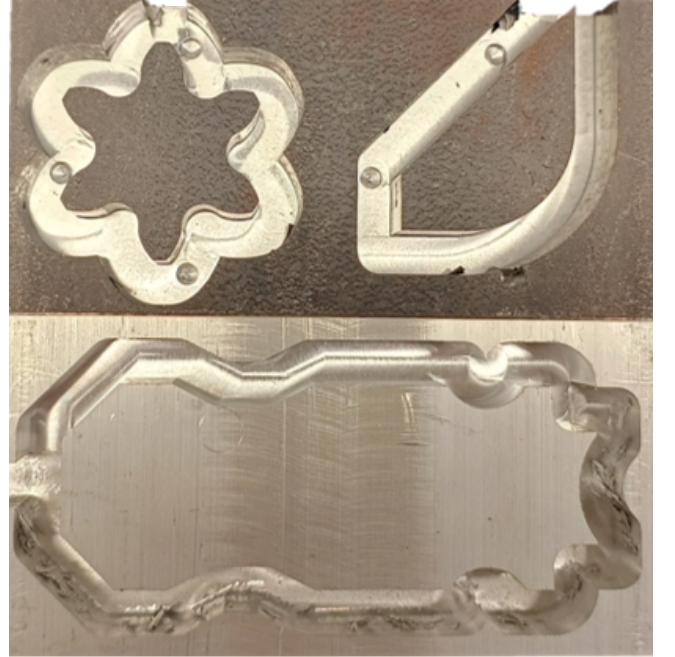


Figure 4: Blowhole series from Gear and Notch (steel, top) and depth series from Plate (aluminum, bottom)

For end-to-end validation under domain shift, the systematic benchmark dataset [38] is used. This dataset covers two series of *Impellers* produced with POM-C on a DMG MORI DMC 60 H (IMP1–IMP12) and a DMG MORI CMX 600 V (IMP1–IMP12 \ IMP4). The processes were planned by CAM, resulting in complex toolpaths and varying radial cutting depths, material removal rates, and process forces. The machining parameters were set $v_c = 126$ mm/min, $v_{sp} = 8000$ min⁻¹, $a_p = 8$ mm, enabling the analysis to focus on adaptation

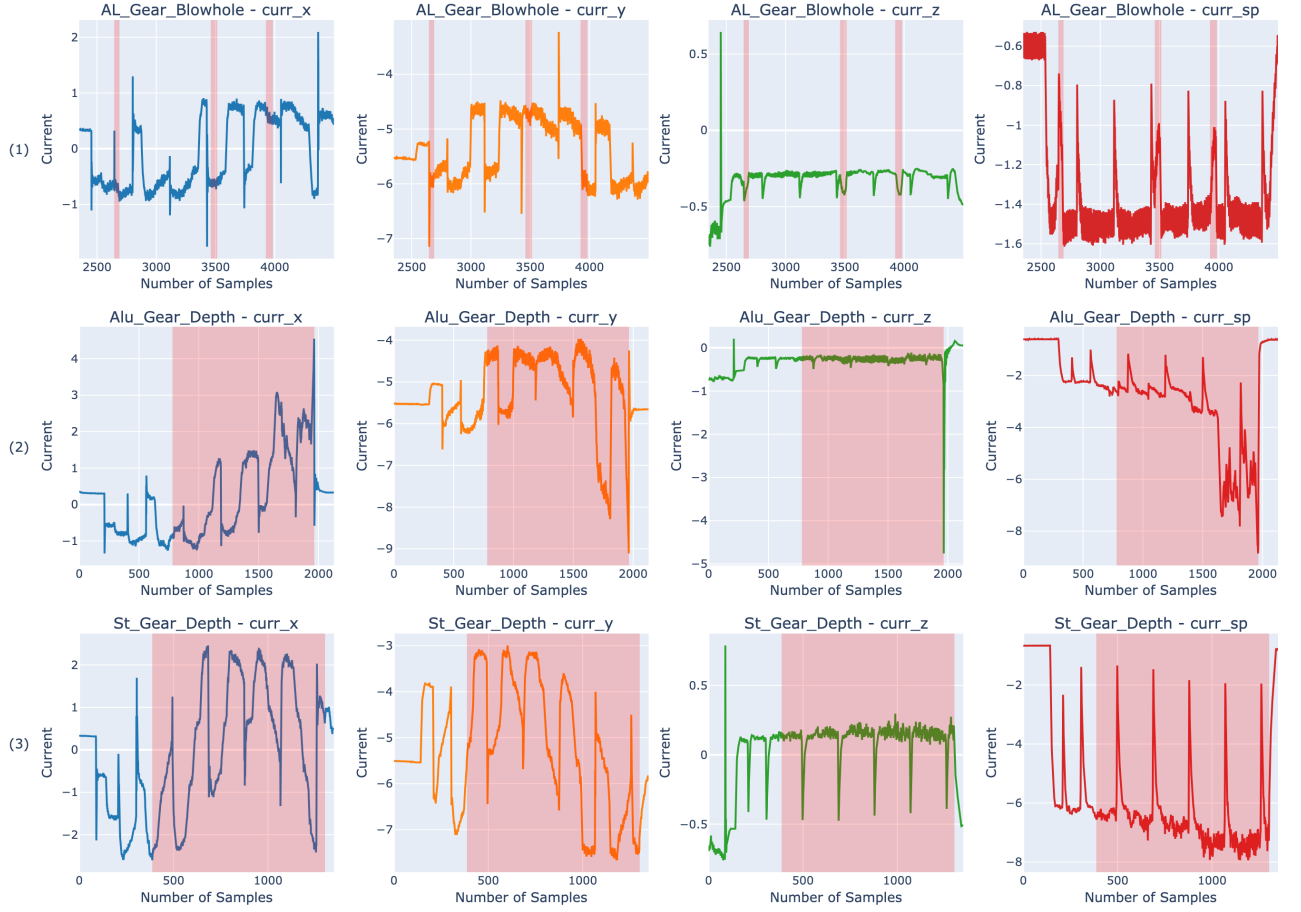


Figure 5: Labeled blowhole and depth anomalies in current signals of the X-, Y-, Z-, and SP-Axis

