IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 18 June 2026, accepted 19 July 2026, date of publication 30 July 2026, date of current version 6 August 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3718352

RESEARCH ARTICLE

Physics-Informed Dataset Optimization for CNC Axis-Current Prediction

ROBIN STRÖBEL[1], KAI NIKLAS HOFMANN[1], HAFEZ KADER[2], (Graduate Student Member, IEEE), JAN BAUMGÄRTNER[1], ALEXANDER PUCHTA[1], BENJAMIN NOACK[2], (Senior Member, IEEE), AND JÜRGEN FLEISCHER[1]

[1]wbk Institute of Production Science, Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany
[2]Autonomous Multisensor Systems (AMS) Research Group, Otto von Guericke University Magdeburg, 39106 Magdeburg, Germany

Corresponding author: Robin Ströbel (robin.stroebel@kit.edu)

This work was supported by the KIT-Publication Fund of Karlsruhe Institute of Technology.

ABSTRACT Accurate prediction of axis and spindle current signals is essential for model-based CNC process monitoring. Although representative datasets are essential, high product diversity, short life cycles, and varying process conditions lead to incrementally collected datasets with imbalanced or shifted distributions. Therefore, this paper proposes Physics-informed Dataset Optimization (PiDO), a data-centric PiML approach that enriches training datasets with physics-consistent samples before model retraining. PiDO uses locally parameterized physical equations (PEs) for axis- and spindle-current dynamics, derived from an energy-balance-based current model, to generate samples in a reduced discretized feature space. Physical consistency is constrained by local PE parameterization, feature-space bounds, NC-code, machine-kinematic information, and statistically observed ranges of unresolved variables. The approach is evaluated on 11 POM-C impeller geometries on two three-axis CNC milling machines (DMC 60H and CMX 600V), resulting in 22 validation datasets, using real-data-only training as baseline. The evaluation is based on RMSE, R^2, significance tests, and dataset-balance measures. Compared with a global PE, the proposed local PE parameterization reduced RMSE by 43.0–74.8% and increased R^2 from 0.072–0.448 to 0.821–0.947. In the downstream prediction task, PiDO improved dataset balance and achieved significant positive RMSE effects in several validation scenarios, particularly in sparse data regimes, without significant deterioration compared to the real-data baseline.

INDEX TERMS CNC machining, process monitoring, incremental learning, data-centric AI, dataset balancing, physics-informed machine learning.

I. INTRODUCTION

The growing demand for customized products and shorter product life cycles places greater demands on the flexibility of industrial manufacturing processes [1]. Consequently, production strategies are shifting towards single-item or low-volume manufacturing. This shift requires rethinking conventional process-monitoring methods to ensure cost-effective manufacturing without compromising quality. Due to their flexibility, CNC milling machines play a central role in the fast and precise manufacture of complex components in this regime. Their performance critically depends on the precise control of their axes. Therefore, their power supply enables continuous data acquisition, providing essential signals for modern process monitoring [2].

Traditional monitoring methods are based on statistical or physical principles. Statistical approaches exploit structures in historical data [3], but reach their limits when batch sizes are small and product characteristics vary widely. Physics-based models can provide simplified but robust system descriptions [4]. However, in dynamic environments where materials, workpieces, and tool conditions can vary

The associate editor coordinating the review of this manuscript and approving it for publication was Rongbo Zhu.

considerably, the conditions for traditional approaches become complex, making their limitations apparent [5]. Therefore, data-driven methods such as machine learning (ML) have become increasingly important [6]. Although ML models can identify complex patterns in extensive, multidimensional datasets, the availability of sufficient training data remains a limiting factor. In industrial practice, data accumulate gradually as more parts are produced, thus requiring incremental learning (IL) [7] approaches. Especially in the early stages, IL often suffers from imbalanced or insufficiently representative datasets, which can lead to biased model behavior. To address these limitations, data-centric AI [8], [9] and dataset balancing [10] are becoming increasingly important. These aim to improve training robustness by optimizing the dataset's composition. Physics-informed machine learning (PiML) [11] can extend this principle by incorporating physics into the dataset optimization pipeline. In particular, physics can be used to extend the dataset, thereby reducing the need for targeted data collection, which would be impractical and uneconomical for one-off or low-volume production.

We therefore introduce Physics-informed Dataset Optimization (PiDO), a data-centric approach that uses a Physical Equation (PE) to enrich training datasets with physics-consistent samples prior to model retraining. Instead of relying on additional targeted data collection, PiDO improves the representation of process-relevant regions within the available dataset. As a result, it can be integrated into incremental learning pipelines as an additional dataset optimization step. The main contributions are summarized as follows:

- A data-centric physics-informed dataset optimization framework, PiDO, is proposed.
- A locally parameterized physical equation for the axis- and spindle-current dynamics is used to generate physics-consistent samples in a reduced feature space.
- Two complementary data-generation mechanisms, Physics-informed Augmentation (PiA) and Physics-informed Extrapolation (PiE), are formulated. PiA augments measured samples, while PiE generates synthetic input-output samples to fill local gaps in sparsely populated process regions.
- The approach is systematically validated across different machines and process conditions using ablation-based comparisons against real-data-only baselines.

The remainder of this paper is organized as follows. Section II reviews the state of the art in IL, data-centric AI, and PiML, and derives the research questions addressed in this work. Section III presents the proposed PiDO methodology, including the hybrid ML-model, feature-space reduction, local PE parameterization, PiA and PiE data generation, and the assumptions defining the scope of applicability. Section IV describes the experimental design used to analyze the methodological components, while Section V reports the corresponding experimental results. Section VI interprets these results and consolidates them into the final PiDO configuration. Section VII validates this configuration on independent datasets across two machines using an ablation-based evaluation and discusses the results, limitations, and practical considerations. Finally, Section VIII summarizes the main findings and outlines directions for future work.

II. STATE OF THE ART

Incrementally growing datasets often exhibit imbalanced and locally sparse distributions, which can distort model training and reduce prediction robustness. Addressing this challenge requires combining insights from IL, data-centric AI, and PiML. While IL focuses on model adaptation under continuously changing datasets, data-centric AI emphasizes the quality and representativeness of training datasets, and PiML introduces physical knowledge into data-driven models. The following sections discuss these perspectives and identify the gap addressed by PiDO.

A. INCREMENTAL LEARNING

Incremental learning (IL) describes the gradual expansion of a database, as well as the acquisition of new knowledge from this non-stationary dataset. This approach is particularly relevant in scenarios where new information is continuously available [7], such as in industrial environments. A key challenge in incremental learning is catastrophic forgetting. This refers to the phenomenon by which a model forgets previously acquired knowledge when trained on new data, severely limiting its learning capability [7], [12].

IL can be categorized into three main types: Task-IL, domain-IL, and class-IL [7]. These categories differ in terms of how changing data interact with the model's context and task structure. *Task-IL* refers to the sequential learning of distinct and clearly separable tasks. Reference [13] presents a concept in which a neural network (NN) is used for multiple tasks. To avoid catastrophic forgetting, the weights of individual layers are modeled as linear combinations of existing weights rather than being overwritten during retraining. In *Domain-IL*, the overall task remains the same, but the context and boundary conditions change over time. A practical example can be found in [14], where a safety warning system is used to predict unstable conditions. The system is retrained using new data and loss functions that reinforce existing knowledge. *Class-IL* addresses the gradual learning of an increasing number of classes or objects. In [15], a system learns new classes by combining new data with stored exemplars of existing classes.

B. DATA-CENTRIC AI

In traditional model-centric training approaches, the dataset used to train an ML model is assumed to be adequate. However, a sufficient and balanced dataset is equally important as an optimized model. The data-centric AI paradigm, therefore, addresses the systematic preparation and improvement of the underlying database [8]. This

involves qualitative and quantitative improvements to the training data using domain-specific knowledge. [16] argues that data-centric approaches must be supplemented with repetitive updates. In practice, this approach has been demonstrated to outperform model-centric baselines [9]. Reference [17] presents a comprehensive survey focused on dataset optimization.

In unbalanced datasets, certain classes or process states are over- or under-represented, which distorts training and biases model behavior. Targeted data augmentation offers a promising solution [18]. However, augmenting the total dataset fails to correct dataset imbalances and can even exacerbate them. Balancing methods address this by adjusting the composition of the dataset via oversampling/undersampling [19], or the generation of synthetic data. For example, Geometric SMOTE synthesizes samples in sparsely populated regions to improve balance [10].

In summary, a data-centric approach prioritizes a diverse and high-quality dataset over increased model complexity or extensive hyperparameter tuning. However, conventional balancing and augmentation methods typically focus on the statistical structure of the dataset without explicitly considering whether generated samples are physically consistent with the underlying process. This is particularly relevant in CNC process monitoring, where sparsely represented regions may correspond to specific tool engagement, feedrate conditions, or machine-axis interactions. Consequently, data-centric optimization alone cannot ensure that samples remain physically meaningful. PiML can address this limitation by contributing physical knowledge to dataset balancing and enrichment.

C. PHYSICS-INFORMED ML

PiML combines data-driven approaches with physical knowledge to improve the robustness and predictive power of ML models, even when training data is limited [20], [21]. The increasing prevalence of new concepts [21], [22], [23] has resulted in the development of taxonomies [11], [21], [24]. The following three levels of increasing physical integration can be distinguished according to [11]:

In *observational bias*, the model is trained on data that reflects the underlying physics, allowing to learn physical functions or vector fields. This can be achieved through feature selection [25], [26], physically consistent data augmentation [27], or by simulating synthetic data in cases where it is sparse [28], [29] or difficult to observe [30].

Learning bias influences model training through the selection of an appropriate loss function, which is composed of data and physical loss terms $\mathcal{L} = \omega_{\text{data}}\mathcal{L}_{\text{data}} + \omega_{\text{physical}}\mathcal{L}_{\text{physical}}$. This method promotes convergence to solutions that correspond to the laws of physics. Although flexible, it only provides an approximation. For example, physical knowledge of Hamilton equations [31], fluid flow [32], heat transfer [33], or fault characteristics [34] can be integrated.

Inductive bias is based on assumptions about physical relationships that are incorporated into the design of a model. Consequently, the model adheres to certain physical laws. This often results in problem-specific implementations that are difficult to transfer. For example, the integration of physical knowledge was achieved in NN [35] in convolution NN [36] or recurrent NN [37] structures.

Although these PiML approaches demonstrate the benefits of incorporating physical knowledge, they predominantly focus on model-centric integration through loss functions, architectures, or feature representations, for example. In contrast, the role of physics in optimizing the composition of an incrementally collected training dataset is less well explored. Although existing physics-informed augmentation and simulation approaches can generate physically meaningful samples, they do not directly address the question of how many samples should be generated for specific sparse and process-relevant regions of a dataset.

D. SUMMARY AND RESEARCH QUESTIONS

Although the IL literature predominantly addresses continuous dataset growth and model adaptation, it provides limited guidance for phases in which data is scarce overall or sparse in regions most relevant to upcoming production. In industrial one-off or low-volume manufacturing, targeted data collection is impractical and uneconomical, making data-centric strategies essential. Existing data-centric methods improve dataset composition, while PiML methods incorporate physical knowledge into learning. However, the combination of both perspectives for a systematic, physics-informed dataset optimization remains insufficiently addressed. In particular, the effective selection of representative training data remains an open question for sampling-based PiML approaches [38].

To address this gap, we present PiDO, which involves two complementary mechanisms to systematically add physically consistent samples to a given dataset. The first, PiA, generates target values based on the inputs of existing data points. The second, PiE, generates physics-consistent input-output samples aligned with the regime of the upcoming part. In contrast to model-centric physics-informed or hybrid modeling approaches, which typically embed physical knowledge into the model architecture, loss function, or input representation, PiDO operates at the dataset level by using physical equations to optimize the training data prior to model learning. Therefore, the main contribution of this paper lies in the methodological formulation and systematic validation of the PiDO approach. Consequently, we pose the following research questions:

1) Can underlying physical equations be locally parametrized for physics-informed dataset optimization?
2) Can PiA and PiE sampling strategies be formulated to enrich datasets with physics-consistent samples?
3) To what extent can PiDO improve dataset balance and model performance by incorporating additional physical knowledge?

The novelty of this work lies in the PiDO framework, a data-centric physics-informed approach that connects *observational bias* and *learning bias*. It combines local parameterization of an underlying physical equation with the complementary data-generation mechanisms PiA and PiE. The challenge of selecting representative training samples is addressed through transferable sampling functions. The approach is validated on two milling machines and a series of parts with varying geometries, providing insight into the effect of optimizing incrementally growing training datasets with physically consistent samples.

III. METHODOLOGY

This section introduces the proposed PiDO methodology, detailing the underlying physical models and local parameterization in a reduced feature space, as well as the dataset optimization strategies.

A. SAMPLE APPLICATION AND SYSTEM OVERVIEW

The development of the proposed approach requires a ML application. For this work, a residual-based monitoring framework was chosen. In such a system, a hybrid ML-model [2] predicts axis-specific current signals $\hat{I}_i$ with the axis $i \in \{\mathrm{x}, \mathrm{y}, \mathrm{z}, \mathrm{sp}\}$ of a CNC milling machine, which serve as reference values and are continuously compared with the measured current signal I_i. The system is shown in Figure 1.

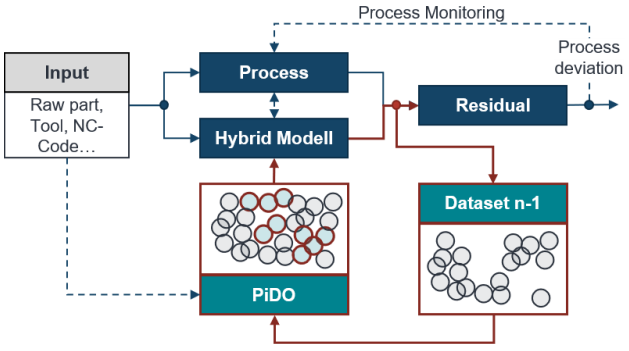

FIGURE 1. Incremental learning approach (red) as part of a residual-based process monitoring.

In line with the data-centric AI paradigm, the objective is to efficiently curate and optimize the hybrid model's training dataset. Therefore, the approach is ML-model agnostic. This ensures that PiDO is compatible with other IL approaches that continuously refine a given model. A specific training dataset is sampled for every new part produced, creating D_{train} in this scenario. It consists of all available measured samples D_{Real} and physically informed samples D_{PiDO}, tailored to the current process. The aim is to enrich and balance D_{train} with physical knowledge in order to achieve a more precise, axis-specific current prediction $\hat{I}_i$.

The specific sequence of the sampling process is shown in Figure 2. First, the *Input* of the upcoming part (NC-Code, tool geometry, and raw part geometry) is analyzed with respect to the given dataset $D_{\mathrm{Real_{n-1}}}$ by determining the relevant areas of the feature space. Subsequently, PEs are parametrized for all relevant areas. The dataset is then optimized with the data generated using the PE (PiA and PiE). Finally, $D_{\mathrm{Real_{n-1}}}$ and D_{PiDO} are combined to D_{train} for model training. New measured data becomes available after production, resulting in an updated $D_{\mathrm{Real_{n+1}}}$, which serves as the basis for the upcoming iterations.

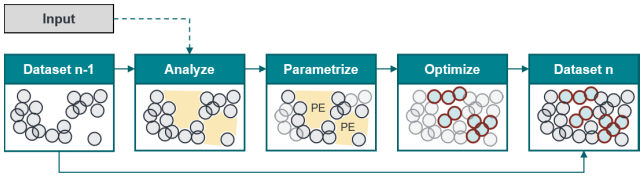

FIGURE 2. PiDO procedure: The existing dataset is optimized based on the input of the upcoming part (NC-Code, tool, and raw part geometry).

B. HYBRID AND PHYSICAL MODEL

The methodology of the hybrid model is based on the axis-level current signal prediction framework presented in [2]. It combines kinematic machine data, including axis position, velocity, and acceleration, with a process simulation (see Figure 3).

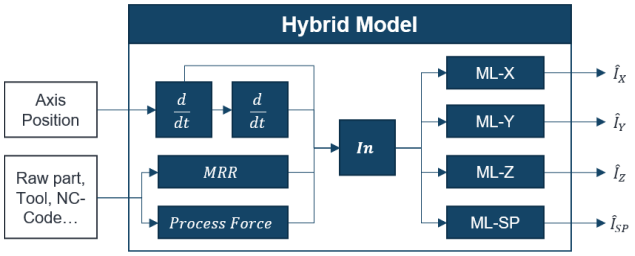

FIGURE 3. Hybrid model structure for reference signal prediction [2].

The simulation derives relevant signals that are difficult or expensive to measure, such as cutting forces and the Material Removal Rate (MRR). Analogously to [2], the process forces are calculated based on the Kienzle force model [39], [40] via $F_{\mathrm{cz}} = b \cdot h_m^{1-z} \cdot k_{c1.1} \cdot K$, $F_{\mathrm{cnz}} = b \cdot h_m^{1-x} \cdot k_{f1.1} \cdot K$, and $F_{\mathrm{pz}} = b \cdot h_m^{1-y} \cdot k_{p1.1} \cdot K$. The parameters $k_{c1.1}$, $k_{f1.1}$, $k_{p1.1}$, x, y, and z are material-specific values. The following [39], b and h_m are calculated based on

$$b = \frac{a_p}{\sin \kappa},$$

and

$$h_m = \frac{114,6}{\varphi_s} \cdot f_z \cdot \frac{a_e}{D} \cdot \sin \kappa.$$

To transform the forces into the machine coordinate system, the rotation formulation from [2] is applied:

$$\mathbf{F} = \begin{bmatrix} F_{\mathrm{x}} \\ F_{\mathrm{y}} \\ F_{\mathrm{z}} \end{bmatrix} = R_z(\vartheta) \cdot \begin{bmatrix} F_{\mathrm{cz}} \\ F_{\mathrm{cnz}} \\ F_{\mathrm{pz}} \end{bmatrix} \tag{1}$$

$R_z(\vartheta)$ describes a rotation around the z-axis. The resulting tangential spindle force F_{sp} is given by

$$F_{sp} = \sqrt{F_x^2 + F_y^2}, \tag{2}$$

which must be multiplied by the tool radius to get M_{sp}.