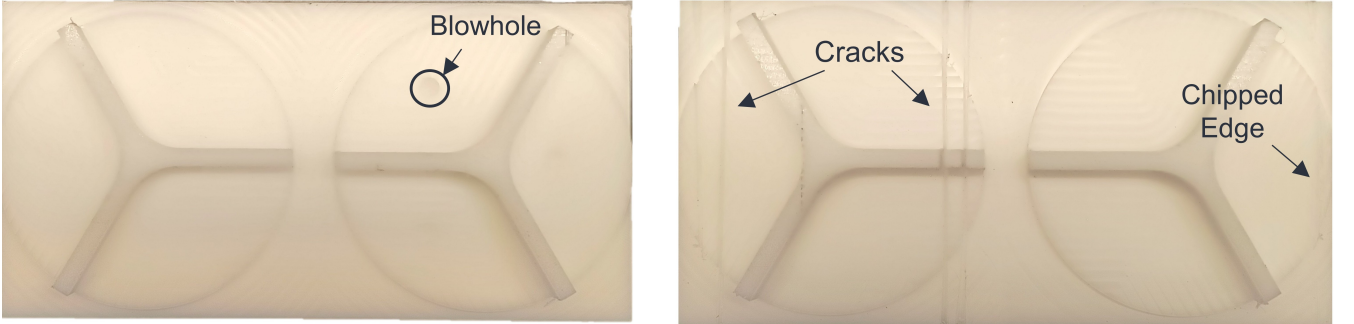


Figure 6: Impeller series with anomalous processes due to cavities, cracks, and chipped edges

across workpieces, toolpath geometries, and machines. Additionally, four anomalous processes involving cavities, cracks, and chipped edges are given on both machines (see figure 6). The dataset was chosen because the soft anomalies make detection particularly complex, allowing for a differentiated evaluation of the learning process. All datasets were collected at a sampling rate of 500 Hz and downsampled to 25 Hz.

4.2 Experiment 1: Identification of the Monitoring Model

This experiment is carried out in a realistic application scenario involving real process data and the challenge of generalizing to unknown workpiece geometries.

Based on the findings of [10], Residual generation is performed by an ExtraTrees model (estimators = 30, max depth = 100, leaves = 5, nodes = 1000) on the transitory axes and a Gradient Boosting model (estimators = 100, nodes = 6000) on the main spindle. This enables a data-centric investigation of the effects on the monitoring model.

The training and validation datasets used are *Gear* and *Notch*. The *Plate* dataset is reserved as a strictly independent test set. Since it differs in both geometry and machining characteristics, it serves as a realistic benchmark for generalization to unseen workpiece geometries. A strict separation between training, validation, and test data is maintained to prevent overfitting and to quantify deployment-relevant performance. The

Table 3: Training-set configurations of Experiment 1. The *Plate* dataset serves as an independent test set.

Config	Training data	Test data
Big	Gear Normal, Gear SF, Notch Normal, Gear Depth, Gear Blowhole	Plate Normal, Plate SF, Plate Depth, Plate Blowhole
Mid	Gear Normal, Notch Normal, Gear Depth, Gear Blowhole	Plate Normal, Plate SF, Plate Depth, Plate Blowhole
Small	Gear Normal (50%), Notch Normal (50%), Gear Depth, Gear Blowhole	Plate Normal, Plate SF, Plate Depth, Plate Blowhole

three dataset configurations shown in table 3 are defined to assess the robustness of the anomaly detection methods with respect to varying training data availability. The ‘Big’ configuration represents a scenario with high data availability. The Mid configuration reduces the amount of training data, simulating moderate availability by excluding variations in feed rate and spindle speed. The Small configuration reduces the effective training data further by sub-sampling to 50%, representing a minimal-data scenario to enable the explicit testing of generalization under data scarcity.

All considered detectors expose hyperparameters that influence their detection behaviour. To ensure a fair comparison, these are optimised separately using Bayesian optimisation in Optuna with 500 optimisation loops. In each trial, Optuna proposes a hyperparameter configuration based on the outcomes of previous trials. Performance is evaluated using leave-one-out cross-validation. Therefore, the prediction model is trained on the anomaly-free files and applied to the validation folds to calculate the residuals. The configured detector is then applied to the residua. The evaluation metrics are computed per fold and averaged. The mean F1-Score is returned to Optuna, which then proposes the next hyperparameter configuration for the subsequent trial. Finally, the performance is evaluated using the plate test files. To ensure a statistically valid result, this procedure is performed 30 times for each detector and material combination.

To standardize the search space across detectors, all statistical methods share three additional hyperparameters, namely the window size, the minimum anomaly duration required for a valid window, and the window patience, the number of consecutive time steps required to confirm a state change. In addition, each method exposes a set of detector-specific hyperparameters, summarized in table 4.

4.3 Experiment 2: End-to-end validation under domain shift

In order to adapt to progressive domain shifts, the pipeline must be able to learn incrementally from a stream of machining runs. Therefore, a realistic production setting must be created to validate the adaptive monitoring system. Unlike Experiment 1, the process model is re-optimized to incorporate an additional model-centric optimization in a realistic application. Component ablations are evaluated for incre-

Table 4: Hyperparameters for the statistical detectors.

Method	Hyperparameters and description
Z-Score	<code>all_axis_threshold</code> (0–1), threshold for the aggregated product of axis-specific rolling Z-scores.
EWMA	<code>std_multiplier</code> (0.1–2.0), multiplier for adaptive EWMA-based deviation bands; <code>abs_offset</code> (0.1–5.0), absolute offset around the EWMA centerline; <code>min_axis_detections</code> (1–4), minimum number of detecting axes.
CUMSUM	<code>std_multiplier</code> (0.1–2.0), multiplier for adaptive CUMSUM-based deviation bands; <code>abs_offset</code> (0.1–5.0), absolute offset around the cumulative-residual reference; <code>min_axis_detections</code> (1–4), minimum number of detecting axes.
SPC	<code>control_limit_mul</code> (1.0–5.0), multiplier for adaptive control limits; <code>method</code> ({ <code>ewma</code> , <code>rolling</code> }), method for estimating the moving centerline and dispersion; <code>enable_rule1</code> ({ <code>True</code> , <code>False</code> }), Western Electric Rule 1 (3 σ); <code>enable_rule2</code> ({ <code>True</code> , <code>False</code> }), Rule 2 (2/3 at 2 σ); <code>enable_rule3</code> ({ <code>True</code> , <code>False</code> }), Rule 3 (4/5 at 1 σ); <code>enable_rule4</code> ({ <code>True</code> , <code>False</code> }), Rule 4 (8 consecutive points); <code>min_axis_detections</code> (1–4), minimum number of detecting axes.

mental learning and in combination with density-based sampling, active learning, and physics-informed dataset optimization. Table 5 provides an overview of the scenarios, namely Baseline, IL, DcIL, DcIL with AL, DcIL with PiDO, and Full. All configurations are consistently evaluated on both machines in a sequential adaptation protocol.

Table 5: Evaluated scenarios in Experiment 2

Scenario	Incremental training	DcIL	PiDO	AL
Baseline				
Incremental	✓			
DcIL	✓	✓		
DcIL with AL	✓	✓		✓
DcIL with PiDO	✓	✓	✓	
Full	✓	✓	✓	✓

An initial training dataset consisting of anomaly-free and anomaly-affected processes is randomly created. The remaining workpieces are then evaluated in randomly permuted sequences of 1-12 training files to emulate different production orders. Within each sequence, the pipeline is updated online before the next workpiece is evaluated. Depending on the scenario, the current training set is reduced via DCIL, adopted via AL, or expanded via PiDO, based on the upcoming test workpiece. The anomaly-free subset is split into two. One split is used to optimize the process models using an Extra Trees regressor within 100 Optuna loops (with

100–600 estimators, a maximum depth of 1–10, and a minimum number of samples per split of 2–10). The optimized model is then applied to the second anomaly-free split and the subset containing anomalous sections to calculate residuals. Subsequently, the SPC detector is optimized on the basis of the Windows Score in 100 loops. The updated process and monitoring model is then applied to subsequent processes in the sequence, and the results are stored. The data is added to the training set, yielding learning curves that quantify adaptation behavior over the production stream.

Algorithm 2 summarizes the incremental validation loop, which is repeated 30 times on both machines. To evaluate the isolated impact on detection performance and prevent bias, test sequences containing both anomaly-free and anomalous segments are analyzed. Therefore, the random permutations result in a total of 2799 evaluable runs across all scenarios on both machines.

Algorithm 2 Incremental validation procedure

```

1: function INCREMENTALVALIDATION( $\mathcal{D}_{\text{dataset}}, \text{scenario}$ )
2:    $\mathcal{F} \leftarrow$  all workpieces in  $\mathcal{D}_{\text{dataset}}$ 
3:    $\mathcal{F}_{\text{anom}} \leftarrow \{f \in \mathcal{F} \mid \mathcal{A}(f)\}$ 
4:    $\mathcal{F}_{\text{norm}} \leftarrow \{f \in \mathcal{F} \mid \neg \mathcal{A}(f)\}$ 
5:    $\mathcal{F}_{\text{init}} \leftarrow \{\text{Pick}_1(\mathcal{F}_{\text{anom}}), \text{Pick}_1(\mathcal{F}_{\text{norm}})\}$ 
6:    $\mathcal{F}_{\text{rem}} \leftarrow \mathcal{F} \setminus \mathcal{F}_{\text{init}}$ 
7:    $\mathcal{R} \leftarrow \emptyset$ 
8:   for all  $P \in \text{Perm}(\mathcal{F}_{\text{rem}})$  do
9:      $\mathcal{T}_{\text{train}} \leftarrow \mathcal{F}_{\text{init}}$ 
10:    for  $t = 1$  to  $\min(12, |P|)$  do
11:       $f_{\text{test}} \leftarrow P_t$ 
12:      if scenario uses IL, DcIL, PiDO, or AL
13:      then
14:         $\mathcal{T}_{\text{train}} \leftarrow \text{DATASAMPLE}(\mathcal{T}_{\text{train}})$ 
15:      end if
16:      if scenario uses PiDO then
17:         $\mathcal{T}_{\text{train}} \leftarrow \text{DATAOPTIMIZE}(\mathcal{T}_{\text{train}}, f_{\text{test}})$ 
18:      end if
19:       $M \leftarrow \text{TRAINPROCESSMODEL}(\mathcal{T}_{\text{train}})$ 
20:       $R_{\text{train}} \leftarrow \text{RESIDUALS}(M, \mathcal{T}_{\text{train}})$ 
21:       $D \leftarrow \text{OPTIMIZEDDETECTOR}(R_{\text{train}})$ 
22:       $R_{\text{test}} \leftarrow \text{RESIDUALS}(M, f_{\text{test}})$ 
23:       $\text{metrics} \leftarrow \text{EVALUATE}(D, R_{\text{test}})$ 
24:       $\mathcal{R} \leftarrow \mathcal{R} \cup \{(\text{scenario}, f_{\text{test}}, t, \text{metrics})\}$ 
25:       $\mathcal{T}_{\text{train}} \leftarrow \mathcal{T}_{\text{train}} \cup \{f_{\text{test}}\}$ 
26:    end for
27:  end for
28:  return  $\mathcal{R}$ 
29: end function

```

5 Results

5.1 Experiment 1

Experiment 1 evaluates the robustness of the statistical detectors under variations in training-set size and dataset configuration. Table 6 provides a comparison of

the evaluated detectors in terms of average performance and stability across all runs.