The voxel-based MRR is simulated as described in [2]. These signals then serve as inputs to axis-specific ML models, which output the current signals from the drives. In the case of a three-axis milling machine, the hybrid model structure can be described by axis- and spindle-specific ML models

$$M_i(\mathbf{In}) = \hat{I}_i, \text{ where } i \in \{x, y, z, sp\}, \tag{3}$$

which map the input vector $\mathbf{In}$ to the respective target $\hat{I}_x$, $\hat{I}_y$, $\hat{I}_z$, and $\hat{I}_{sp}$. The input vector is defined as

$$\mathbf{In} = \left[\dot{\mathbf{x}}^T \ \ddot{\mathbf{x}}^T \ \dot{\varphi} \ \ddot{\varphi} \ \mathbf{F}^T \ M_{sp} \ \mathrm{MRR}\right]^T, \tag{4}$$

where $\dot{\mathbf{x}}, \ddot{\mathbf{x}} \in \mathbb{R}^3$ denote the translational velocity and acceleration vectors, $\dot{\varphi}$ and $\ddot{\varphi}$ the angular velocity and acceleration, $\mathbf{F} \in \mathbb{R}^3$ the force vector, and M_{sp} the spindle moment.

A detailed derivation of the physical models $M_i(\mathbf{In})$ and the resulting current-dynamics equations can be found in [2]. The derivation relies on several simplifying assumptions that are relevant for the present work: The axis system is described by an energy balance, mass flows such as removed chips or adhering coolant are neglected, the internal energy of the axis system is assumed to be constant, and a constant voltage approximation is used to relate power consumption to current. Furthermore, coupling between axes and the spindle is approximated by first-order perturbation terms. Based on these assumptions and the following [2], PEs for axis-specific current consumption are obtained. For a coupled axis system, the exemplary approximate x-axis formulation derived in [2] is given by

$$\begin{aligned}
\hat{I}_x \approx\ & c_1 \cdot \ddot{x}_x \cdot \dot{x}_x \cdot \mathrm{sgn}(\dot{x}_x) + c_2 \cdot \dot{x}_x^2 \cdot \mathrm{sgn}(\dot{x}_x) \\
& + c_3 \cdot F_x \cdot \mathrm{MRR} + c_4 \cdot \mathrm{sgn}(\dot{x}_x) + c_5 \\
& + c_6 \cdot \ddot{x}_y \cdot \dot{x}_y \cdot \mathrm{sgn}(\dot{x}_y) + c_7 \cdot \dot{x}_y^2 \cdot \mathrm{sgn}(\dot{x}_y) \\
& + c_8 \cdot F_y \cdot \mathrm{MRR} + c_9 \cdot \mathrm{sgn}(\dot{x}_y) \\
& + c_{10} \cdot \ddot{x}_z \cdot \dot{x}_z \cdot \mathrm{sgn}(\dot{x}_z) + c_{11} \cdot \dot{x}_z^2 \cdot \mathrm{sgn}(\dot{x}_z) \\
& + c_{12} \cdot F_z \cdot \mathrm{MRR} + c_{13} \cdot \mathrm{sgn}(\dot{x}_z) \\
& + c_{14} \cdot \ddot{\varphi} \cdot \dot{\varphi} \cdot \mathrm{sgn}(\dot{x}_\varphi) + c_{15} \cdot \dot{\varphi}^2 \cdot \mathrm{sgn}(\dot{\varphi}) \\
& + c_{16} \cdot M_{sp} \cdot \mathrm{MRR} + c_{17} \cdot \mathrm{sgn}(\dot{\varphi})
\end{aligned} \tag{5}$$

Thus coupled PEs can be described by

$$M_i(\mathbf{In}) = \sum_{p=1}^{17} c_{p,i} X_p = \hat{I}_i, \tag{6}$$

with X_p listed in the appendix as formula 17. For the application within PiDO, these physical equations are not interpreted as globally exact descriptions across all operating regimes. Instead, they are assumed to provide locally valid approximations within constrained regions and are therefore parametrized accordingly.

C. INTEGRATION OF PIML FOR DATASET OPTIMIZATION

The approach aims to enrich an incrementally growing dataset with physically consistent data, which can therefore be classified as a combination of *learning* and *observational bias* [11]. *Learning biases* integrate a loss function to which an ML model M adapts during training, thus incorporating physical knowledge. As the PE shown in formula 6 is analytically solvable in the given case, there is no need for numerical differentiation during model training. This has the advantage that explicit data points can be generated from the analytical solution instead of using a residual term in the loss function to account for the physical equation. This can be formalized by

$$\tilde{\mathcal{L}} = \sum_{c \in \{\text{data, PE}\}} \bar{\omega}_c \, \mathbb{E}_{k \sim c}\left[\left\| M(\mathbf{In}_{c,k}) - I_{c,k} \right\|^2\right], \tag{7}$$

where $\bar{\omega}_{\text{data}}$ and $\bar{\omega}_{\text{PE}}$ are the weights used to balance the two loss terms. Each base weight ω_c is scaled by the fraction of data points in its category:

$$\bar{\omega}_c = \frac{N_c}{N_{\text{data}} + N_{\text{PE}}} \omega_c$$

D. FEATURE SPACE REDUCTION

As the identification of relevant target areas for PiDO is computationally intensive in a high-dimensional feature space, it must be reduced. To ensure clarity for a human interpreter, the resulting space should be representable through three clear and understandable dimensions. As [2] identified tree-based models as well suited the impurity-based feature importance of a random forest regressor is used to evaluate relevance. In addition, the widely used importance of permutation can be applied. Figure 4 illustrates the feature importance analysis of dataset DALG (aluminum) and DSTG (steel) from [41]. The evenly distributed tool path direction provides information about the relevance of individual input features for the respective axis. For I_x, the feature $\dot{x}_x$ has the greatest influence. I_y is most influenced by $\dot{x}_y$. And for both I_z and I_{sp}, MRR is highly influential.

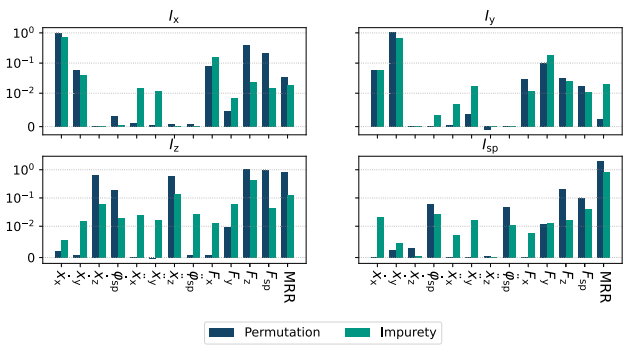

FIGURE 4. Feature importance on DALG and DSTG.

The MRR depends significantly on the components $\dot{x}_\mathrm{x}$, $\dot{x}_\mathrm{y}$, and $\dot{x}_\mathrm{z}$. In order to establish an orthogonally independent feature space $\mathbb{M}$ from the three most relevant features, MRR is normalized with the feedrate $\|\dot{x}\|$. This results in the area removal AR at a given time to

$$\mathrm{AR} = \frac{\mathrm{MRR}}{\|\dot{x}\|} = a_\mathrm{e} \cdot a_\mathrm{p} \cdot g_v,$$

where g_v describes the geometric filling degree, $g_\mathrm{v} \in (0, 1]$ (see figure 5). As AR represents a geometric quantity, it is easy to understand. The resulting feature space $\mathbb{M}$ is given by the dimensions $\dot{x}_\mathrm{x}$, $\dot{x}_\mathrm{y}$, and AR. It should be noted that it is used exclusively for the generation of PiDO data points. The ML models M_i still operate on the 13 features.

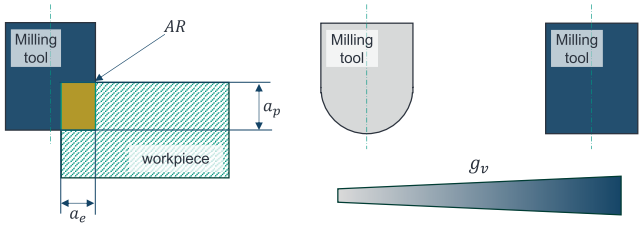

FIGURE 5. Visualization of AR and g_v.

The velocities $\dot{x}_\mathrm{x}$ and $\dot{x}_\mathrm{y}$ and the area removal AR are used to identify the input process of the upcoming part. To check where a process is located in $\mathbb{M}$, it is divided into individual areas B_{ijk}

$$B_{ijk} = \left\{ (\dot{x}_\mathrm{x},\ \dot{x}_\mathrm{y},\ \mathrm{AR}) \in \mathbb{R}^3 \;\middle|\; \begin{array}{l} \dot{x}_\mathrm{x} \in [\xi_i,\ \xi_{i+1}), \\ \dot{x}_\mathrm{y} \in [\eta_j,\ \eta_{j+1}), \\ \mathrm{AR} \in [\zeta_k,\ \zeta_{k+1}) \end{array} \right\}, \tag{8}$$

where ξ_i, η_j, and ζ_k denote the respective interval limits.

By focusing on dominant kinematic and geometric quantities, $\mathbb{M}$ enables the identification of consistent process states across varying operating conditions, allowing the parameterization of locally valid PEs. The smaller the bins, the more accurately the data points can be generated. However, since computational effort also increases, a suitable compromise must be found.

E. PE PARAMETERIZATION

To explicitly enforce locality, the PE-parameterization is adapted to the toolpath and process parameters within each relevant region B_{ijk} of the reduced feature space. As local data points are more similar to each other than to globally distributed data points, a range-specific PE is created for each range B_{ijk} in which the new process is located. To give more weight to local data points than to global data, the weighting function $w_{B_{ijk}}$ is introduced. The euclidean distance $\|q-p\|_2$ between two vectors $q, p \in \mathbb{R}^n$ is a suitable metric to distinguish between different areas within the feature space $\mathbb{M}$. To determine the minimum distance between individual areas of $\mathbb{M}$, the respective area center

$$\bar{\mathbf{b}}_{ijk} = \mathrm{mid}(B_{ijk}) = \frac{1}{2}\begin{bmatrix} \xi_i + \xi_{i+1} \\ \eta_j + \eta_{j+1} \\ \zeta_k + \zeta_{k+1} \end{bmatrix} \tag{9}$$

is used. The PE is parameterized on the basis of the entire measured dataset D_real. To weight local data points with respect to the spatial structure, a modified Gaussian weighting function based on [42] is used

$$w^{(n)}_{B_{ijk}} = \mathcal{N}\left(b^{(n)}_{b_{ijk}} \;\middle|\; \bar{\mathbf{b}}_{ijk}, \sigma^2\ \mathrm{diag}(1, 1, \alpha^{-2}\mathbf{I})\right) + \beta. \tag{10}$$

$\mathcal{N}$ represents a gaussian and $b^{(n)}_{b_{ijk}}$ describes the n-th data point ($n = 1, \ldots, N$) of D_real located in range B_{ijk}. The parameters are to be interpreted as follows:

- α: Scales the dimension AR relative to $\dot{x}_\mathrm{x}$ and $\dot{x}_\mathrm{y}$. This parameter determines how a change in the cross-sectional area of the milling tool affects the fit.
- β: Offset that ensures a minimum weight. This parameter determines how much the PE fit is affected by data points recorded under significantly different conditions and component geometries.
- σ: Controls the width of the weighting function, and thus balances local and global influences. A small σ results in a focus on data recorded under similar conditions. As σ increases, data from different components is given greater weight in the parameterization.

The coefficients $c_{p,i,B_{ijk}}$ of the PEs are determined for each range B_{ijk} containing data for the new component using a weighted least squares approach. For each range and each axis $i \in \{\mathrm{x},\ \mathrm{y},\ \mathrm{z},\ \mathrm{sp}\}$, a linear model of

$$\check{I}^{(n)}_{i,B_{ijk}} = \sum_{p=1}^{17} c_{p,i,B_{ijk}} X^{(n)}_p \tag{11}$$

is set up. The vector $\check{I}^{(n)}_{i,B_{ijk}}$ represents the current calculated by the PE for i, with one entry for each of the N data points located in the range B_{ijk}. The p entries of $X^{(n)}_p$ are listed in formula (17) in the appendix. The coefficients to be optimized for the range B_{ijk} are described by $c_{p,i,B_{ijk}}$. The optimal coefficients $\hat{c}_{p,i,B_{ijk}}$ are determined by minimizing the weighted quadratic deviation

$$\hat{c}_{p,i,B_{ijk}} = \underset{\tilde{c}_{p,i,B_{ijk}} \in \mathbb{R}}{\arg\min} \sum_{n=1}^{N} w^{(n)}_{B_{ijk}} \left[I^{(n)}_{i,B_{ijk}} - \check{I}^{(n)}_{i,B_{ijk}} \right]^2, \tag{12}$$

where $w^{(n)}_{B_{ijk}}$ denotes the weighting factor of the n-th real data point. $I^{(n)}_{i,B_{ijk}}$ are the real values, which are compared with the calculated values $\check{I}^{(n)}_{i,B_{ijk}}$ from formula 11. Given the optimal coefficients $\hat{c}_{p,i,B_{ijk}}$, the physically consistent currents

$$\hat{I}_{i,B_{ijk}} = \sum_{p=1}^{17} \hat{c}_{p,i,B_{ijk}} X_p, \qquad \forall X_p \in B_{ijk} \tag{13}$$

for a region B_{ijk} can be calculated. To use this approach, the optimal parameters α, β, and σ for scaling the weighting

function must be determined experimentally. Figure 6 visualizes the feature space $\mathbb{M}$ and the weighting of the data points for exemplary ranges B_{ijk}.

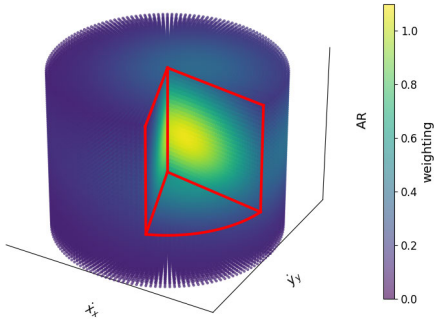

FIGURE 6. Exemplary weighting within the reduced feature space.

F. DATASET OPTIMIZATION

By checking the areas B_{ijk} of the characteristic space in which a process is located, it is possible to determine the number of real data points $n_{\text{real},B_{ijk}}$. Based on this, the dataset can be balanced with n PiDO data points in each area according to two methods.

PiA is the generation of augmented data points $\hat{I}$ within a defined target range. In contrast to PiE, measurement data $\mathcal{X}$ are used as input for the PE.

$$PE(\mathcal{X}) \rightarrow \hat{I}$$

PiE is the generation of physics-consistent input-output samples $\hat{\mathcal{X}}$ and $\hat{I}$ within a defined target range. Therefore, simulated parameters $\hat{\mathcal{X}}$ are used as the PE input.

$$PE(\hat{\mathcal{X}}) \rightarrow \hat{I}$$

Figure 7 illustrates the balancing via PiA and PiE for $A_{\text{real},B_{ijk}}$ with 30% and 70% of a given target density exemplarily.

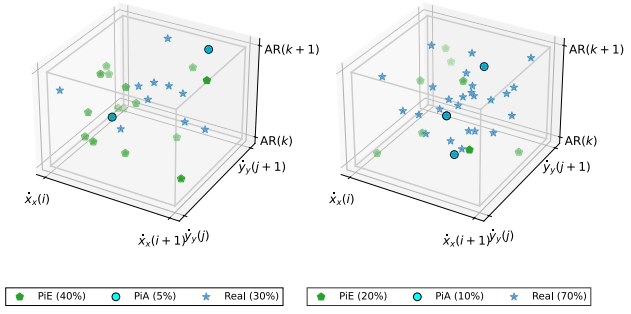

FIGURE 7. Areas B_{ijk} with different proportions of real, PiA, and PiE data. Data is exemplarily added for $A_{\text{real},B_{ijk}} = 30\%$, $A_{\text{real},B_{ijk}} = 70\%$ relative to a targed density.

The optimal ratio of n_{PiA} and n_{PiE} data points must be determined experimentally. To achieve a generally applicable approach across varying conditions, the sampling strategy must depend on the amount of real data available in the area B_{ijk}. The complementary approaches are introduced below.

1) PIA

To generate PiA data D_{PiA}, the real data D_{real} is used as input for the PE to calculate the current $\hat{I}_{\mathbf{P}}$. For each range $B_{ijk} \subset \mathbb{M}$ in which the new component G_{new} is located ($G_{\text{new}} \in B_{ijk}$), a sample of size n_{PiA} is generated from the associated real data $D_{\text{real}} \subset B_{ijk}$ via an equally distributed random selection without replacement. Let the total number of real data points available in this range be denoted by $n_{\text{real}} = |D_{\text{real}}|$. The selected sample $D_{\text{PiA}} \subset D_{\text{real}}$ is given by

$$D_{\text{PiA}} = \{z_1, \ldots, z_{n_{\text{PiA}}}\}, \quad z_i \sim \mathcal{U}(D_{\text{real}}), \quad \text{for } i = 1, \ldots, n_{\text{PiA}},$$

where the vectors $z_i = [\dot{x}_{\text{x}}, \dot{x}_{\text{y}}, \text{AR}]^T$ are interpreted as independent and identically distributed realizations of a uniform distribution $\mathcal{U}$ on the discrete set D_{real}. No further calculations of features are necessary since z_i refers to all relevant input parameters of **In**.

2) PIE

By contrast, generating PiE data D_{PiE} is more complex. Here, only the range limits of $\dot{x}_{\text{x}}$, $\dot{x}_{\text{y}}$, and AR are given for B_{ijk}. The components of the input vector **In** are not given and must be determined. Therefore, n_{PiE} data points

$$D_{\text{PiE}} = \{z_1, \ldots, z_{n_{\text{PiE}}}\}, \quad z_i \sim \mathcal{U}(B_{ijk}), \quad \text{for } i = 1, \ldots, n_{\text{PiE}},$$

are generated within a given range B_{ijk}, where $\mathcal{U}(B_{ijk})$ denotes a continuous uniform distribution. Thus, the points are uniformly distributed in the reduced feature space $\mathbb{M}$ and the features $\dot{x}_{\text{x}}$, $\dot{x}_{\text{y}}$ and AR can be determined. The process forces $F_{\mathbf{P}}$ for $i \in \{\text{x}, \text{y}, \text{z}, \text{sp}\}$ are given by formula 1 and 2. The acceleration $\ddot{x}_i$ for $i \in \{\text{x}, \text{y}, \text{z}\}$ cannot be specified unambiguously within B_{ijk}, and $\dot{x}_{\text{z}}$ is subject to fluctuations in the passive force. To ensure robustness under varying operating conditions, these are approximated statistically based on the given training data. The mean μ and the variance σ^2 within the 5%-95%-quantile are used to enable a robust determination of the relevant velocity and acceleration components. For $\dot{x}_{\text{z}}$, $\ddot{x}_{\text{x}}$, $\ddot{x}_{\text{y}}$ and $\ddot{x}_z$ all $X = \{x_1, x_2, \ldots, x_n\}$ within the 5%-95%-quantile $Q_{0.05}$ and $Q_{0.95}$ are used. The input parameters are approximated by $\dot{x}_{\text{z}}, \ddot{x}_{\text{x}}, \ddot{x}_{\text{y}}, \ddot{x}_{\text{z}} \sim \mathcal{N}(\mu, \sigma^2)$ with

$$\mu = \frac{1}{|X_Q|} \sum_{x_i \in X_Q} x_i \quad \text{and} \quad \sigma^2 = \frac{1}{|X_Q| - 1} \sum_{x_i \in X_Q} (x_i - \mu)^2.$$

A normal approximation is used as a pragmatic representation of the empirical variability within the selected quantile range. The spindle speed $\dot{\varphi}_{\text{sp}}$ is given by the NC code, which is controlled during the milling process. Thus, $\mathcal{N}(\mu = 0, \sigma^2)$ is applied to approximate the acceleration $\ddot{\varphi}_{\text{sp}}$. This statistical

representation reflects the physically admissible variability observed in previous measurements and avoids imposing deterministic trajectories that would reduce generalization.

PiE conducts a controlled exploration in sparsely populated regions relevant to the upcoming part, within the bounds of physics. Although this constitutes local extrapolation in $\mathbb{M}$, it remains constrained by analytically derived physical relationships, locality, machine kinematics, and information relating to the upcoming process.

G. ASSUMPTIONS AND SCOPE OF APPLICABILITY

The proposed PiDO-framework is based on the following assumptions, which determine its intended scope of application:

- Reduced feature space: The reduced feature space $\mathbb{M} = (\dot{x}_x, \dot{x}_y, \mathrm{AR})$ is assumed to capture the dominant kinematic and geometric influences needed to characterize the local data distribution with respect to the upcoming process. It is used to identify process-relevant regions for PiDO-based enrichment, while the ML models continue to operate on the full input space to ensure that no predictive information is lost.
- Local validity: The PEs describing the axis-specific current dynamics are assumed to provide locally valid approximations within constrained regions of the reduced feature space $\mathbb{M}$, rather than globally exact descriptions across all operating regimes. This locality is enforced by partitioning $\mathbb{M}$ into regions B_{ijk} and by weighting nearby data more strongly during PE parametrization.
- Statistical approximation of unresolved PiE variables: For PiE, input variables that cannot be uniquely specified from $\mathbb{M}$, such as accelerations, are assumed to be adequately represented by empirical distributions estimated from prior measurements. Since the inverse mapping from $\mathbb{M}$ to these variables is ill-posed, this probabilistic representation captures physically admissible variability while avoiding arbitrary deterministic trajectory assumptions.
- Controlled, physics-bounded extrapolation: For PiE, sparsely populated regions in $\mathbb{M}$ relevant for upcoming production are assumed to remain physically meaningful if the generated samples remain within the respective region B_{ijk} and are constrained by machine kinematics, NC-code information, locally parametrized PEs, and the statistical ranges observed in prior measurements. Under this assumption, PiE represents controlled local extrapolation rather than unconstrained model extrapolation.

Together, these assumptions define the scope within which PiDO can generate physically consistent samples and improve robustness across varying operating conditions.

IV. EXPERIMENTAL DESIGN

The scaling parameters α, β, and σ of the weighting functions, as well as the function to generate PiA and PiE data points, must be determined. Therefore, two experiments are conducted. The first analyzes the parameterization of the physical equations and the second investigates the effects of different dataset optimization strategies.

A. DATASETS AND SETUP

Two workpiece geometries (Gear G and Plate P) are selected from the dataset described in [41]. Both were recorded on a DECKEL MAHO DMC 60H milling machine using a 10 mm diameter end mill. Each dataset consists of three iterations: three in the baseline configuration, or a low, a baseline, and a high configuration. For a_p and SF, the high value is referred to as h and the low value as l. The baseline axial cutting depth is $a_p = 6\,\mathrm{mm}$ and the radial cutting depth is $a_e = 10\,\mathrm{mm}$. For the steel material S235JR (ST), the feedrate $F = 576\,\mathrm{mm/min}$ and the spindle speed $S = 3200\,\mathrm{min}^{-1}$, while for the aluminum AL 2007 T4 (AL) $F = 350\,\mathrm{mm/min}$ and $S = 3800\,\mathrm{min}^{-1}$ in the baseline setup. In the datasets DAL/DST_{Xa_p}, a_p ranges from 3 mm to 12 mm. The feedrate F varies in $DAL_{PSF/GSF}$ between 267 mm/min and 442 mm/min, while S ranges between 2900 min^{-1} and 4800 min^{-1}. For S235JR, the feedrate F varies in $DST_{PSF/GSF}$ between 432 mm/min and 720 mm/min, while S varies between 2400 min^{-1} and 4000 min^{-1}. The datasets are listed in table 1.

TABLE 1. Datasets extracted from [41].

Machine	Material	Name of the datasets
DMC 60H	AL 2007 T4	DAL_G, DAL_P, DAL_{GSF}, DAL_{PSF}, DAL_{Ga_p}, DAL_{Pa_p}
	S235 JR	DST_G, DST_N, DST_P, DST_{GSF}, DST_{PSF}, DST_{Ga_p}, DST_{Pa_p}

To extract the tool engagement sections in the reduced feature space $\mathbb{M}$, $0 < \mathrm{AR} < 200$ and $\|\dot{x}\| \neq 0$ is applied to the datasets, which are down sampled to 50Hz. MRR and the forces F_{cz}, F_{cnz} and F_{pz} (see section III-B) are calculated based on [2] using the coefficients $k_{c1.1} = 1990\mathrm{N/mm}^2$, $z = 0.7019$, $k_{f1.1} = 351\mathrm{N/mm}^2$, $x = 0.4911$ and $k_{p1.1} = 247\mathrm{N/mm}^2$, $y = 0.2$ for S235 JR and $k_{c1.1} = 472\mathrm{N/mm}^2$, $z = 1.08$ $k_{f1.1} = 20\mathrm{N/mm}^2$, $x = 0.75$ and $k_{p1.1} = 32\mathrm{N/mm}^2$, $y = 0.2$ for AL 2007 T4.

In accordance [2], an extra tree model (estimators = 30, max depth = 100, leafs = 5, nodes = 1000) is selected for the x-, y-, and z-axes. Concurrently, a gradient boosting model (estimators = 100, nodes = 6000) is selected for the main spindle, with both operating at 50Hz.

B. EXPERIMENT 1: PE PARAMETERIZATION

The experiment aims to determine the optimal parameters (α, β, and σ) of the weighting function $w_{B_{ijk}}^{(n)}$ (formula 10).

The experimental design considers different characteristics, including material, part geometry, feedrate v_c, spindle speed $\dot{\varphi}$, and depth of cut a_p. This configuration results in four main experiments for DAL_G, DAL_P, DST_G, and DST_P.

Algorithm 1 Setup for the Determination of α, β, and σ

for $m = 1$ **to** 30 **do**
  $D_{\text{test}}, D_{\text{train}}$
  for all $B_{ijk} \in \mathbb{M}$ **do**
    if $B_{ijk} \cap D_{\text{test}}$ **then**
      $\text{PE}_{ijk} \leftarrow B_{ijk} \cap D_{\text{test}}$
    end if

$$\begin{aligned} \underset{\alpha,\beta,\sigma}{\text{minimize}} \quad & \text{RMSE}\left(I_{B_{ijk}}, \hat{I}_{B_{ijk}}\right) \\ \text{subject to} \quad & 0 \le \alpha \le 2.5, \\ & 0 \le \beta \le 1, \\ & 0 \le \sigma \le 5, \end{aligned}$$

    use Optuna with $n_{\text{optuna}} = 100$ to solve the optimization problem above for (α, β, σ)
    $w_{B_{ijk}} \leftarrow w_{B_{ijk}}(\alpha, \beta, \sigma)$
    $\hat{c}_{B_{ijk}} \leftarrow \hat{c}_{B_{ijk}}(D_{\text{train}}, w_{B_{ijk}})$
    $\hat{I}_{B_{ijk}} \leftarrow \hat{I}_{B_{ijk}}(\hat{c}_{B_{ijk}}, D_{\text{test},B_{ijk}})$
  end for
end for

A complex but almost evenly distributed component in the $\dot{x}_\text{x}$-$\dot{x}_\text{y}$-plane is given by G. By contrast, P enables examination of the parameters characteristics of a component that is not uniformly distributed with respect to this plane. For each experiment, a leave-one-out strategy is applied to the datasets SF_l, SF_h, D_l, and D_h and one baseline iteration. Therefore, one dataset is used as the test dataset D_{test} and the remaining ones serve as the training dataset D_{train}. The variation of v_c and a_p enables a differentiated analysis of the impact of all relevant dimensions of the reduced feature space.

The predicted values $\hat{I}_{B_{ijk}}$ are obtained using the corresponding parameterized PEs, which can be compared to the actual values $I_{B_{ijk}}$. For all currents signals I_x, I_y, I_z, and I_sp, RMSE is calculated according to

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{l=1}^{n}\left(I_{B_{ijk,l}} - \hat{I}_{B_{ijk,l}}\right)^2}, \tag{14}$$

where n denotes the number of observations.

The range limits of B_{ijk} are defined according to formula 8 over the feature space $\mathbb{M} = \left(\dot{x}_\text{x}, \dot{x}_\text{y}, \text{AR}\right)$. The value ranges are defined as $\dot{x}_\text{x}$, $\dot{x}_\text{y} \in \{-15, -14, \ldots, 15\}$ and $\text{AR} \in \{0, 2, 4, \ldots, 200\}$, according dataset limits. The grid size is set to $\xi_{i+1} - \xi_i = 1\,\text{mm/s}$, $\eta_{j+1} - \eta_j = 1\,\text{mm/s}$, $\zeta_{k+1} - \zeta_k = 10\,\text{mm}^2$. All areas B_{ijk} of D_{test} in $\mathbb{M}$ are determined, and a separate PE is set up for every area. These are parameterized on the respective training data D_{train} with the weighting function $w_{B_{ijk}}^{(n)}$ (see formula 10).

The optimization framework Optuna is used to optimize α, β, and σ in $n_{\text{optuna}} = 100$ loops. To generate statistically valid results, all trials are repeated 30 times. The procedure is shown in algorithm 1.

C. EXPERIMENT 2: DATASET OPTIMIZATION

The aim of this experiment is to derive a general sampling strategy for PiA and PiE data, depending on the real data proportion per range B_{ijk}. Therefore, the influence of the PiA proportion A_{PiA} and the PiE proportion A_{PiE} is investigated for varying real data proportions A_{real}.

DAL_{GSF_l} is selected as test dataset because it contains the highest number of data points per area and is evenly distributed across the feature space. To obtain a universal function, it must be as independent as possible from toolpath directions, material, and technological parameters. Therefore, the Wasserstein metric [43] is used to examine to what extent different training datasets differ from DAL_{GSF_l} (see figure 33 in the appendix). To ensure uniform coverage and thus a representative selection, a total of nine equidistantly distributed datasets are selected (with Wasserstein distances $DAL_{GN} = 5.730$, $DST_{GSF_l} = 14.803$, $DST_{PD_l} = 32.948$, $DST_{PN} = 52.602$, $DST_{GSF} = 62.066$, $DST_{PSF_h} = 76.161$, $DST_{GSF_h} = 83.051$, $DST_{PD_h} = 113.471$ and $DST_{GD_h} = 125.544$). Each of these is used individually as training dataset D_{train}, while $\text{DAL}_{\text{GSF}_\text{l}}$ serves as a constant test dataset $D_{\text{test}} = D_{\text{next}}$.

To obtain the necessary training data within the areas of the test dataset D_{next}, a 70/30 train-test split is performed, resulting in $D_{\text{next,train}}$ and $D_{\text{next,test}}$ (see algorithm 3 in the appendix). To ensure uniform training datasets $D_{\text{next,train}}$ and D_{train}, all areas must have the same number of data points for $A_{\text{real}} = 100\,\%$. Therefore, an upper limit is introduced to which the data within each area is randomly reduced. Figure 8 shows that this condition is satisfied with an upper limit of 40, resulting in 12 areas that form a complete circle in the $\dot{x}_\text{x}$-$\dot{x}_\text{y}$-plane. The resulting areas are given by the range limits $\xi_{i+1} - \xi_i = 2\,\text{mm/s}$, $\eta_{j+1} - \eta_j = 2\,\text{mm/s}$, $\zeta_{k+1} - \zeta_k = 2\,\text{mm}^2$. This is a compromise between the upper limits and the area that can be covered.

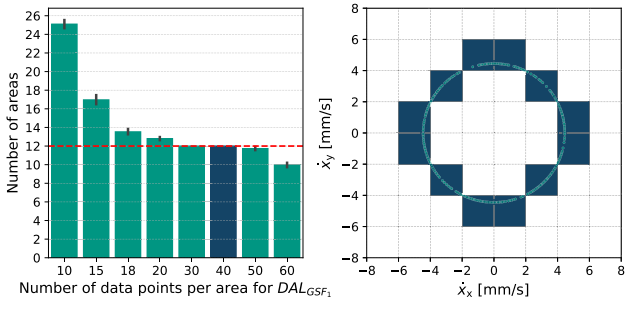

FIGURE 8. Analysis of dataset DAL_{GSF_l} for selecting the upper limit of data points per area B_{ijk} in the reduced feature space. The left-hand plot shows the number of available data points per area, and the right-hand plot shows the resulting area distribution when an upper limit of 40 data points is applied.

The varying real data proportions are achieved by reducing $D_{\text{next,train}}$ to a proportion A_{real} of $\{0\,\%, 10\,\%, \ldots, 100\,\%\}$ resulting in $D_{\text{next,red}}$. PE data points are generated for each

area B_{ijk} of $D_{\text{next,test}}$ with the weighting function $w_{B_{ijk}}$ and α, β, and σ determined in Experiment 1.

A full factorial experimental design is carried out for all real data proportions. Therefore, the percentage proportions for PiA is set to $\{0, 3, 5, 8, 10, 15, 25\}$ and for PiE to $\{0, 3, 5, 10, 15, 20, 35, 50, 65, 80, 95, 110\}$ and combined in all configurations. The range for PiA is set to a higher resolution for lower proportions to capture finer effects. Furthermore, the maximum amount is limited to 115% of the target density, and the PiA proportion is limited to a maximum of the size of the respective real data (1:1 ratio) to avoid overrepresentation and structural distortions. This results in a PiE proportion between 0 % and 110 %, and a PiA proportion between 0 % and 25 %. Figure 9 visualizes the PiA and PiE proportions.

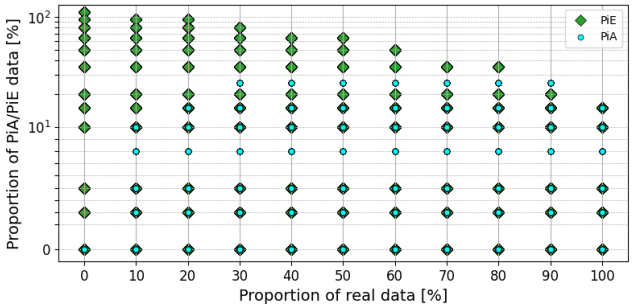

FIGURE 9. Examined PiA and PiE proportions as a function of the real data proportion.

The corresponding number of PiA and PiE data is generated using the PE and combined with $D_{\text{real,red}}$ and D_{train} to form D_{model}. This dataset is used for model training, whereby the predicted current values $\hat{I}_i$ are compared with the current values I_i from $D_{\text{test,test}}$ using RMSE.

V. EXPERIMENTAL RESULTS

This section presents the experimental results for the PE-parameterization and the sampling strategies, serving as the empirical basis for the subsequent methodological consolidation.