Table 6: Detector-level summary sorted by mean F_W

Detector	F_W (std)	F1	Precision	Recall
SPC	0.7309 (0.1708)	0.6595	0.5564	0.8256
ZScore+SPC	0.6486 (0.2438)	0.5912	0.5268	0.7070
EWMA+SPC	0.6441 (0.2384)	0.5868	0.5198	0.7060
ZScore+ZScore	0.6082 (0.2530)	0.5519	0.4960	0.6576
EWMA	0.5564 (0.2382)	0.5151	0.4742	0.5953
ZScore	0.5538 (0.2525)	0.5076	0.4740	0.5931
CumSum	0.4138 (0.2064)	0.3858	0.4756	0.3649

SPC achieves the strongest overall results, with the highest mean window-level score ($\bar{F}_W = 0.7309$) and the best balanced classification metrics (F1 = 0.6595, precision = 0.5564, recall = 0.8256), while also exhibiting the lowest variability ($\sigma = 0.1708$). The composite variants improve over their single-component counterparts in most metrics (e.g., Composite (ZScore+SPC): $\bar{F}_W = 0.6486$, F1 = 0.5912), but remain less stable than SPC ($\sigma = 0.2384$ –0.2530). Among the remaining individual detectors, EWMA and ZScore yield mid-range performance ($\bar{F}_W = 0.5564$ and 0.5538) with comparatively higher variability ($\sigma = 0.2382$ and 0.2525). In contrast, CumSum shows the weakest mean detection performance ($\bar{F}_W = 0.4138$, F1 = 0.3858) despite moderate variability ($\sigma = 0.2064$). Overall, the results indicate that SPC offers the most favorable trade-off between detection quality and reproducibility.



Figure 7: Window-Score across training-set configurations (big/mid/small) and material.

Figure 7 visualizes performance across dataset configurations and provides evidence on data efficiency. The single detectors remain largely stable across big, mid, and small training sets, indicating robustness under reduced training data. This is particularly relevant for small-batch production, where only limited training data is often available. Composite detectors show similarly stable behavior across configurations, supporting their suitability for deployment under data constraints.

Figure 8 compares the methods in precision-recall space and highlights characteristic detection profiles. SPC lies in the highest recall and precision region, resulting in an F1 of 0.66. In contrast, the composite detectors (EWMA+SPC, Z-Score+SPC) cluster in a less balanced region, resulting in lower F1-Scores.

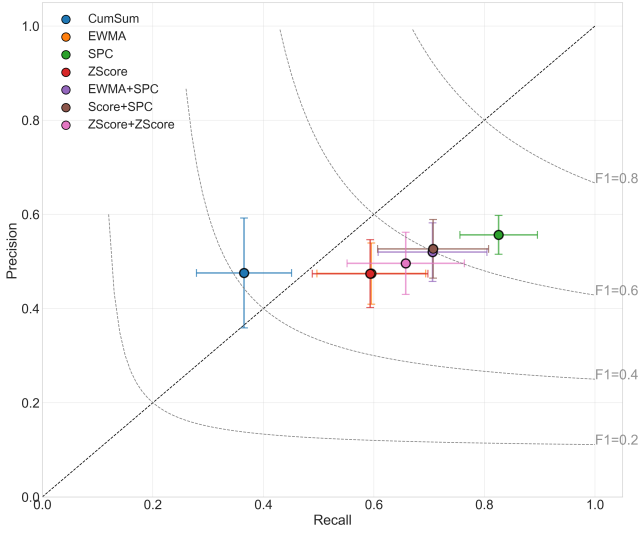


Figure 8: Precision–recall characteristics with F1-Score isolines.

5.2 Experiment 2

This experiment evaluates the proposed monitoring framework under incremental model and dataset maintenance strategies. Figure 9 presents the scenario-specific evaluation, aggregated across both machines and all iterations.

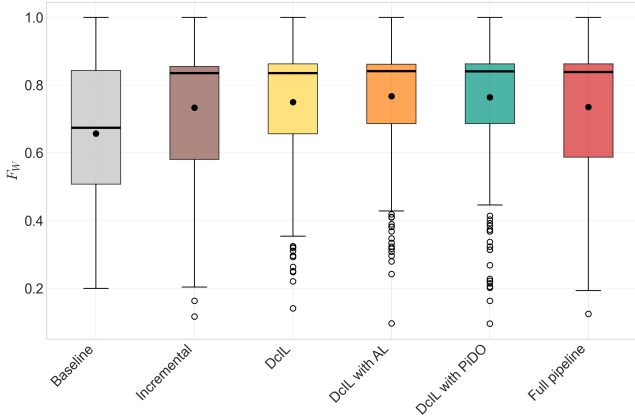


Figure 9: F_W -Distributions of the evaluated scenarios.

All incremental variants show an improvement on the baseline. The baseline reaches a median F_W of 0.674, whereas the incremental scenarios cluster in a narrow range between 0.836 and 0.841. The highest median is obtained by DcIL with AL (0.841) and DcIL with PiDO (0.841), followed by the Full pipeline (0.839), while Incremental and DcIL form the lower end of the incremental group (both 0.836). Overall, all incremental scenarios improve the median F_W by approximately 0.16 compared with the baseline, indicating that incremental updating provides the dominant performance gain.

The statistical significance is assessed using Welch’s t test for pairwise comparisons, as summarized in table 7. All incremental variants achieve a significantly higher window score than the baseline. The mean improvement relative to the baseline ranges from $\Delta\mu = 0.076$

Table 7: Results of the Welch test for pairwise comparisons of the window scores for the overall scenario

Scenario 1	Scenario 2	$\mu - \mu$	t	p	$p < 0.05$
Baseline	Incremental	0.076	-4.52	7.72×10^{-6}	✓
	DcIL	0.092	-5.78	1.21×10^{-8}	✓
	DcIL with AL	0.110	-7.05	5.46×10^{-12}	✓
	DcIL with PiDO	0.107	-6.64	7.29×10^{-11}	✓
	Full pipeline	0.078	-4.66	4.03×10^{-6}	✓

(Incremental) to $\Delta\mu = 0.110$ (DcIL with AL), with all corresponding baseline comparisons being highly significant ($p < 0.001$). At the same time, most pairwise comparisons among the incremental variants are not significant. Only Incremental vs. DcIL with AL ($p = 0.032$) and DcIL with AL vs. Full pipeline ($p = 0.041$) reach nominal significance, motivating a detailed analysis of their progression over the learning process.