A. EXPERIMENT 1: PE PARAMETERIZATION

Figure 10 presents the KDE distributions of the scaling parameters α, β, and σ. The top row compares the two materials aluminum and steel. The middle row further differentiates this representation based on dataset D . . .$_{\text{G}}$ and D . . .$_{\text{P}}$. The bottom row shows the results of DAL_{G} and DST_{G} for each axis.

Table 2 shows the mean, the modal value, the median, and the standard deviation of the distributions of α, β, and σ separated by axis.

All distributions show the same overall trend, particularly in terms of part and material specifications. The values of α and β are concentrated close to 0 with decreasing density, while σ is located in the range around 4.

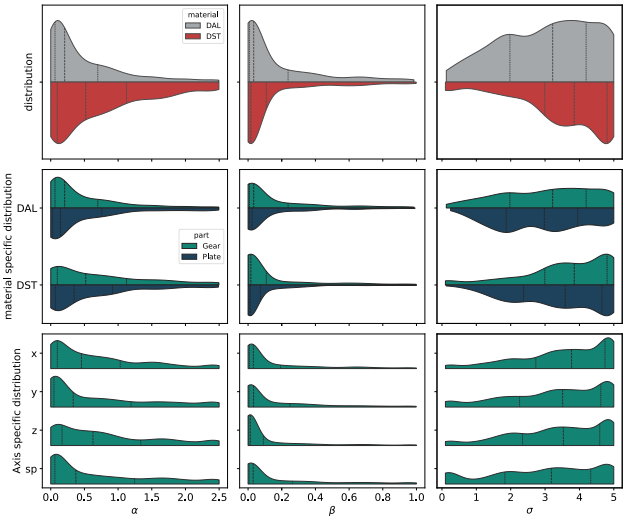

FIGURE 10. KDE distribution of α, β, and σ for varying materials (aluminum and steel) part geometries and axes (DAL_{G} and DST_{G}).

TABLE 2. Comparison of mean, modal value, median, and standard deviation of the KDE distribution per axis for the parameters.

Parameter	Axis	Mean	Modal value	Median	Std.
α	x	0.7539	**0.0976**	0.4556	0.6848
	y	0.8078	**0.0701**	0.3336	0.7836
	z	0.8758	**0.1126**	0.6269	0.7465
	sp	0.8139	**0.0826**	0.3730	0.7769
β	x	0.2020	**0.0200**	0.0313	0.2364
	y	0.2212	**0.0170**	0.0325	0.2471
	z	0.1641	**0.0130**	0.0133	0.2136
	sp	0.2395	**0.0170**	0.0304	0.2704
σ	x	3.3809	**4.8381**	3.7711	1.3052
	y	3.1811	**4.8086**	3.5114	1.4255
	z	3.2309	**4.8327**	3.5329	1.3410
	sp	2.9423	**4.7882**	3.1845	1.5635

B. EXPERIMENT 2: DATASET OPTIMIZATION

The parameters for α, β, and σ determined in experiment 1 serve as the basis for the second experiment and for each axis the respective modal value is selected.

Figure 11 and 12 show the mean PiA and PiE values for the selected datasets and the varying real data ratios. For each entry, the PiA and PiE proportion with the lowest RMSE is determined 30 times and the mean is calculated. Figure 11 shows the axis-specific results for the PiA proportions. It is noticeable that the PiA proportions of I_{sp} and I_{z} tend to be larger (up to 8%) than those of I_{x} and I_{y} (up to 3%). Figure 12 presents the PiE proportions, which are generally larger than the PiA proportions. Compared to PiA, higher similarity emerges for the different axes, whereby the PiE proportions tend to be slightly higher for I_{x}, I_{y}, and I_{z} than for I_{sp}.

To identify a sampling function independent of a specific axis configuration, it must be averaged. The resulting proportions are shown in figure 13. The proportion of PiA data increases with an increase in the amount of real data.

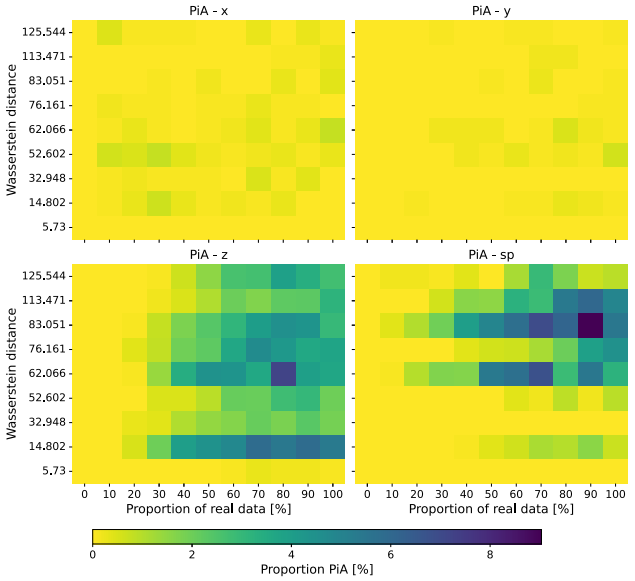

FIGURE 11. Optimal PiA proportions as a function of the real data proportion of the target density within a voxel across datasets with varying Wasserstein distances.

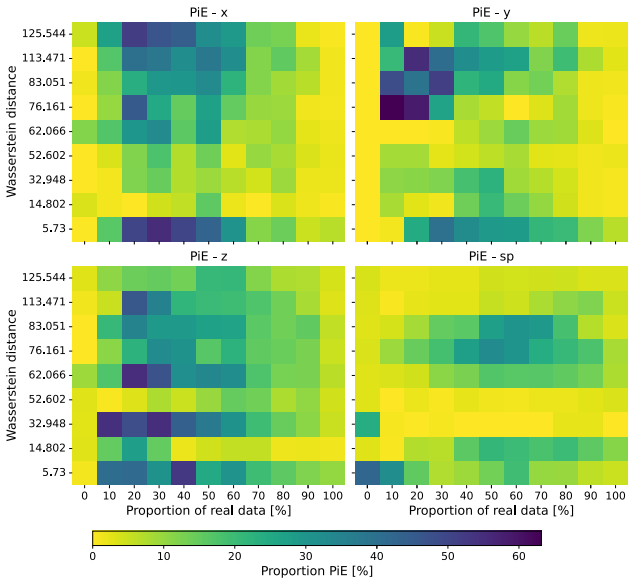

FIGURE 12. Optimal PiE proportions as a function of the real data proportion of the target density within a voxel across datasets with varying Wasserstein distances.

In contrast, the proportion of PiE starts low, rises, and finally decreases as the proportion of real data increases. Figure 14 shows the distribution of the mean PiA and PiE proportions of figure 13. It can be seen that the dispersion of the PiA proportions is comparatively small, whereas the range of the PiE proportions is higher.

VI. DISCUSSION AND METHOD CONSOLIDATION

This section interprets the experimental results and consolidates the PE parameterization and sampling strategies

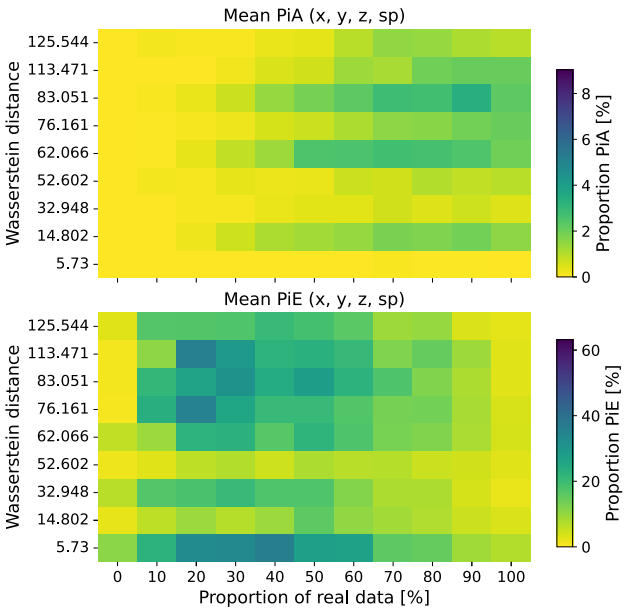

FIGURE 13. Mean PiA and PiE proportions as a function of the real data proportion of the target density within a voxel across datasets with varying Wasserstein distances.

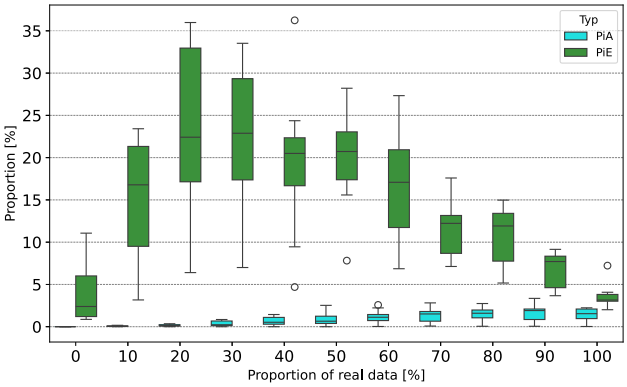

FIGURE 14. Mean PiA and PiE distributions across all axes.

into a final PiDO configuration. The discussion focuses on explaining how the observed parameter distributions of the weighting function and the PiA/PiE sampling proportions support the configuration selected for validation.

A. EXPERIMENT 1: PE PARAMETERIZATION

The parameter distributions of α, β, and σ follow the multi-factorial patterns resulting from material, component geometry, axis, feedrate, and cutting depth. Axis-specific parameters are selected to obtain functions that are independent of toolpath and technological parameters (see table 2). The modal value is used since it represents the most frequent value. This choice favors transferability over dataset-specific optimization, which is essential for applying the same PiDO configuration to varying geometries and process conditions without re-identify the scaling parameters.

Comparing materials, aluminum has lower values for α and σ than steel, which is related to the distributions of the feedrate v_c in the datasets. In the steel dataset, the values of v_c range from -12 to $+12mm/s^{-1}$, while in the aluminum dataset they vary between -8 and $+8mm/s^{-1}$. Figure 15 shows the feedrate distributions for aluminum and steel for the G part. Since the range of AR is similar in both datasets, a different α is required to weight the components comparably. Lower β occurs for steel sugesting a reduced weighting of distant data points.

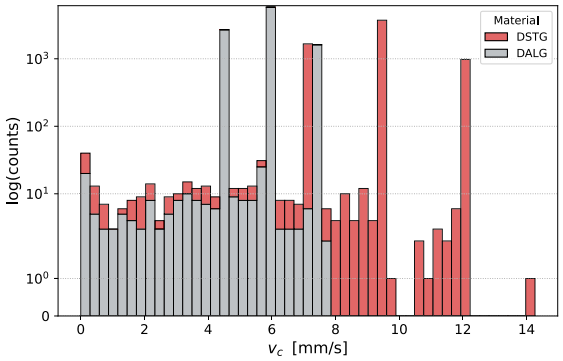

FIGURE 15. Feedrate distribution v_c for DAL_G and DST_G.

When comparing G and P for both materials, the parameters show similar distributions. This similarity supports uniform parameterization for α, β, and σ. A deviation can be seen in σ, where slightly lower σ values occur for P than for G. This behavior can be explained by the density in the feature space (see figure 16), where the accumulation in the positive and negative $\dot{x}_x$ direction increases the local influence, to which σ responds with a lower value. This observation confirms the use of G for the identification of universal parameters based on an evenly distributed dataset.

The axis specific evaluation of the G dataset differentiates the superimposed distributions.

B. EXPERIMENT 2: DATASET OPTIMIZATION

The difference in the axis-specific PiA proportions can be explained by the fact that the interpolated feed axes x and y exhibit more complex dynamics, resulting in a higher information content of a given dataset. The z axis is passive and influenced by its coupling to the other axes, resulting in a higher benefit for augmented data and the resulting increased variability. The qualitative progression of the PiE proportions is largely similar for x, y, and z, whereas the sp axis shows slightly lower PiE proportions. This suggests that PiE has a similar effect on all axes. The distribution of the mean PiA and PiE proportions over all axes is shown in Figure 17 based on the real data proportion (left) and the Wasserstein distance (right).

The distance correlation according to [44] is used as metric to investigate the influences on the PiA and PiE proportions. In contrast to other coefficients, the distance correlation can capture nonlinear relationships. A value of 0 indicates

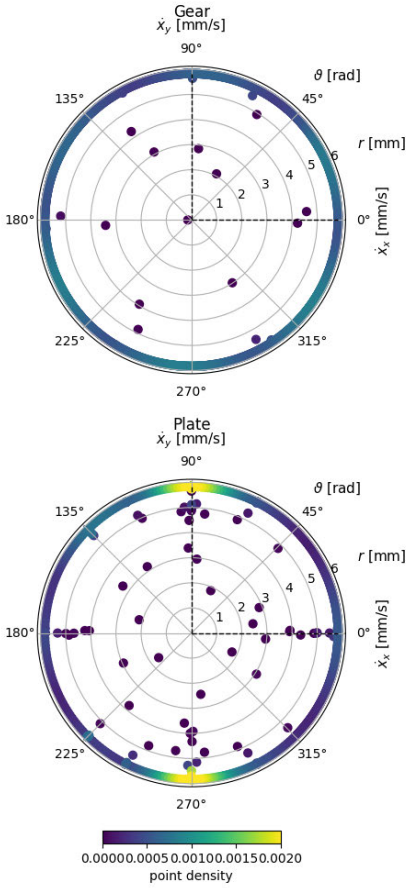

FIGURE 16. Density distribution of datasets DAL_{GN_1} and DAL_{PN_1}.

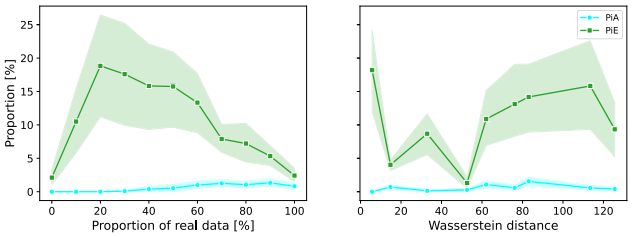

FIGURE 17. Mean and 95% quantile of median PiA and PiE proportions.

no relationship, and values close to 1 indicate a strong correlation. For the real data proportion, the correlations are 0.6832 and 0.4552 for PiA and PiE, respectively, while the Wasserstein distances are 0.2830 and 0.2185, respectively. Unified modeling based on the real data proportion to reduce complexity is supported by the higher values. This indicates that the real-data proportion is the dominant practical indicator for determining the need for PiA and PiE enrichment, whereas the Wasserstein distance provides additional information but leads to a less compact and transferable sampling rule.

Based on this, a general function for the PiA and PiE proportions is determined. Table 3 shows the R^2 and *RMSE*

TABLE 3. R^2 and *RMSE* for PiA and PiA with a polynomial fit of order *n*.

Order n	R^2_{PiA}	R^2_{PiE}	$RMSE_{\mathrm{PiA}}$	$RMSE_{\mathrm{PiE}}$
1	0.131068	0.036838	1.472389	13.312309
2	0.133493	0.192792	1.470333	12.186990
3	0.141676	0.230116	1.463374	11.901897
4	**0.141752**	**0.234954**	**1.463309**	**11.864446**
5	0.141753	0.235201	1.463309	11.862530
6	0.141822	0.235233	1.463249	11.862281
7	0.141827	0.236800	1.463246	11.850110

for a fit with polynomial degrees $n = 1, \ldots, 7$. The low R^2 highlights the difficulty of deriving a universally applicable sampling function due to the superposition of axis-specific, geometric, and process-dependent effects. Nevertheless, the gradual decrease in RMSE indicates that the polynomial fit captures a relevant trend in the required PiA and PiE proportions. Therefore, a fourth-order polynomial function is used, as it provides a compromise between flexibility, interpretability, and moderate model complexity. Thus, the function is given by $f_{PiA/PiE}(x) = a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$, where a_n with $n \in \{0, 1, 2, 3, 4\}$ are the coefficients, and x is the proportion of real data. A least-squares fit results in

$$f_{\mathrm{PiA}}(x) = -5.908 \cdot 10^{-8}x^4 + 3.653 \cdot 10^{-6}x^3 + 4.098 \cdot 10^{-4}x^2 - 1.017 \cdot 10^{-2}x + 1.768 \cdot 10^{-2} \quad (15)$$

for PiA and

$$f_{\mathrm{PiE}}(x) = -1.119 \cdot 10^{-6}x^4 + 3.314 \cdot 10^{-4}x^3 - 3.522 \cdot 10^{-2}x^2 + 1.335\,x + 1.766 \quad (16)$$

for PiE. The regression curves of formula 15 and 16 are shown in figure 18.

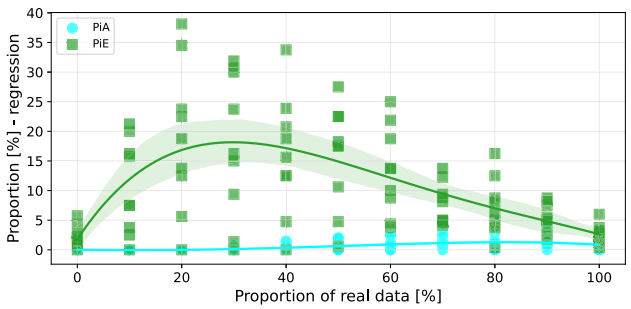

FIGURE 18. Regression of the sampling proportions as a function of the real data proportion of the target density within a voxel.

The resulting PiA and PiE functions are shown in Figure 19 in combination with the real data proportion. The upper figure shows the absolute sampling proportions based on the given upper limit of 40. The lower figure shows the proportion relative to the total amount of data in a given bin.