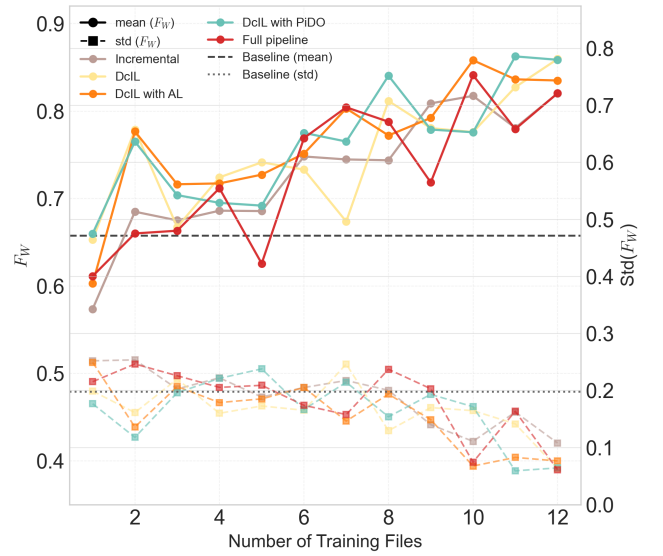


Figure 10: Evolution of the window score F_W and its standard deviation over the incremental learning process for the evaluated scenarios.

Figure 10 summarizes both trend and variability of F_W . All scenarios improve over the sequence and finish above the baseline mean (0.657), but with varying trajectories. The largest gain is observed for Incremental (0.574 to 0.820), followed by DcIL with AL (0.603 to 0.835), while DcIL and DcIL with PiDO reach the highest final means (0.860 and 0.859). The Full pipeline also improves (0.611 to 0.820) but peaks earlier (around $x = 10$). At the same time, $\text{Std}(F_W)$ decreases for all methods, indicating stabilization from 0.177–0.252 at $x = 1$ to 0.061–0.108 at $x = 12$. The lowest final variability is obtained by Full pipeline (0.061), whereas Incremental remains the most variable (0.108).

The F_W -Distributions of the first two and last two sequences are compared to quantify the progression patterns. Figure 11 shows an upward shift for all scenarios. The largest shifts are observed for the Incremental and Full pipeline, followed by DcIL with PiDO. DcIL with AL and especially DcIL exhibit smaller shifts due to their initial performance. A Welch’s t-test is applied to

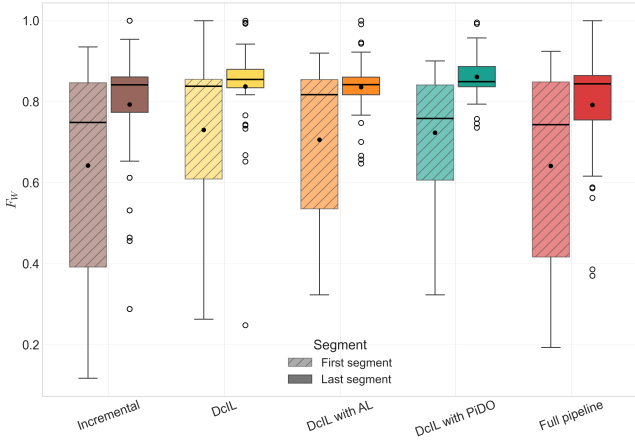


Figure 11: F_W -Distribution of the first two and the last two sequences.

statistically assess the differences (see table 8). *DcIL* shows a significant improvement from $\mu_A = 0.731$ in the early to $\mu_B = 0.837$ in the late segment. The same holds for *DcIL with AL* (0.706 to 0.836) and the *Full pipeline* (0.641 to 0.792). *Incremental* also shows a significant increase from 0.642 to 0.793. The strongest gain is observed for *DcIL with PiDO*, which improves from 0.724 to 0.861 with $p = 3.61 \times 10^{-6}$. Overall, all within-scenario early-to-late differences are statistically significant at $\alpha = 0.05$.

Table 8: Welch significance tests for early and late learning-curve sequences of window score F_W .

First vs. Last Segment	μ_A	μ_B	t	p	$p < 0.05$
Incremental	0.642	0.793	-3.19	0.0023	✓
DcIL	0.731	0.837	-3.01	0.0037	✓
DcIL with AL	0.706	0.836	-3.58	8.30×10^{-4}	✓
DcIL with PiDO	0.724	0.861	-5.23	3.61×10^{-6}	✓
Full pipeline	0.641	0.792	-3.45	9.99×10^{-4}	✓

6 Discussion

6.1 Experiment 1

The results indicate that SPC achieves the highest overall performance with a score of 0.731, followed by ZScore+SPC (0.649) and EWMA+SPC (0.644). These top methods also show comparatively high recall (SPC: 0.826, Score+SPC: 0.707, EWMA+SPC: 0.706), indicating that most true anomalies are detected. Precision is moderate (SPC: 0.556, Score+SPC: 0.527, EWMA+SPC: 0.520), implying the presence of false alarms. This must be interpreted alongside the conservative evaluation logic, which prioritizes the quality of anomaly detection in sequences containing both normal and anomalous segments. Consequently, the level of detail for purely normal segments can be reduced, and false positives are penalized severely. Therefore, these metrics are most informative for mixed-process sequences containing both anomaly-free and anomalous parts.

Based on these findings, SPC is selected as the detection method for the developed process monitoring sys-

tem. This choice is primarily motivated by its superior overall performance among the evaluated methods. In the present monitoring context, high recall is particularly important to minimize missed anomalies, while the achieved precision remains sufficient to keep false alarms at an acceptable level. Compared with more complex ML- or DL-based approaches, SPC is straightforward to implement and interpret, and its transparent rule-based logic supports acceptance in industrial practice. Moreover, SPC does not require a dedicated training phase on normal data, which simplifies commissioning and facilitates adaptation to new workpiece geometries, materials, and process conditions.