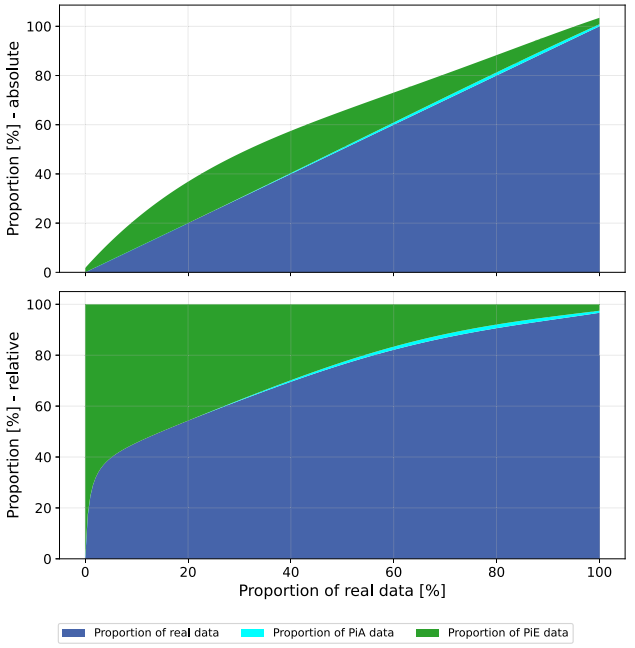

FIGURE 19. Absolute proportions (top) and proportions relative to the total data (PiA, PiE, and real) within a voxel (bottom).

VII. VALIDATION

This section validates the consolidated PiDO configuration in an ablation-based evaluation on independent datasets across two machines.

A. MACHINES AND DATASETS

The validation aims to identify the influence of PiDO on dataset and model performance in a realistic production scenario. Therefore, the series of impeller components from [45] is used, which is characterized by highly variable geometries. The series is manufactured from POM-C on a DMC 60H [46] and a CMX 600V [47]. In comparison to the datasets of the previous experiments, it contains different toolpaths, machines, and materials, resulting in different technological parameters, such as feedrate, spindle speed, and cutting tool. A total of 12 geometries $DPOM_{\mathrm{IMP1..12}}$ are available for the DMC 60H. $CPOM_{\mathrm{IMP4}}$ was excluded for the CMX 600V due to incomplete measurement data, resulting in 11 usable datasets. To ensure comparability between both machines, the 11 geometries available in both machines are used for validation. Figure 20 visualizes exemplarily POM_{IMP1}, POM_{IMP2}, POM_{IMP3} and POM_{IMP12}.

The process forces and the MMR are calculated according to section III-B. The forces were determined using a force measurement platform. The coefficients for F_{cz}, F_{cnz} and F_{pz} are obtained via an *Optuna* optimization, resulting in $k_{c1.1} = 260N/mm^2$, $z = 0.246$, $k_{f1.1} = 320N/mm^2$, $x = 0.23$, and $y = 0.2$. The coefficient $k_{p1.1}$ is set to $1N/mm^2$ since it scales within the parameterization of the PEs and the ML model.

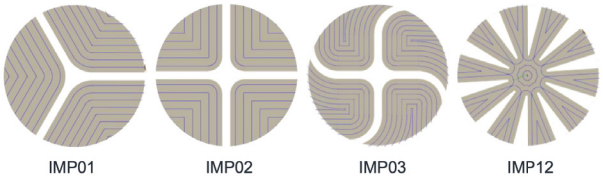

FIGURE 20. Geometries of exemplary POM_{IMP} parts [45].

Algorithm 2 - Val1: Validation of the PE-Parametrization

$D_{\text{test}}, D_{\text{train}}$
for alle $B_{ijk} \in \mathbb{M}$ **do**
  if $B_{ijk} \cap D_{\text{test}}$ **then**
    $\text{PE}_{ijk} \leftarrow B_{ijk} \cap D_{\text{test}}$
  end if
end for
$\hat{c}_{B_{ijk}} \leftarrow PE(D_{\text{train}} \cdot w_{B_{ijk}}(\alpha, \beta, \sigma))$
$\hat{I} \leftarrow PE_{\hat{c}_{B_{ijk}}}(D_{\text{test},B_{ijk}})$
$\text{RMSE}(I, \hat{I})$
$R^2(I, \hat{I})$

B. EXPERIMENTAL SETUP

The following describes the validation process for the PE parameterization and the sampling strategy.

1) VALIDATION 1: PE PARAMETERIZATION

The aim is to evaluate the predictive performance of the proposed approach for several PEs in comparison to a global PE parameterized on the entire dataset. To create a challenging scenario, the datasets used for PE parameterization are selected to have limited coverage of the test dataset areas. Therefore, $\text{DPOM}_{\text{IMP01}}$, $\text{DPOM}_{\text{IMP02}}$ and $\text{DPOM}_{\text{IMP03}}$ are combined to D_{train} and $\text{DPOM}_{\text{IMP12}}$ serves as test dataset D_{test} (see [46]). Thus, D_{train} does not cover all movement directions of D_{test}.

For the benchmark (global PE), one PE is parameterized on D_{train}. To validate approach presented (local PEs), a separate PE is created for each area B_{ijk} of D_{test}. The coefficients $\hat{c}_{B_{ijk}}$ of the respective PE are determined via the weighting function $w_{B_{ijk}}$ and the parameters α, β, and σ. The PEs of both approaches are used to calculate the current value $\hat{I}$, which is compared to the actual current values I using RMSE and R^2. Algorithm 2 describes the procedure schematically.

2) VALIDATION 2: ABLATION STUDY OF THE SAMPLING STRATEGY

This experiment aims to verify the transferability of the derived sampling functions and the impact of the PiDO approach. The individual and combined effects are therefore systematically assessed using the methodological principles of an ablation study. Therefore, the performance of models trained on optimized datasets are compared with that of models trained solely on real data. The examined training data configurations include:

- Real data D_{real} as baseline
- Real data enriched by PiA data $D_{\text{real+PiA}}$
- Real data enriched by PiE data $D_{\text{real+PiE}}$
- Real data enriched by PiA and PiE data $D_{\text{real+PiA+PiE}}$

To evaluate the impact on performance, the $\text{RMSE}_{PiDO,i}$ of the respective axis i with $i \in \{\text{x}, \text{y}, \text{z}, \text{sp}\}$ is subtracted from its baseline, thus $\Delta\text{RMSE}_{PiDO,i} = \text{RMSE}_{\text{D}_{\text{real}},i} - \text{RMSE}_{\text{D}_{\text{PiDO}},i}$. The mean over all axis ΔRMSE_{PiDO} is given by $\frac{1}{4}\sum_{i\in\{\text{x,y,z,sp}\}} \Delta\text{RMSE}_{PiDO,i}$.

The IMP datasets from both machines (DMC and CMX) are used in identical validation protocols. These differ from the development datasets in terms of toolpath and technological parameters, and in case of the CMX dataset, in terms of machine and axis configuration. Therefore, they enable an evaluation of the robustness and cross-machine generalisation. The investigation will consider two scenarios:

Scenario 1 addresses the influence of data poverty. Therefore, $n = 1$ resulting in one training file D_{train} and one test file D_{test}. Both are randomly selected for every $m = 30$ iterations. Data of D_{train} is added successively in steps of 10 % to simulate incremental data avaliability.

In **Scenario 2**, a random sequence $D_{\text{train,list}} = \text{POM}_{\text{IMP}t} \mid t \in (\mathbb{N} \cap [1, 12]) \setminus \{4\}$ is generated based on the given datasets. Initially, the first dataset of $D_{\text{train,list}}$ is used for training and the second for testing. To ensure comparability between both scenarios, an identical initial dataset is used. Subsequently, the first and second are combined and used for training while the third is used for testing. Thus, the number of impellers represented in D_{train} increases successively. To generate statistically valid results, the procedure was repeated 30 times with varying sequences $D_{\text{train,list}}$.

Scenario 1 focuses on evaluating robustness in the context of limeted data and incremental dataset growth. In contrast, Scenario 2 aims to assess generalization across different part geometries and toolpaths. As scenario 2 progresses, the amount of real data available for model training will increase. This enables an investigation into whether the ML model can continue to capture the complex relationships given by real data, or whether it is biased towards the simplified PEs. The use of unseen datasets, cross-machine evaluation, and random sequences enables a statistically robust assessment. A Shapiro–Wilk test is used to determine whether the 30 samples are normally distributed. If so, a two-tailed sample test is performed to analyze significance. If the distribution deviates from normal, a two-tailed Wilcoxon signed-rank test is performed. The significance level is set to $\alpha = 0.05$. [48]

In both scenarios, the training data is reduced to the upper limit $\rho_{target} = 40$, and the areas B_{ijk} where the test data is located are identified. For those, the sampling function is used to calculate the respective proportion and thus the number of PiA and PiE data to be generated. The PEs coefficients $\hat{c}_{B_{ijk}}$ are derived from the training data via the weighting function $w_{B_{ijk}}$ and the identified α, β, σ. These are then used to generate the additional PiA and PiE data.

The IMP datasets show a high data density for a radius of $r = 8$, while a lower density is given for $r < 8$. As can be seen in figure 21, the distribution of data points along the AR axis is not uniform. This allows for a differentiated analysis in two spaces. Besides the entire feature space $\mathbb{M}$ for which $C := \mathbb{M}$ applies, a subspace can be introduced as $D := \{\mathbf{m} \in \mathbb{M} \mid 12 \leq m_{\mathrm{AR}} \leq 30 \ \wedge \ r(\mathbf{m}) \leq 6\}$, representing a cylindrical range with radius 6 in the interval $12 \leq \mathrm{AR} \leq 30$. Figure 21 illustrates D with its limits in the $\dot{x}_\mathrm{x}$-$\dot{x}_\mathrm{y}$-plane on the left side. The boundaries of the AR interval are defined as a symmetric band around the modal value of the AR distribution. The modal value M_{AR} is given by the histogram bin with the highest frequency, and the band is given by $\left[M_{\mathrm{AR}} - \frac{1}{4}\sigma_{\mathrm{AR}},\ M_{\mathrm{AR}} + \frac{1}{4}\sigma_{\mathrm{AR}}\right]$, where σ_{AR} denotes the standard deviation of the AR distribution.

The validation procedure is described in Algorithm 5 in the appendix.

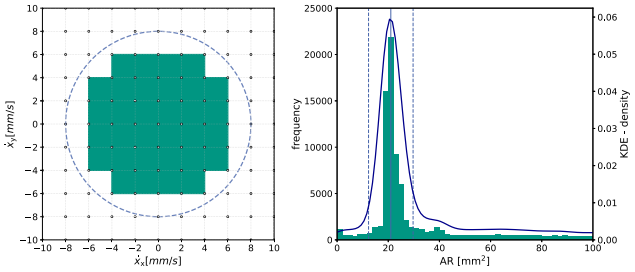

FIGURE 21. Illustration of D of $\mathrm{DPOM}_{\mathrm{IMP}t} \mid t \in (\mathbb{N} \cap [1, 12]) \setminus \{4\}$.

C. VALIDATION RESULTS

1) VALIDATION 1: PE PARAMETERIZATION

Figure 22 shows the results for a global PE and multiple local PEs for the axis currents I_x, I_y, I_z and I_sp. The proposed approach consistently yields lower RMSE. If R^2 is used to examine the variance described, the improvement becomes even more evident (see table 4). The increased prediction quality forms a necessary basis for the transfer of physical knowledge via PiDO.

TABLE 4. RMSE and R^2 for one global PE and the proposed approach.

Approach	Metric	x-axis	y-axis	z-axis	sp-axis
global PE	RMSE	0.9726	0.9286	1.0053	3.0344
	R^2	0.2077	0.1657	0.0718	0.4479
proposed	RMSE	**0.3029**	**0.2340**	**0.3378**	**1.7297**
	R^2	**0.9232**	**0.9470**	**0.8952**	**0.8206**

2) VALIDATION 2: ABLATION STUDY OF THE SAMPLING STRATEGY

Figure 23 (DMC) and 24 (CMX) show the standard deviations σ of the datasets in the target areas of the test parts. These show the extent to which it is possible to balance the data distribution. It can be seen that the standard deviation

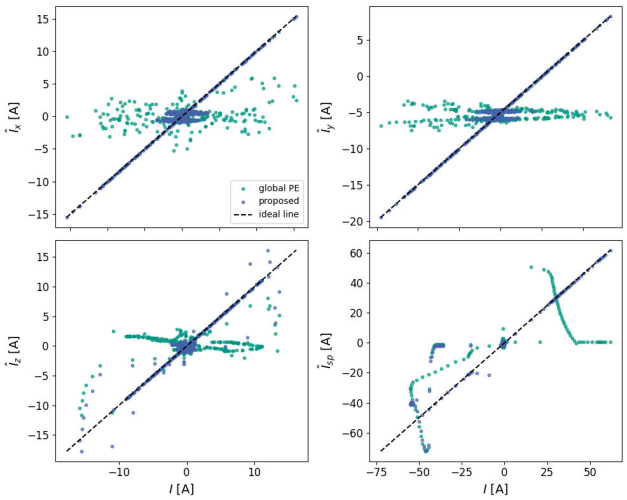

FIGURE 22. Comparison of a global and local PEs.

increases with increasing data volume. Overall, $D_{\mathrm{real+PiA+PiE}}$ results in lower σ than D_{real}. This effect can be seen on both machines.

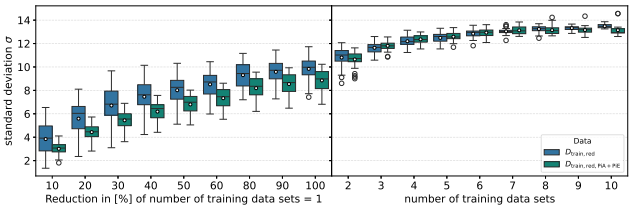

FIGURE 23. Standard deviation σ in $B_{ijk} \subset D_{\mathrm{test}}$ for DMC.

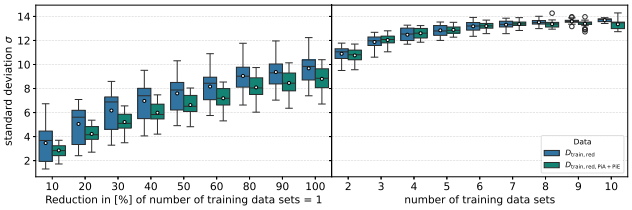

FIGURE 24. Standard deviation σ in $B_{ijk} \subset D_{\mathrm{test}}$ for CMX.

To investigate the overall benefit of PiDO, Table 5 shows the averaged RMSE values and the corresponding standard deviations σ for both machines. Scenario 1 and 2 are considered together within the entire feature space (C). The analysis reveals axis-specific differences and machine-dependent deviations in model quality. Overall, it can be seen that the lowest RMSE values for both machine types are achieved predominantly with $D_{\mathrm{real+PiA+PiE}}$. In addition, the standard deviation σ of the PiDO datasets is consistently lower.

Figure 25 provides a more concrete insight into the current signal, exemplarily showing the axis-specific current curves

TABLE 5. Axis-specific comparison for both machines.

axis	data	DMC		CMX	
		RMSE	Std	RMSE	Std
I_x	Real	0.26650	0.03795	0.12971	0.05017
	Real PiA	0.26662	0.03862	0.13000	0.04959
	Real PiE	**0.26428**	0.03699	0.12947	0.04849
	Real PiA PiE	0.26450	**0.03696**	**0.12946**	**0.04721**
I_y	Real	0.17373	0.05053	0.13636	0.06196
	Real PiA	0.17376	0.05069	0.13543	0.05934
	Real PiE	0.17119	0.04879	0.13523	0.05936
	Real PiA PiE	**0.17124**	**0.04870**	**0.13504**	**0.05783**
I_z	Real	0.05500	0.02300	0.12596	0.05564
	Real PiA	0.05488	0.02291	0.12578	0.05556
	Real PiE	0.05467	**0.02280**	0.12581	0.05532
	Real PiA PiE	**0.05450**	0.02287	**0.12532**	**0.05506**
I_{sp}	Real	0.37285	0.04237	0.02527	0.00701
	Real PiA	0.37265	0.04235	0.02522	0.00699
	Real PiE	0.37099	0.04135	0.02509	0.00691
	Real PiA PiE	**0.37091**	**0.04130**	**0.02505**	**0.00687**

for the first 500 data points for $D_{\text{train}} = \text{DPOM}_{\text{IMP06}}$ and $D_{\text{test}} = \text{DPOM}_{\text{IMP11}}$, which corresponds to $m = 1$ and $n = 1$ in Algorithm 5. Overall, it can be seen that the curves are of high quality and that all approaches deliver similar results. If the prediction is more complex, this is reflected in all approaches, as can be seen in the z-axis. Differences become apparent in the details.

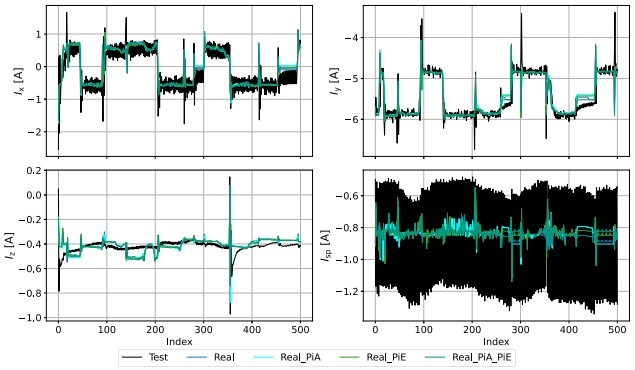

FIGURE 25. Current signal for the individual axes (DMC).

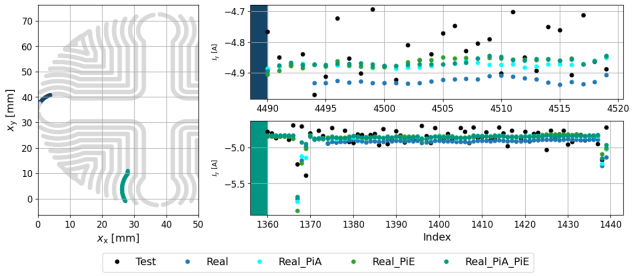

FIGURE 26. Exemplary sections of I_y for $D_{\text{test}} = \text{DPOM}_{\text{IMP11}}$.

Figure 26 shows excerpts from $D_{\text{test}} = \text{DPOM}_{\text{IMP11}}$ and the corresponding current curves I_y, illustrating the impact of PiDO. It demonstrates the positive impact of PiDO, which shifts the predicted signal towards the true values. The generated data points and their influence on the axes for two exemplary voxels on the DMC are shown in Figure 27. However, the effects of data points from other voxels on the prediction accuracy in these two voxels must be considered.

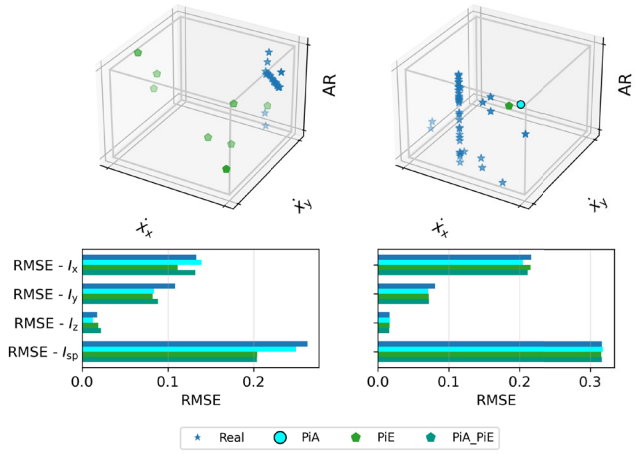

FIGURE 27. Exemplary DMC voxels and voxel-specific *RMSE* for $A_{\text{real},B_{ijk}} = 70\%$ on the left ($\dot{x}_x \in [4, 6)$, $\dot{x}_y \in [4, 6)$, AR $\in [18, 20)$) and $A_{\text{real},B_{ijk}} = 95\%$ on the right ($\dot{x}_x \in [-2, 0)$, $\dot{x}_y \in [-8, -6)$, AR $\in [20, 22)$).

Figure 28 (DMC) and 29 (CMX) show the mean ΔRMSE_{PiDO} for $\Delta\text{RMSE}_{\text{D}_{\text{real+PiA}}}$, $\Delta\text{RMSE}_{\text{D}_{\text{real+PiE}}}$ and $\Delta\text{RMSED}_{\text{real+PiA+PiE}}$. Scenario 1 (reduction of the training file in 10 % steps) is shown on the left, while scenario 2 (gradual addition of further training files) is shown on the right. The subspace D is shown at the top and the entire feature space C at the bottom. It can be seen that a greater improvement is achieved within the areas D than within C. Significant improvement occurs more frequently on the DMC than on the CMX. An improvement can also be achieved in the entire feature space. This effect decreases as the amount of training data increases.

Figures 30 (DMC) and 31 (CMX) show the impact of $\Delta\text{RMSE}_{\text{D}_{\text{real+PiA}}}$ on an axis-specific level. For the DMC $i \in \{\text{x}, \text{y}, \text{z}, \text{sp}\}$ and for the CMX $i \in \{\text{x}, \text{z}\}$ are shown. It can be seen that the impact of PiDO on the DMC is driven by all axes except the z axis, resulting in a similar trend if beneficial. Thus, a stronger impact is achieved with a reduced amount of training data, which continuously decreases. In contrast, the added value of PiDO for the z axis of the CMX increases with a larger dataset, resulting in significant improvements.

D. DISCUSSION OF THE VALIDATION RESULTS

The validation results are discussed with respect to three aspects: The quality of the local PE parameterization, the ability of PiDO to improve dataset balance, and the resulting effect on prediction accuracy and robustness. This structure enables the validation to be interpreted in terms not only of model error, but also of how the generated samples alter the distribution of the training data.

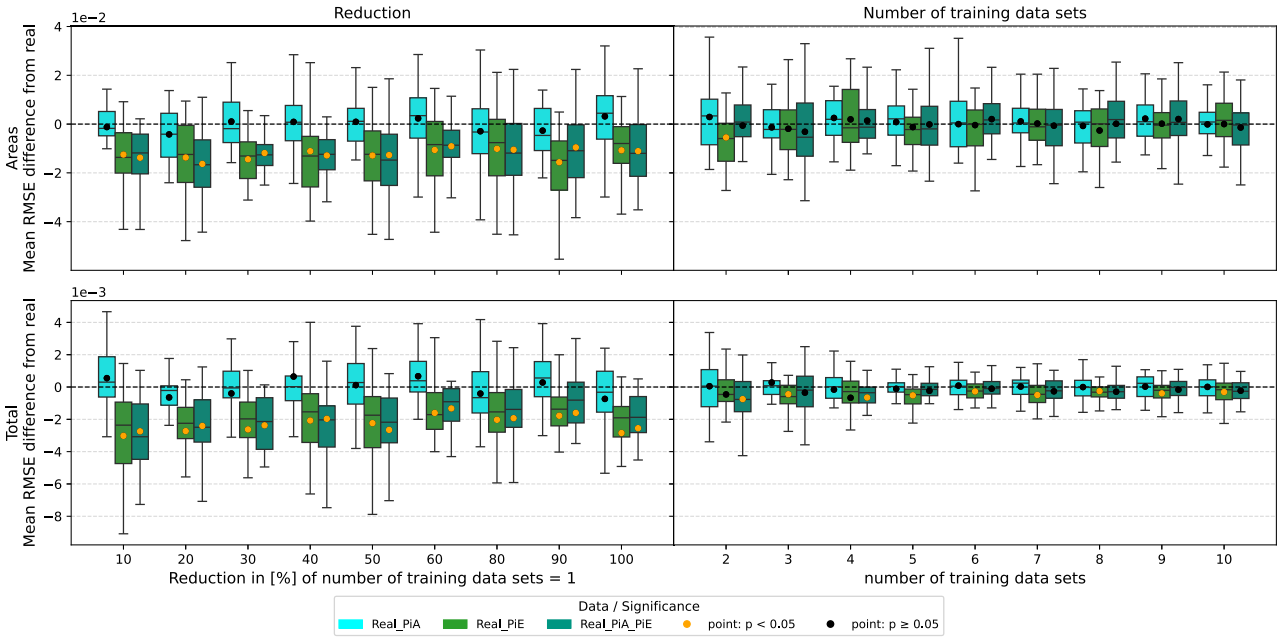

FIGURE 28. Averaged RMSE differences to D_{real} over reduction and number of training files (DMC).

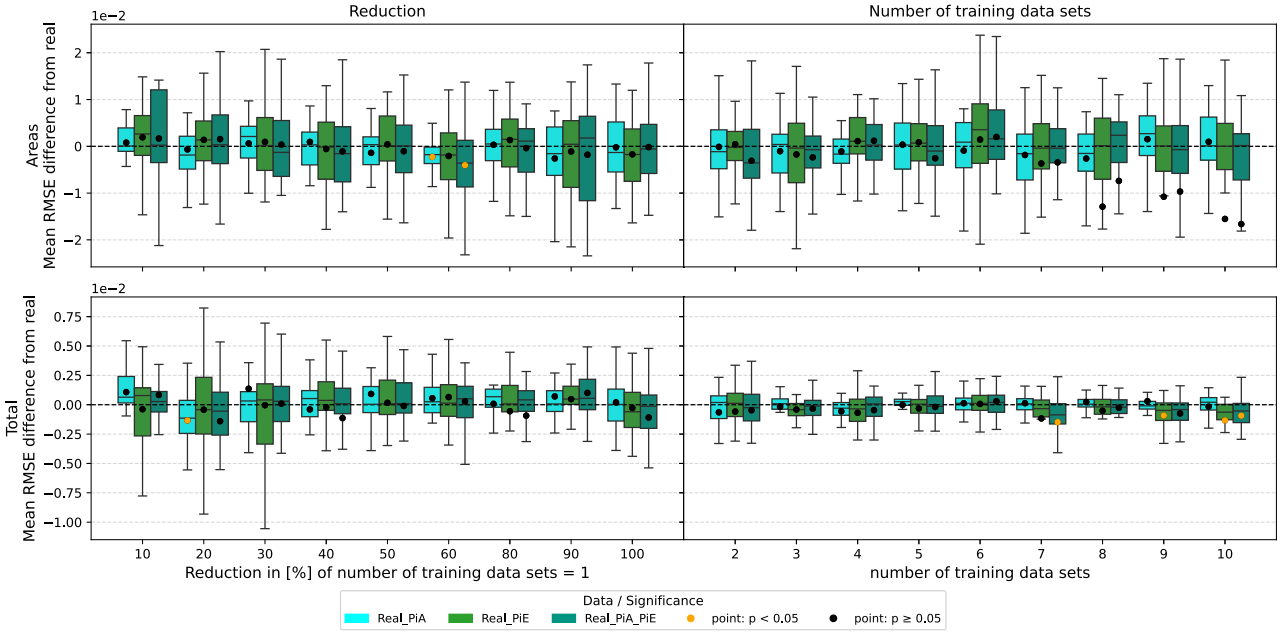

FIGURE 29. Averaged RMSE differences to D_{real} over reduction and number of training files (CMX).

The results show that the proposed approach for PE parameterization consistently achieves lower RMSE and higher R^2 across all axes. This improvement is due to local parameterization. A separate PE is determined for each range B_{ijk} and used exclusively for that range. This reduces the influence of global effects and individual outliers on the parameterization.

The balancing analysis in figure 23 and 24 shows a small initial standard deviation σ, which increases as the amount of data increases. At first, there are few data points in many areas, resulting in a low σ. When the total amount of data increases, there will still be areas with comparatively few data points, resulting in an increased standard deviation as existing heterogeneities become more apparent. The results show that the proposed sampling strategy achieves a more balanced dataset during this transition. The addition of PiDO data homogenizes the data distribution, resulting in a lower standard deviation for $B_{ijk} \subset D_{\text{test}}$. σ is not influenced by

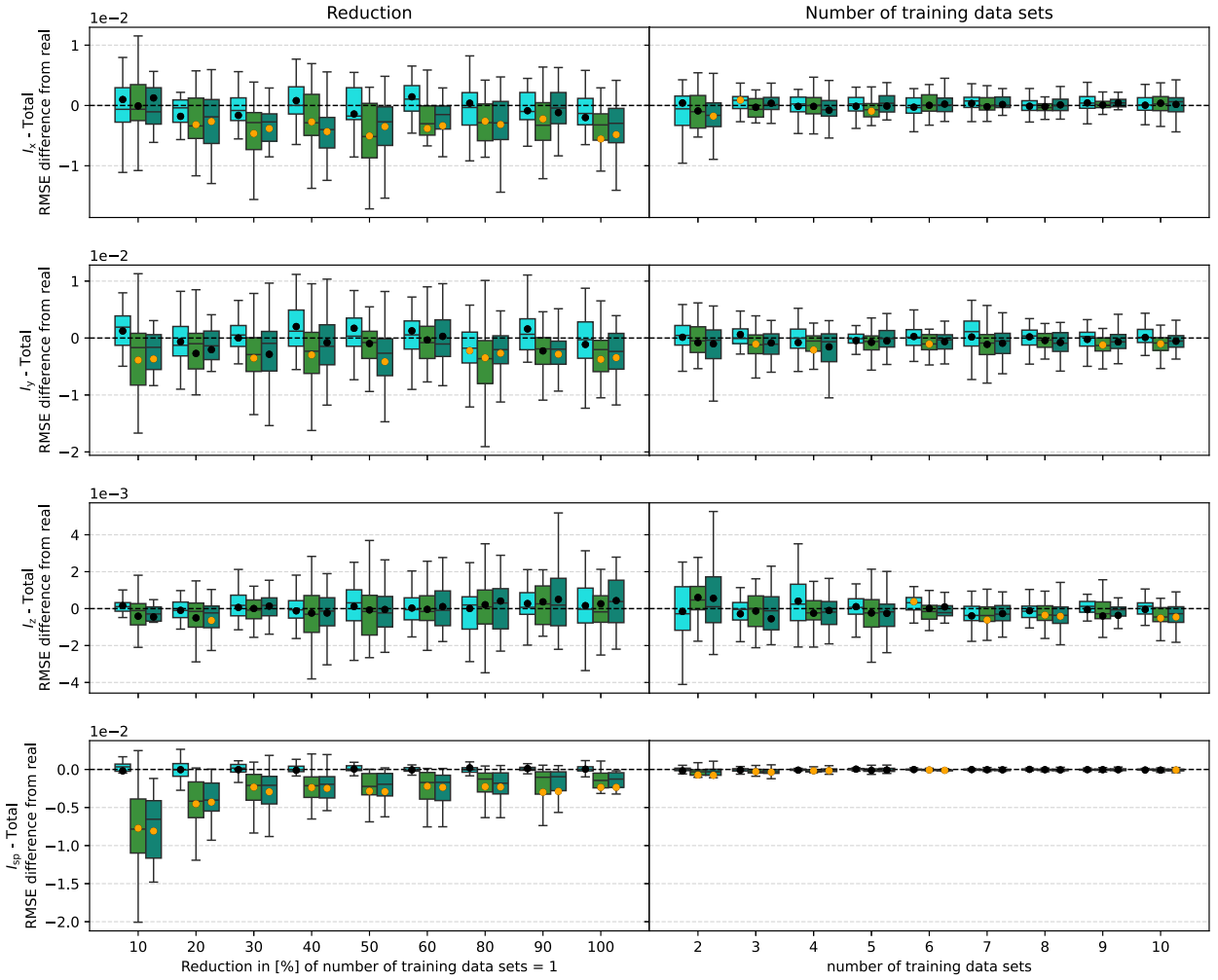

FIGURE 30. Axis-specific RMSE differences to D_{real} over reduction and number of training files (DMC).

PiDO from sequence 3 to 7. This is due to areas in which little real data is added. According to the sampling strategy, only a few PiA and PiE data points are generated, resulting in limited effects. From sequence 9 onward, the approach shows a reduced σ, indicating that the threshold for generating a relevant amount of data has been reached. It can be assumed that this trend will continue. Overall, it can be demonstrated that the sampling strategy leads to a more balanced dataset.

The RMSE values shown in Table 5 indicate that the addition of PiDO data (PiE and the combination of real, PiA, and PiE) consistently leads to lower RMSE on both machines, while PiA alone already results in a slight improvement in some cases. This illustrates that synthesized and augmented data integrate explicit physical knowledge into the dataset, which can improve a model's prediction quality. This effect can be attributed to more accurate predictions for specific time periods. As shown in figure 26, the added PiDO data is closer to the actual signal than using real data alone. At the same time, the models become more robust, as reflected by the consistently lower standard deviation.

Investigations of the benefit of the supplementary PiA and PiE data in figure 28 and 29 show that PiE in particular has a significant positive effect. This effect intensifies with a reduced amount of real data, such as when a new series with significant changes is manufactured. This is supported by the fact that the largest improvement is achieved with PiE and PiA+PiE in D on the DMC. The effect decreases as the amount of measured data increases since less PiDO data is generated based on the sampling functions. Furthermore, the influence of PiDO data decreases because measured data contains higher-quality information. PiDO does not result in a decline in model performance. Therefore, it can be concluded that the ML model is still capable of identifying the relationships within the measured data and that there is no significant bias towards the PEs. Overall, PiA has no significant influence due to its limited impact.

The higher impact of PiDO on the DMC can be explained by the machine kinematics and the associated complexity of the resulting axis-specific signals. Furthermore, as the DMC dates from 1997, its axes are worn, resulting in an additional

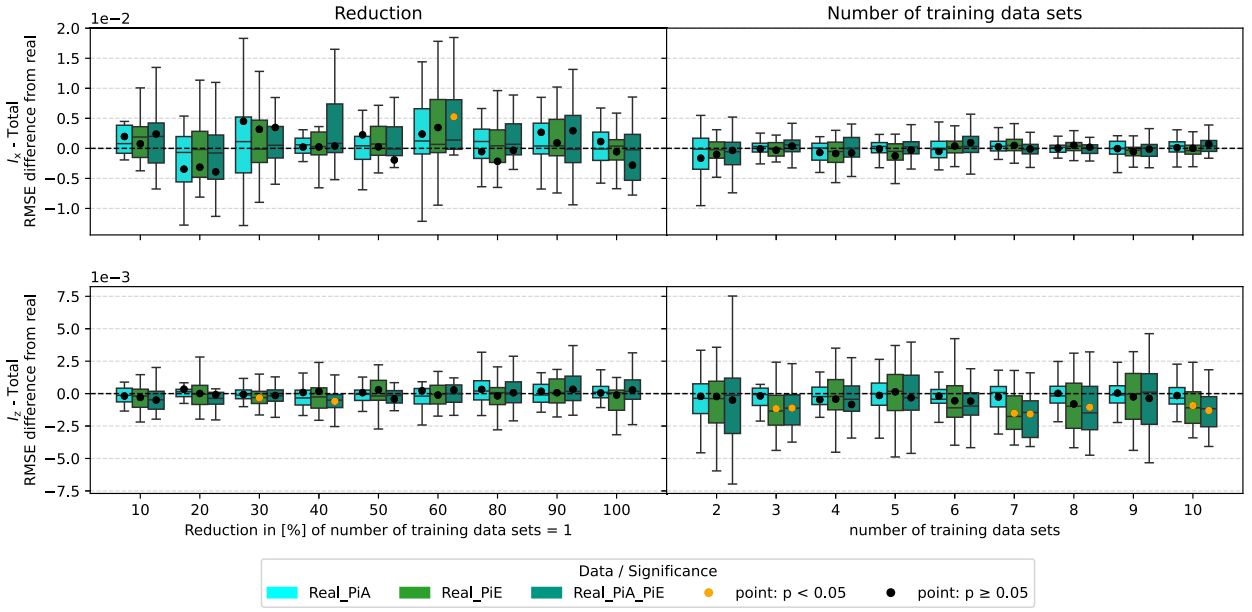

FIGURE 31. Axis-specific RMSE differences to D_{real} over reduction and number of training files for the x and z axis (CMX).

layer of complexity. Therefore, it can be concluded that physical knowledge from PiDO is particularly advantageous for predicting accuracy when the interaction to be mapped is influenced by many superimposed effects. This is further confirmed by the slight improvements achieved on the z axis of the CMX, the most complex due to its kinematic chain.