6.2 End-to-End Validation

All incremental strategies outperform the baseline, confirming the essential nature of updating the monitoring system when operating domain shifts. The baseline achieves a median F_W of 0.674, whereas all incremental variants lie within the narrow range (0.836 to 0.841). The main performance gain is achieved by moving from a static baseline to an incremental strategy. At the same time, the data-centric extensions mainly affect how this adaptation unfolds over the learning process.

The learning curves show that incremental variants differ in their temporal evolution. The significant improvement in performance of Incremental indicates that online updating provides a robust baseline. DcIL improves performance and is therefore effective in managing the growing training dataset. Adding AL results in the highest median, suggesting that the oversampling interface can be beneficial. PiDO contributes to stabilizing the adaptation and achieving the strongest late-stage detection performance. The full pipeline combines all methods, but individual gains are lost, and performance approaches the incremental setting. This indicates that the methods influence each other when combined, reducing the isolated positive effects. Furthermore, all scenarios show reduced standard deviation with DcIL with AL ($\text{Std}(F_W) = 0.079$) and DcIL with PiDO ($\text{Std}(F_W) = 0.059$) reaching the lowest late-stage variability. This indicates that the additional methods, especially, stabilize performance.

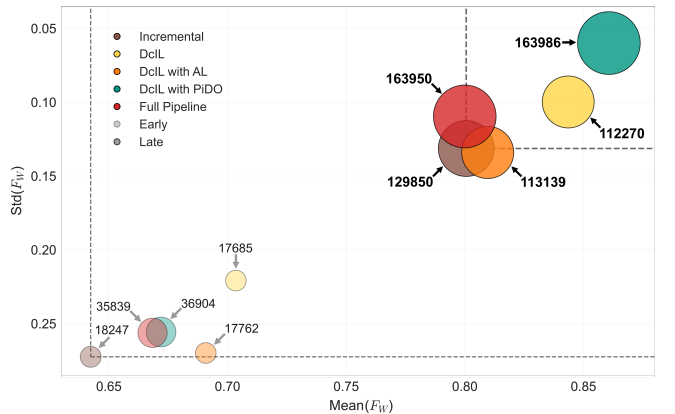


Figure 12: F_W and $\text{Std}(F_W)$ combined with storage requirements.

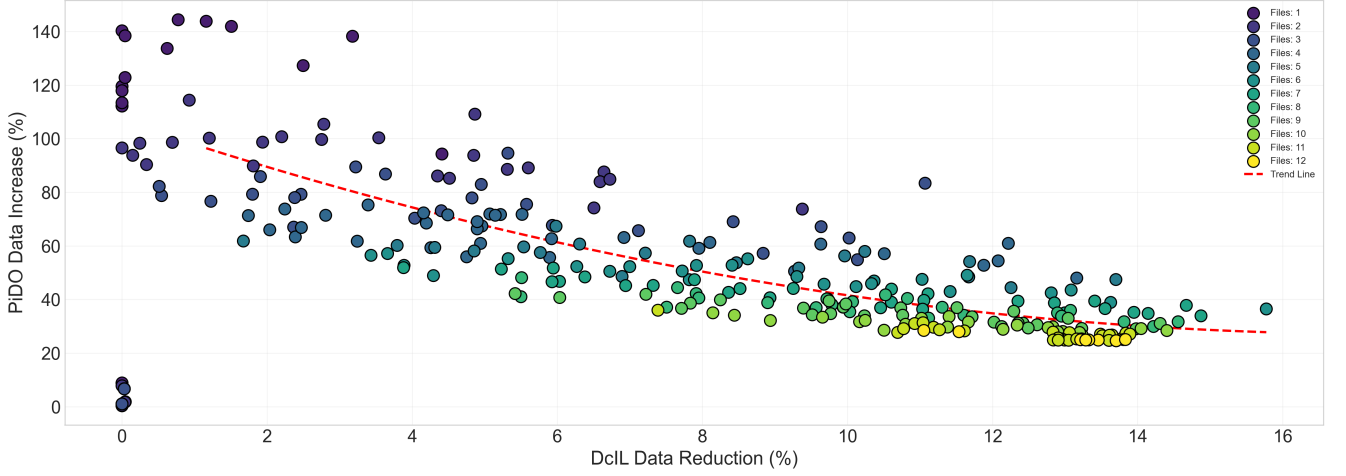


Figure 13: Relation between DcIL-based data reduction and PiDO-based data increase over incremental sequences.

To further condense the findings, figure 12 relates F_W and $\text{Std}(F_W)$ with the storage requirements. A comparison between the early (bottom left) and late (top right) iterations clearly demonstrates the improved detection performance. The mean F_W increases by $0.107 \leq \Delta F_W \leq 0.151$, while variability decreases by $-0.124 \leq \Delta \text{std}(F_W) \leq -0.061$, indicating that incremental adaptation is recommended. These shifts are accompanied by an increase in data volume from roughly 25-54k to 110-161k. Figure 13 visualizes the relationship between DcIL-based data reduction and PiDO-based data generation during this process. Data generation is relevant in the low-data regime, while data-centric selection becomes increasingly important once the memory contains a broader range of real workpieces. This can be addressed by the two complementary methods since PiDO contributes a relatively large amount of data at the beginning. In contrast, DcIL reduces the amount of data in the later stages.

Compared to Incremental ($F_W = 0.642$), DcIL ($F_W = 0.731$) yields improvement, particularly in the initial phase. Since 15.6% more data has to be stored for Incremental, it can be concluded that DcIL is able to select high-quality data and reduce storage requirements. When PiDO is integrated into DcIL, the highest detection performance is achieved in both the early ($F_W = 0.724$) and late ($F_W = 0.861$) segments. This demonstrates that PiDO can be successfully applied to DcIL-storage and that dataset management and optimization can improve detection performance. When the AL-interface is activated for DcIL or the full pipeline, performance approaches Incremental. Consequently, no clear benefit of the oversampling interface can be isolated. This suggests that adding data based on false detections does not necessarily enhance detection performance.