An axis-specific evaluation shows that axes influenced by complex superimposed effects in particular benefit from the additional physical knowledge provided by PiDO. This is especially evident for the x, y, and sp axes of the DMC and, to a limited extent, for the z axis of the CMX. If, on the other hand, the axes are not influenced by complex effects, the data's information content is sufficient with fewer samples. This is the case for the z axis of the DMC. It stands alone, does not form a kinematic chain with other axes, and has a high weight due to the workpiece it carries. For the CMX, a reduced complexity is the case for all other axes.

Overall, the validation results suggest that PiDO is most effective in prediction tasks involving complex axis interactions and sparse datasets. Its influence decreases when sufficient real data covers the relevant operating regimes. Figure 32 provides a graphical summary, which compares the number of improvements for both machines. Scenario 1 and 2 are combined for clarity. The total amount of improvements (< 0) and deteriorations (> 0) is counted. Significant effects are highlighted as darker colored bars. Significant improvements predominate on the DMC, particularly for PiE and PiA+PiE, with frequent significant increases and no significant deterioration. There is also a tendency towards improvement on the CMX, although these improvements are rarely significant. When the impact is compared across all axes, a clear picture of improvement emerges. It can be summarized that model quality can be significantly improved by PiDO and its use can be considered advantageous.

E. LIMITATIONS AND PRACTICAL CONSIDERATIONS

Although the validation results suggest that PiDO could improve both predictive performance and dataset balance, several limitations must be considered.

With respect to data availability, it should be noted that improvements can only be achieved to a certain extent. In particular, the application of PiDO depends on the availability of a minimum amount of real data within the relevant region of the reduced feature space. If very little data is available, it may no longer be possible to perform a sufficient PE fit. This limitation can be exacerbated by drastic domain changes where operating conditions differ significantly from previously observed states. If a substantial amount of data exists for a given state but a new product is produced under completely different conditions, the amount of data generated and the impact of the approach will be limited (see figure 19).

In addition to limitations related to data availability, the use of simplified PEs introduces an inherent model-related limitation. The PE may introduce bias towards an abstract system description that cannot fully represent all process- or machine-specific effects. However, this bias is explicitly bounded by the sampling strategy, particularly for $A_{\text{real}} = 100\%$, where limited PiDO data is generated. Closely related is the inherent limitation of synthetic data itself. The quality of PiDO data is lower than of real data, as it approximates

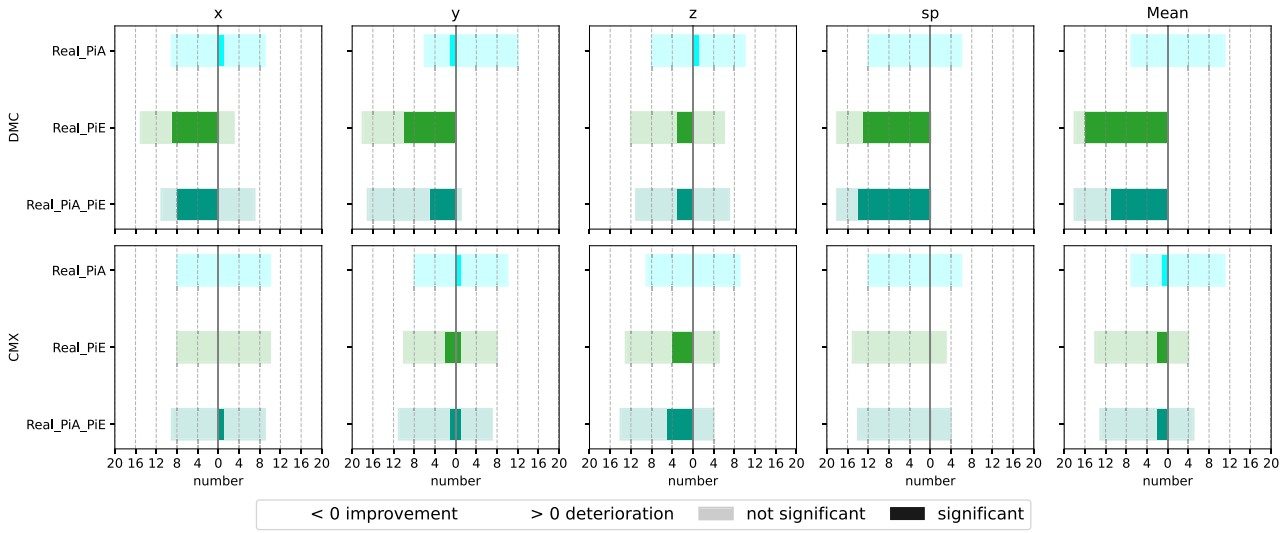

FIGURE 32. Summary of improvements and deteriorations for the DMC and the CMX during validation.

the underlying process based on simplified physical models. Nevertheless, it can improve performance when used as a complement to real data rather than as a replacement.

Further limitations arise from the choice of general sampling functions. As these are defined independently of the axis configuration and the dataset, they are subject to superimposed effects. These cannot be resolved within the current formulation, which reduces the fit of the sampling functions (see table 3). The sampling quality can be improved by introducing axis-specific or shift-dependent functions. However, this would reduce the approach's general applicability and increase the effort required for a transfer to other machines. Furthermore, the discretization affects the quality of the generated PiDO data. Reducing the bin size is expected to improve quality, but results in increased computational effort.

With regard to transferability and practical deployment, the general structure of the reduced feature space enables the PE parameterization and the sampling functions to be transferred to other three-axis milling machines. This was demonstrated by the validation on two machines with different axis configurations. When transferred to other machines and processes, the general methodology can be applied, but the parameters and functions must be appropriately adapted to the new conditions.

From a computational standpoint, parameterizing multiple local PEs is challenging and must be considered when selecting the bin size. Larger bins reduce the number of PEs that must be parametrized, whereas smaller bin sizes improve modeling accuracy at a higher computational cost. The computational cost of PiE and PiA is proportional to the number of data points generated. This must be taken into account in the model training strategy. Since PiDO only interferes with the training process, the model can be used for real-time monitoring once it has been trained.

VIII. SUMMARY AND OUTLOOK

This study proposed and validated a PiDO data-generation framework for physics-informed dataset optimization in ML-based CNC process monitoring. The framework used a physical equation to describe the axis-specific current dynamics in order to generate targeted, physics-consistent samples within a reduced feature space. By locally parameterizing the physical equation through Gaussian weighting and introducing two complementary data-generation modes (PiA and PiE), the approach was able to enrich incrementally collected training datasets in a domain-specific way. Furthermore, a transferable sampling strategy was developed to determine the proportions of real, PiA and PiE samples required for a given target density. Validation on multi-part datasets across different machines demonstrated that PiDO enhanced the balance of datasets and improved the predictive performance and robustness of ML models.

Future work should focus on several actionable directions. First, adaptive strategies for selecting bin sizes, target densities, and PiA/PiE ratios should be investigated in order to balance prediction quality, dataset balance, and computational effort. In this context, the use of the entire available training dataset should be evaluated more systematically, for example in combination with data-weighting strategies similar to the PE-parameterization scheme. Additionally, the impact of discrepancies between training and inference data should be analyzed in greater detail. Second, PiDO should be systematically compared with established data-centric and learning-based approaches for handling data imbalance and distribution shifts, such as SMOTE, time-series augmentation, domain adaptation, and cost-sensitive learning. Third, the interaction between data-centric optimization and model-centric IL should be analyzed to determine how PiDO can be combined with other IL methods. Fourth, the approach should be transferred to additional machines and evaluated with respect to cross-machine scalability

under varying industrial conditions, including temporal drift, tool-wear progression, different materials, and additional machining operations. Finally, to better evaluate the practical benefits of PiDO in industrial applications, future studies should incorporate additional application-oriented metrics, such as mean absolute error, peak-load prediction error, transient accuracy, and robustness under load changes. The proposed methodology could also be generalized beyond current prediction to other physical signals, manufacturing processes, and data-scarce industrial domains.

Overall, the results confirm that physics-informed samples are most beneficial in data-sparse and extrapolative regions of the feature space, as they can increase model robustness and predictive accuracy in these areas. The proposed framework complements existing model-centric approaches by providing a practical method for incorporating physically consistent domain knowledge into manufacturing datasets. Therefore, it offers a generalizable strategy for improving learning in conditions involving sparse, imbalanced or drifting datasets.

APPENDIX

This section presents the PE, the Wasserstein distances of the selected datasets, and several algorithms.

For the PE, the terms

$$\begin{aligned}
X_1 &= \ddot{x}_\mathrm{x} \cdot \dot{x}_\mathrm{x} \cdot \mathrm{sgn}(\dot{x}_\mathrm{x}) \\
X_2 &= (\dot{x}_\mathrm{x})^2 \cdot \mathrm{sgn}(\dot{x}_\mathrm{x}) \\
X_3 &= F_\mathrm{x} \cdot \mathrm{MRR} \\
X_4 &= \mathrm{sgn}(\dot{x}_\mathrm{x}) \\
X_5 &= 1 \\
X_6 &= \ddot{x}_\mathrm{y} \cdot \dot{x}_\mathrm{y} \cdot \mathrm{sgn}(\dot{x}_\mathrm{y}) \\
X_7 &= (\dot{x}_\mathrm{y})^2 \cdot \mathrm{sgn}(\dot{x}_\mathrm{y}) \\
X_8 &= F_\mathrm{y} \cdot \mathrm{MRR} \\
X_9 &= \mathrm{sgn}(\dot{x}_\mathrm{y}) \\
X_{10} &= \ddot{x}_\mathrm{z} \cdot \dot{x}_\mathrm{z} \cdot \mathrm{sgn}(\dot{x}_\mathrm{z}) \\
X_{11} &= (\dot{x}_\mathrm{z})^2 \cdot \mathrm{sgn}(\dot{x}_\mathrm{z}) \\
X_{12} &= F_\mathrm{z} \cdot \mathrm{MRR} \\
X_{13} &= \mathrm{sgn}(\dot{x}_\mathrm{z}) \\
X_{14} &= \ddot{\varphi}_\mathrm{sp} \cdot \dot{\varphi}_\mathrm{sp} \cdot \mathrm{sgn}(\dot{\varphi}_\mathrm{sp}) \\
X_{15} &= (\dot{\varphi}_\mathrm{sp})^2 \cdot \mathrm{sgn}(\dot{\varphi}_\mathrm{sp}) \\
X_{16} &= M_\mathrm{sp} \cdot \mathrm{MRR} \\
X_{17} &= \mathrm{sgn}(\dot{\varphi}_\mathrm{sp})
\end{aligned} \tag{17}$$

for $i \in \{\mathrm{x}, \mathrm{y}, \mathrm{z}, \mathrm{sp}\}$ with

- x_i: Position
- $\dot{x}_i$, $\dot{\varphi}_\mathrm{sp}$: Axis specific velocity
- $\ddot{x}_i$, $\ddot{\varphi}_\mathrm{sp}$: Axis specific acceleration
- $\mathrm{sgn}(\cdot)$: Sign function used to model the current direction
- F_i: Axis specific process force
- MRR: Material removal rate
- M_sp: Spindle torque

are given.

Algorithm 3 outlines the schematic procedure of the development experiment designed to investigate the influence of PiA and PiE data on the model performance.

Algorithm 3 Varying Proportions of D_real, D_PiA and D_PiE

for $m = 1$ **to** 30 **do**
  D_next, D_train
  $D_\mathrm{next,train} \leftarrow 0.7 \cdot D_\mathrm{next}$
  $D_\mathrm{next,test} \leftarrow D_\mathrm{next} \setminus D_\mathrm{next,train}$

  for all $B_{ijk} \in \mathbb{M}$ **do**
    $D_\mathrm{train} \leftarrow D_\mathrm{train}$ reduce to upper limit
    $D_\mathrm{next,train} \leftarrow D_\mathrm{next,train}$ reduce to upper limit
  end for

  proportion of next,train $\leftarrow \{0\%, 10\%, \ldots, 100\%\}$
  for next proportion (A_real) **in** proportion of next,train data **do**
    $D_\mathrm{next,red} \leftarrow D_\mathrm{next,train} \cdot A_\mathrm{real}$
    $D_\mathrm{real} \leftarrow (D_\mathrm{train} \wedge D_\mathrm{next,red})$
    for all $B_{ijk} \in \mathbb{M}$ **do**
      if $B_{ijk} \cap D_\mathrm{real}$ **then**
        $\hat{c}_{B_{ijk}} \leftarrow (D_\mathrm{real}) \cdot w_{B_{ijk}}$ with α, β and σ
      end if
    end for

    PiA proportion $\leftarrow \{0\%, \ldots, 25\%\}$
    PiE proportion $\leftarrow \{0\%, \ldots, 110\%\}$
    for A_PiA **in** proportion PiA data **do**
      for A_PiE **in** proportion PiE data **do**
        if $(A_\mathrm{real} + A_\mathrm{PiA} + A_\mathrm{PiE}) \geq 115\%$ **or** $(A_\mathrm{PiA} > A_\mathrm{real})$ **then**
          Skip loop iteration
        end if

        Number of PiA (A_PiA) $\leftarrow$ #target $\cdot A_\mathrm{PiA}$
        Number of PiE (A_PiE) $\leftarrow$ #target $\cdot A_\mathrm{PiE}$
        $D_\mathrm{PiA} \leftarrow \hat{c}_{B_{ijk}} \cdot \mathrm{PE}_{B_{ijk}}$
        $D_\mathrm{PiE} \leftarrow \hat{c}_{B_{ijk}} \cdot \mathrm{PE}_{B_{ijk}}$
        $D_\mathrm{Modell} \leftarrow D_\mathrm{real} \wedge D_\mathrm{PiA} \wedge D_\mathrm{PiE}$

        currents$\leftarrow \{\mathrm{x}, \mathrm{y}, \mathrm{z}, \mathrm{sp}\}$
        for current (i) **in** currents **do**
          $\hat{I}_i \leftarrow$ **ML** $-$ **Modell**(D_Modell)
          return $\mathrm{RMSE}(I_i, \hat{I}_i)$
        end for

      end for
    end for
  end for
end for

Figure 33 shows the Wasserstein distances of the individual training datasets to the test dataset $\mathrm{DALG}_\mathrm{SF1}$.

Algorithm 4 describes validation scenario 1 and 2.

In Algorithm 5, the general structure of the validation of the sampling strategy is described.

Algorithm 4 Validation Scenario 1 and 2

for $m = 1$ **to** 30 **do**

  Variant 1:
  $n = 1$
  Reduction rates $\leftarrow \{10\%, 20\%, \ldots, 100\%\}$
  for n_{red} **in** Reduction rates **do**
    $D_{\text{test,m,n=1}}$, $D_{\text{training,m,n=1}}$
    $D_{\text{training,m,n=1,red}} \leftarrow D_{\text{training,m,n=1}} \cdot n_{\text{red}}$
    $D_{\text{train,m,n=1,target}} \leftarrow D_{\text{training,m,n=1,red}} \leq \rho_{target}$ {# Data reduction to target density}
  end for

  Variant 2:
  for $n = 1$ **to** 10 **do**
    $D_{\text{test,m,n}}$, $D_{\text{training,m,n}}$
    $D_{\text{train,m,n,target}} \leftarrow D_{\text{training,m,n}} \leq \rho_{target}$ {# Data reduction to target density}
  end for

end for

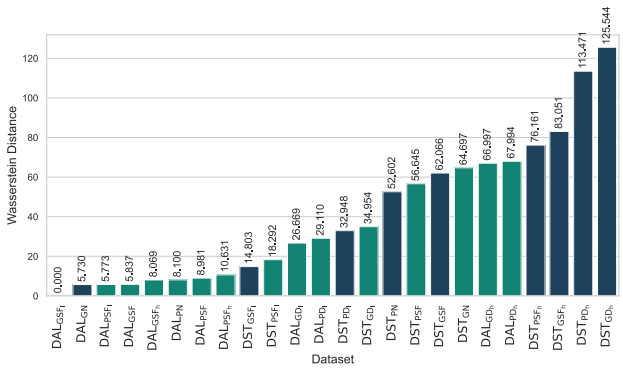

FIGURE 33. Waserstein distances of a given D_{train} to $DALG_{SF1}$.