7 Conclusion and Future Work

This work demonstrates that residual-based monitoring is an effective and deployment-oriented strategy

for CNC milling. By combining an axis- and spindle-specific predictive process model with a lightweight statistical monitoring layer, the proposed architecture suppresses nominal process variability and highlights deviations caused by disturbances and faults. The results show that predictive reference generation and residual-based monitoring form a flexible condition-monitoring approach that remains effective under varying geometries, materials, and machines.

Across the experiments, the results confirm both the effectiveness of the residual-based monitoring concept itself and the importance of adaptive model maintenance. For detector identification, SPC achieved the best overall trade-off between detection quality, robustness, and interpretability. All adaptive scenarios significantly outperformed the static baseline during validation under domain shift. The baseline reached a median F_W of 0.674, whereas all incremental variants achieved median values between 0.836 and 0.841. This shows that the dominant performance gain is obtained by moving from a static strategy to incremental updating. With regard to data-centric model maintenance, the results show that the investigated extensions influence the monitoring system through their effect on temporal progression and stability. In summary, it can be concluded that

1. residual-based monitoring with predictive signal references provides a foundation for CNC condition monitoring.
2. incremental model maintenance is essential once the production context evolves and previously unseen workpieces are introduced.
3. the learning effects of the prediction model is beneficial for the monitoring model.
4. the added value of more complex data-centric extensions lies primarily the temporal evolution and stability of the prediction performance. In this context, DcIL is particularly relevant when selective retention of informative samples and reduced storage capacities become important. PiDO can contribute

to a performance improvement at the expense of increased data volume.

Future work should broaden the empirical basis of validation and further analyze the interaction between predictive reference quality and downstream monitoring performance. So far, the validation has been limited to one main benchmark geometry and a restricted set of materials and machine settings. Additional geometries, materials, process settings, and anomaly classes are therefore required to determine the limits of generalization more clearly. Beyond this, future studies should systematically investigate longer incremental-learning horizons, the role of repeated versus highly diverse part spectra, and the extent to which active learning becomes beneficial under different recurrence patterns of workpieces. A further promising direction is the targeted analysis of how prediction-model quality, data-centric memory composition, and false-alarm behavior interact in residual-based monitoring systems.

References

- [1] Nicole Ottersböck and Isabella Urban. *Trends und Entwicklungen in der Arbeitswelt – Hintergründe zur Erforschung von KI-basiertem Wissenstransfer im Projekt KI_{eper}*. Dec. 2025. DOI: 10.1007/978-3-662-71591-8_1.
- [2] Stefan Sauer, Andreas Peichl, and Klaus Wohlrabe. “Fachkräftemangel in Deutschland und Europa – Historie, Status quo und was getan werden muss”. In: *ifo Schnelldienst* (2022).
- [3] Sven-Vegard Buer, Jan Ola Strandhagen, and Felix Chan. “The link between Industry 4.0 and lean manufacturing: mapping current research and establishing a research agenda”. In: *International Journal of Production Research* 56 (Mar. 2018), pp. 1–17. DOI: 10.1080/00207543.2018.1442945.
- [4] W. Elmaraghy et al. “Complexity in engineering design and manufacturing”. In: *CIRP Annals - Manufacturing Technology* 61 (Jan. 2012), pp. 793–814.
- [5] R. Teti et al. “Advanced monitoring of machining operations”. In: *CIRP Annals - Manufacturing Technology* 59 (Dec. 2010), pp. 717–739. DOI: 10.1016/j.cirp.2010.05.010.
- [6] Rui Zhao et al. *Deep Learning and Its Applications to Machine Health Monitoring: A Survey*. 2016. arXiv: 1612.07640 [cs.LG]. URL: <https://arxiv.org/abs/1612.07640>.
- [7] Robin Ströbel et al. “Intelligente Prozessüberwachung für die flexible Produktion”. In: *Zeitschrift für wirtschaftlichen Fabrikbetrieb* 120 (Mar. 2025), pp. 224–231. DOI: 10.1515/zwf-2024-0154.
- [8] Jinjiang Wang et al. “Deep Learning for Smart Manufacturing: Methods and Applications”. In: *Journal of Manufacturing Systems* 48 (Jan. 2018), pp. 144–156. DOI: 10.1016/j.jmsy.2018.01.003.
- [9] William Woodall. “Controversies contradictions in statistical process control”. In: 1 (Oct. 2000). DOI: 10.1080/00224065.2000.11980013.
- [10] Robin Ströbel et al. “Hybrid Machine Learning for CNC Process Monitoring”. In: *IEEE Access* PP (Jan. 2025), pp. 1–1. DOI: 10.1109/ACCESS.2025.3573400.
- [11] Jay Lee et al. “Intelligent Maintenance Systems and Predictive Manufacturing”. In: *Journal of Manufacturing Science and Engineering* 142.11 (Aug. 2020), p. 110805. ISSN: 1087-1357. DOI: 10.1115/1.4047856. eprint: https://asmedigitalcollection.asme.org/manufacturingscience/article-pdf/142/11/110805/6710166/manu_142_11_110805.pdf. URL: <https://doi.org/10.1115/1.4047856>.
- [12] Xiaoli Li, Shiu Tso, and Jun Wang. “Real-time tool condition monitoring using wavelet transforms and fuzzy techniques”. In: *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 30 (Sept. 2000), pp. 352–357. DOI: 10.1109/5326.885116.
- [13] Robin Ströbel et al. “Density-Based Dataset Balancing for Incremental Retraining in CNC-Axis and Spindle Current Prediction”. In: *IEEE Access* 14 (2026), pp. 103195–103214. DOI: 10.1109/ACCESS.2026.3708218.
- [14] Robin Ströbel et al. “Physics-Informed Dataset Optimization for CNC Axis-Current Prediction”. In: *IEEE Access* (2026), pp. 1–1. DOI: 10.1109/ACCESS.2026.3718352.
- [15] Changfu Liu, Lida Zhu, and Chenbing Ni. “Chatter detection in milling process based on VMD and energy entropy”. In: *Mechanical Systems and Signal Processing* 105 (May 2018), pp. 169–182. DOI: 10.1016/j.ymssp.2017.11.046.
- [16] Lars Johannsen, Wolfgang Boos, Kristian Arntz, and Marcel Prümmer. *Erfolgreich Fräsen : im Werkzeugbau*. Aachen: WBA Aachener Werkzeugbau Akademie GmbH, 2022.
- [17] Luh Asrini and Kung-Jeng Wang. “Integration Model of Residual-Based Mixed CUSUM-EWMA Chart with Deep Learning-Based Automatic Optical Inspection”. In: *International Journal of Modeling and Optimization* 14 (Jan. 2024), pp. 129–134. DOI: 10.7763/IJMO.2024.V14.861.
- [18] Kyle Hundman et al. “Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’18. ACM, July 2018, pp. 387–395. DOI: 10.1145/3219819.3219845. URL: <http://dx.doi.org/10.1145/3219819.3219845>.

- [19] Mohsin Munir et al. “DeepAnT: A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series”. In: *IEEE Access* 7 (2019), pp. 1991–2005. DOI: 10.1109/ACCESS.2018.2886457.
- [20] Taewook Kang et al. “A real-time anomaly detection method for robots based on a flexible and sparse latent space”. In: *Engineering Applications of Artificial Intelligence* 158 (Oct. 2025), p. 111310. ISSN: 0952-1976. DOI: 10.1016/j.engappai.2025.111310. URL: <http://dx.doi.org/10.1016/j.engappai.2025.111310>.
- [21] Daehyung Park, Yuuna Hoshi, and Charles C. Kemp. *A Multimodal Anomaly Detector for Robot-Assisted Feeding Using an LSTM-based Variational Autoencoder*. 2017. arXiv: 1711.00614 [cs.R0]. URL: <https://arxiv.org/abs/1711.00614>.
- [22] Mazen Alamir and Sacha Clavel. “An algorithm for sparse piece-wise polynomial residual generation for anomalies detection in industrial manipulator robots”. In: *Control Engineering Practice* 165 (Dec. 2025), p. 106545. DOI: 10.1016/j.conengprac.2025.106545.
- [23] Federico Pittino et al. “Automatic Anomaly Detection on In-Production Manufacturing Machines Using Statistical Learning Methods”. In: *Sensors* 20 (Apr. 2020), p. 2344. DOI: 10.3390/s20082344.
- [24] Mustafa Demetgul et al. “Monitoring the misalignment of machine tools with autoencoders after they are trained with transfer learning data”. In: *The International Journal of Advanced Manufacturing Technology* 128 (Aug. 2023), pp. 1–17. DOI: 10.1007/s00170-023-12060-2.
- [25] Carlos Lauro et al. “Monitoring and processing signal applied in machining processes – A review”. In: *Measurement* 58 (Dec. 2014), pp. 73–86. DOI: 10.1016/j.measurement.2014.08.035.
- [26] Jose Abellan-Nebot and Fernando Subirón. “A review of machining monitoring systems based on artificial intelligence process models. International Journal of Advanced Manufacturing Technology 47: 237-257”. In: *The International Journal of Advanced Manufacturing Technology* 47 (Mar. 2010), pp. 237–257. DOI: 10.1007/s00170-009-2191-8.
- [27] Dragos A. Axinte and N. Gindy. “Assessment of the effectiveness of a spindle power signal for tool condition monitoring in machining processes”. In: *International Journal of Production Research* 42 (2004), pp. 2679–2691. URL: <https://api.semanticscholar.org/CorpusID:110766528>.
- [28] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. “Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations”. In: *Journal of Computational Physics* 378 (2019), pp. 686–707.
- [29] Yusuf Altintas and Deniz Aslan. “Integration of virtual and on-line machining process control and monitoring”. In: *CIRP Annals - Manufacturing Technology* 66 (2017). DOI: 10.1016/j.cirp.2017.04.047.
- [30] Felix Doerrer et al. “Virtual Sensor for Accuracy Monitoring in CNC Machines”. In: *Journal of Manufacturing and Materials Processing* 6 (2022). DOI: 10.3390/jmmp6060137.
- [31] Chao-Wen Lu Marion R. Reynolds. “Control charts for monitoring processes with autocorrelated data”. In: *Nonlinear Analysis: Theory, Methods Applications* (1997).
- [32] Yuandi Wu, Brett Sicard, and Stephen Gadsden. “Physics-informed machine learning: A comprehensive review on applications in anomaly detection and condition monitoring”. In: *Expert Systems with Applications* 255 (2024). DOI: 10.1016/j.eswa.2024.124678.
- [33] Moschos Papananias et al. “Inspection by exception: A new machine learning-based approach for multistage manufacturing”. In: *Applied Soft Computing* 97 (Oct. 2020), p. 106787. DOI: 10.1016/j.asoc.2020.106787.
- [34] Knoll A Baumgartner B Rödel K. “A data mining approach to reduce the false alarm rate of patient monitors”. In: *IEEE* (2012).
- [35] Marcus Perry. “The Exponentially Weighted Moving Average”. In: June 2010. ISBN: 9780470400630. DOI: 10.1002/9780470400531.eorms0314.
- [36] Walter A. Shewhart. *Economic Control of Quality of Manufactured Product*. American Society for Quality Control, 1930.
- [37] Robin Ströbel et al. *Part series dataset of milling processes for time series prediction*. In collab. with Robin Ströbel et al. 2025. DOI: 10.35097/CTVUJ6DGZEPZMKOG.
- [38] Robin Ströbel et al. “A multimodal dataset for process monitoring and anomaly detection in industrial CNC milling”. In: *Data in Brief* 63 (2025), p. 112207. ISSN: 2352-3409. DOI: <https://doi.org/10.1016/j.dib.2025.112207>. URL: <https://www.sciencedirect.com/science/article/pii/S235234092500928X>.