Algorithm 5 Validation Structure

for $m = 1$ **to** 30 **do**
  for $n = 1$ **to** 10 **do**
    Variant 1: $D_{\text{train,m,n=1,target}}$, $D_{\text{test,m,n=1}}$
    Variant 2: $D_{\text{train,m,n,target}}$, $D_{\text{test,m,n}}$

    for alle $B_{ijk} \in \mathbb{M}$ **do**
      if $B_{ijk} \cap D_{\text{test,m,n}}$ **then**
        $\text{A}_{\text{Rea,B}_{\text{ijk}}} = D_{\text{train,m,n,target}} / \rho_{target}$ {# Proportion of real data in regions of the test set}
        $D_{\text{real,B}_{\text{ijk}}} \leftarrow D_{\text{train,m,n,target}} \in B_{ijk}$ {#Training data in the region of the test workpiece}

        Computation of the proportion of PiA and PiE samples using the sampling strategy:
        $A_{\text{PiA}} \leftarrow f_{PiA}(\text{A}_{\text{Rea,B}_{\text{ijk}}})$ {#proportion of PiA samples}
        $A_{\text{PiE}} \leftarrow f_{PiE}(\text{A}_{\text{Rea,B}_{\text{ijk}}})$ {#proportion of PiE samples}
        if A_{PiA} **or** $A_{\text{PiE}} > 0$ **then**
          $\hat{c}_{B_{ijk}} \leftarrow D_{\text{train,m,n,target}} \cdot w_{B_{ijk}}$ with α, β and σ
          $D_{\text{PiA}} \leftarrow \hat{c}_{B_{ijk}} \cdot \text{PE}_{B_{ijk}}$
          $D_{\text{PiE}} \leftarrow \hat{c}_{B_{ijk}} \cdot \text{PE}_{B_{ijk}}$
        end if
      end if

      Investigation of different data compositions:
      $\text{D}_{\text{real}} \leftarrow D_{\text{train,m,n,target}}$
      $\text{D}_{\text{real+PiA}} \leftarrow D_{\text{train,m,n,target}} \wedge D_{\text{PiA}}$
      $\text{D}_{\text{real+PiE}} \leftarrow D_{\text{train,m,n,target}} \wedge D_{\text{PiE}}$
      $\text{D}_{\text{real+PiA+PiE}} \leftarrow D_{\text{train,m,n,target}} \wedge D_{\text{PiA}} \wedge D_{\text{PiE}}$

      for D_{Modell} in $\{\text{D}_{\text{real}}, \text{D}_{\text{real+PiA}}, \text{D}_{\text{real+PiE}}, \text{D}_{\text{real+PiA+PiE}}\}$ **do**
        currents $\leftarrow \{\text{x, y, z, sp}\}$
        for current (i) **in** currents **do**
          $\hat{I}_i \leftarrow \textbf{ML} - \textbf{Modell}(D_{\text{Modell}})$
          return min $\text{RMSE}(I_i, \hat{I}_i)$
        end for

      end for
    end for
  end for
end for

ACKNOWLEDGMENT

The authors acknowledge the use of DeepL Write for spelling and grammar enhancement.

REFERENCES

[1] M. Bortolini, F. G. Galizia, and C. Mora, "Reconfigurable manufacturing systems: Literature review and research trend," *J. Manuf. Syst.*, vol. 49, pp. 93–106, Oct. 2018.

[2] R. Ströbel, S. Deucker, H. Zhou, H. Kader, A. Puchta, B. Noack, and J. Fleischer, "Hybrid machine learning for CNC process monitoring," *IEEE Access*, vol. 13, pp. 91875–91888, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/11015456/

[3] B. Denkena, M.-A. Dittrich, H. Noske, and M. Witt, "Statistical approaches for semi-supervised anomaly detection in machining," *Prod. Eng.*, vol. 14, no. 3, pp. 385–393, Jun. 2020.

[4] X. Zhang, Y. Gao, Z. Guo, W. Zhang, J. Yin, and W. Zhao, "Physical model-based tool wear and breakage monitoring in milling process," *Mech. Syst. Signal Process.*, vol. 184, Feb. 2023, Art. no. 109641.

[5] C. Legaard, T. Schranz, G. Schweiger, J. Drgoňa, B. Falay, C. Gomes, A. Iosifidis, M. Abkar, and P. Larsen, "Constructing neural network based models for simulating dynamical systems," *ACM Comput. Surv.*, vol. 55, no. 11, pp. 1–34, Nov. 2023.

[6] R. X. Gao, J. Krüger, M. Merklein, H.-C. Möhring, and J. Váncza, "Artificial intelligence in manufacturing: State of the art, perspectives, and future directions," *CIRP Ann.*, vol. 73, no. 2, pp. 723–749, 2024.

[7] G. M. van de Ven, T. Tuytelaars, and A. S. Tolias, "Three types of incremental learning," *Nature Mach. Intell.*, vol. 4, no. 12, pp. 1185–1197, Dec. 2022. [Online]. Available: https://www.nature.com/articles/s42256-022-00568-3

[8] J. Jakubik, M. Vössing, N. Kühl, J. Walk, and G. Satzger, "Data-centric artificial intelligence," *Bus. Inf. Syst. Eng.*, vol. 66, no. 4, pp. 507–515, Aug. 2024, doi: 10.1007/s12599-024-00857-8.

[9] N. Bhatt, N. Bhatt, P. Prajapati, V. Sorathiya, S. Alshathri, and W. El-Shafai, "A data-centric approach to improve performance of deep learning models," *Sci. Rep.*, vol. 14, no. 1, p. 22329, Sep. 2024.

[10] L. Camacho, G. Douzas, and F. Bacao, "Geometric SMOTE for regression," *Expert Syst. Appl.*, vol. 193, May 2022, Art. no. 116387. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095741742101678X

[11] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Rev. Phys.*, vol. 3, no. 6, pp. 422–440, Jun. 2021. [Online]. Available: https://www.nature.com/articles/s42254-021-00314-5

[12] G. M. van de Ven, N. Soures, and D. Kudithipudi, "Continual learning and catastrophic forgetting," 2024, *arXiv:2403.05175*.

[13] R. Hyder, K. Shao, B. Hou, P. Markopoulos, A. Prater-Bennette, and M. S. Asif, "Incremental task learning with incremental rank updates," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 566–582.

[14] J. Yan, C. Li, Y. Liu, D. Yu, and Z. Jia, "Incremental model evolution for power system security early warning based on knowledge distillation and active learning," *IEEE Trans. Ind. Informat.*, vol. 20, no. 11, pp. 12958–12968, Nov. 2024.

[15] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "ICaRL: Incremental classifier and representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5533–5542.

[16] N. Polyzotis and M. Zaharia, "What can data-centric AI learn from data and ML engineering?" 2021, *arXiv:2112.06439*.

[17] S. E. Whang, Y. Roh, H. Song, and J.-G. Lee, "Data collection and quality challenges in deep learning: A data-centric AI perspective," *VLDB J.*, vol. 32, no. 4, pp. 791–813, Jul. 2023.

[18] R. Ströbel, M. Mau, A. Puchta, and J. Fleischer, "Improving time series regression model accuracy via systematic training dataset augmentation and sampling," *Mach. Learn. Knowl. Extraction*, vol. 6, no. 2, pp. 1072–1086, May 2024.

[19] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: Overview study and experimental results," in *Proc. 11th Int. Conf. Inf. Commun. Syst. (ICICS)*, Apr. 2020, pp. 243–248.

[20] C. Meng, S. Griesemer, D. Cao, S. Seo, and Y. Liu, "When physics meets machine learning: A survey of physics-informed machine learning," *Mach. Learn. Comput. Sci. Eng.*, vol. 1, no. 1, pp. 1–23, Jun. 2025.

[21] Y. Wu, B. Sicard, and S. A. Gadsden, "Physics-informed machine learning: A comprehensive review on applications in anomaly detection and condition monitoring," *Expert Syst. Appl.*, vol. 255, Dec. 2024, Art. no. 124678.

[22] P. Sharma, W. T. Chung, B. Akoush, and M. Ihme, "A review of physics-informed machine learning in fluid mechanics," *Energies*, vol. 16, no. 5, p. 2343, Feb. 2023.

[23] Y. Xu, S. Kohtz, J. Boakye, P. Gardoni, and P. Wang, "Physics-informed machine learning for reliability and systems safety applications: State of the art and challenges," *Rel. Eng. Syst. Saf.*, vol. 230, Feb. 2023, Art. no. 108900.

[24] J. Pateras, P. Rana, and P. Ghosh, "A taxonomic survey of physics-informed machine learning," *Appl. Sci.*, vol. 13, no. 12, p. 6892, Jun. 2023.

[25] Z. Ma, H. Liao, J. Gao, S. Nie, and Y. Geng, "Physics-informed machine learning for degradation modeling of an electro-hydrostatic actuator system," *Rel. Eng. Syst. Saf.*, vol. 229, Jan. 2023, Art. no. 108898.

[26] D. H. Green, A. W. Langham, R. A. Agustin, D. W. Quinn, and S. B. Leeb, "Physics-informed feature space evaluation for diagnostic power monitoring," *IEEE Trans. Ind. Informat.*, vol. 19, no. 3, pp. 2363–2373, Mar. 2023.

[27] P. Maheshwari, W. Wang, S. Triest, M. Sivaprakasam, S. Aich, J. G. Rogers, J. M. Gregory, and S. Scherer, "PIAug—Physics informed augmentation for learning vehicle dynamics for off-road navigation," 2023, *arXiv:2311.00815*.

[28] L. Schröder, N. K. Dimitrov, D. R. Verelst, and J. A. Sørensen, "Using transfer learning to build physics-informed machine learning models for improved wind farm monitoring," *Energies*, vol. 15, no. 2, p. 558, Jan. 2022.

[29] S. Teng, X. Chen, G. Chen, and L. Cheng, "Structural damage detection based on transfer learning strategy using digital twins of bridges," *Mech. Syst. Signal Process.*, vol. 191, May 2023, Art. no. 110160.

[30] M. Arias Chao, C. Kulkarni, K. Goebel, and O. Fink, "Fusing physics-based and deep learning models for prognostics," *Rel. Eng. Syst. Saf.*, vol. 217, Jan. 2022, Art. no. 107961.

[31] J. Shen, J. Chowdhury, S. Banerjee, and G. Terejanu, "Machine fault classification using Hamiltonian neural networks," 2023, *arXiv:2301.02243*.

[32] A. K. Kumar, S. Jain, S. Jain, M. Ritam, Y. Xia, and R. Chandra, "Physics-informed neural entangled-ladder network for inhalation impedance of the respiratory system," *Comput. Methods Programs Biomed.*, vol. 231, Apr. 2023, Art. no. 107421.

[33] S. Cai, Z. Wang, S. Wang, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks for heat transfer problems," *J. Heat Transf.*, vol. 143, no. 6, Jun. 2021, Art. no. 060801.

[34] S. Shen, H. Lu, M. Sadoughi, C. Hu, V. Nemani, A. Thelen, K. Webster, M. Darr, J. Sidon, and S. Kenny, "A physics-informed deep learning approach for bearing fault detection," *Eng. Appl. Artif. Intell.*, vol. 103, Aug. 2021, Art. no. 104295.

[35] D. Chen, Y. Li, K. Liu, and Y. Li, "A physics-informed neural network approach to fatigue life prediction using small quantity of samples," *Int. J. Fatigue*, vol. 166, Jan. 2023, Art. no. 107270.

[36] W. Deng, K. T. P. Nguyen, C. Gogu, J. Morio, and K. Medjaher, "Physics-informed lightweight temporal convolution networks for fault prognostics associated to bearing stiffness degradation," *PHM Soc. Eur. Conf.*, vol. 7, no. 1, pp. 118–125, Jun. 2022.

[37] F. A. C. Viana, R. G. Nascimento, A. Dourado, and Y. A. Yucesan, "Estimating model inadequacy in ordinary differential equations with physics-informed neural networks," *Comput. Struct.*, vol. 245, Mar. 2021, Art. no. 106458.

[38] T. X. Nghiem, J. Drgona, C. Jones, Z. Nagy, R. Schwan, B. Dey, A. Chakrabarty, S. Di Cairano, J. A. Paulson, A. Carron, M. N. Zeilinger, W. Shaw Cortez, and D. L. Vrabie, "Physics-informed machine learning for modeling and control of dynamical systems," in *Proc. Amer. Control Conf. (ACC)*, Jun. 2023, pp. 3735–3750.

[39] R. Neugebauer, Ed., *Werkzeugmaschinen: Aufbau, Funktion und Anwendung Von Spanenden Und Abtragenden Werkzeugmaschinen*. Berlin, Germany: Springer, 2012, doi: 10.1007/978-3-642-30078-3.

[40] C. Croitoru, "An empirical estimation of cutting force for face milling using a stationary dynamometer," *Bull. Polytech. Inst. Iasi. Mach. Construct. Sect.*, vol. 67, no. 4, pp. 53–67, Dec. 2021.

[41] R. Ströbel, M. Mau, H. Kader, D. Erd, D. Bless, S. Deucker, A. Puchta, J. Fleischer, and B. Noack. (2024). *Training and Validation Dataset 3 of Milling Processes for Time Series Prediction*. Artwork Size: 120,3 MB Pages: 120,3 MB. [Online]. Available: https://radar.kit.edu/radar/en/dataset/feFwILjideOropmh
[42] L. Zhang, H. Ai, S. Yan, H. Chen, J. Zou, and J. Zhang, "Gaussian distance weighted algorithm for geometric characteristics of three-dimensional discrete curves," *J. Appl. Math. Phys.*, vol. 12, no. 10, pp. 3599–3612, 2024.
[43] F. Mémoli, "Gromov–Wasserstein distances and the metric approach to object matching," *Found. Comput. Math.*, vol. 11, no. 4, pp. 417–487, Aug. 2011, doi: 10.1007/s10208-011-9093-5.
[44] C. Ramos-Carreño and J. L. Torrecilla, "Dcor: Distance correlation and energy statistics in Python," *SoftwareX*, vol. 22, May 2023, Art. no. 101326. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352711023000225
[45] R. Ströbel, M. Kuck, F. Oexle, H. Kader, A. Puchta, B. Noack, and J. Fleischer, "A multimodal dataset for process monitoring and anomaly detection in industrial CNC milling," *Data Brief*, vol. 63, Dec. 2025, Art. no. 112207. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235234092500928X
[46] R. Ströbel, M. Kuck, F. Oexle, and H. Kader. (2025). *A Multimodal Dataset for Process Monitoring and Anomaly Detection in Industrial CNC Milling*. [Online]. Available: https://radar.kit.edu/radar/en/dataset/hvvwn1kfwf7qt48z
[47] R. Ströbel, M. Kuck, and F. Oexle. (2025). *A Multimodal Dataset 2 for Process Monitoring and Anomaly Detection in Industrial CNC Milling*. [Online]. Available: https://radar.kit.edu/radar/en/dataset/vnnu3n9z7ndsnhfd
[48] J. Hedderich and L. Sachs, *Angewandte Statistik*. Berlin, Germany: Springer, 2016, doi: 10.1007/978-3-662-45691-0.

ROBIN STRÖBEL received the M.Sc. degree in mechanical engineering from Karlsruhe Institute of Technology (KIT), in 2022. He is currently pursuing the Ph.D. degree in incremental model training for model based anomaly detection. Since 2022, he has been a Research Associate with the Intelligent Machines and Components Research Group, wbk Institute of Production Science, where he works on various projects in the field of data science in manufacturing.

KAI NIKLAS HOFMANN received the M.Sc. degree in mechanical engineering from Karlsruhe Institute of Technology (KIT), in 2026. His research interests include production technology, process monitoring, and data science.

HAFEZ KADER (Graduate Student Member, IEEE) is currently pursuing the Ph.D. degree with the Autonomous Multisensor Systems (AMS) Research Group, Faculty of Computer Science, Otto von Guericke University Magdeburg (OVGU). His research focuses on analyzing the relevance and significance of different features and detecting anomalies in sensor data to make processes more efficient and optimize the performance of machine learning models.

JAN BAUMGÄRTNER received the M.Sc. degree in computer engineering from Heidelberg University, in 2021. Since 2022, he has been a Research Associate with the wbk Institute of Production Science, where he works on various projects in the field of intelligent manufacturing systems and robotics. Since 2026, he has been the Head of Research for industrial robotics with the wbk and the Scientific Coordinator of CRC 1574.

ALEXANDER PUCHTA received the M.Sc. degree in mechanical engineering from Karlsruhe Institute of Technology, in 2020. He is currently pursuing the Ph.D. degree in machine, plant, and process automation, focusing on the topic of autonomous modeling of feed axes for machine tools. He has been the Head of the Intelligent Machines and Components Group, wbk Institute of Production Science, since 2023.

BENJAMIN NOACK (Senior Member, IEEE) received the Ph.D. degree in computer science from the Intelligent Sensor-Actuator Systems (ISAS) Research Group, University of Karlsruhe (TH), in 2013. Since 2021, he has been a Professor with the Faculty of Computer Science, Otto von Guericke University Magdeburg (OVGU), where he heads the Autonomous Multisensor Systems (AMS) Research Group. His research focuses on distributed sensor data fusion and its applications in industrial process monitoring and autonomous mobile robotics.

JÜRGEN FLEISCHER received the degree in mechanical engineering from the Universität Karlsruhe (TH), and the Ph.D. degree from the Institute for Machine Tools and Industrial Engineering (iwb), in 1989. Since 1992, he has been holding several leading positions in industry before being appointed as a Professor and the Head of the wbk Institute for Production Technology, Karlsruhe Institute of Technology (KIT), in 2003. In addition, he has been a Visiting Professor with Tongji University, Shanghai, since 2012. His current scientific research focuses on intelligent components for machine tools and handling systems, the automation of immature processes, and agile production facilities. As a recognized member of the scientific community, he is active in German Academy of Science and Engineering (acatech), and is a member of several scientific and industrial advisory boards.