Journal of Machine Engineering, 2025, Vol. 25, No. 4, 33–52
ISSN 1895-7595 (Print) ISSN 2391-8071 (Online)

Received*: 20 June 2025* / Accepted: *01 October 2025* / Published online: *19 November 2025*

machine learning;
anomaly detection;
flexible manufacturing;
milling processes

Robin STRÖBEL[1], Marcus MAU[*1],
Marcel DIEBOLD[1], Alexander PUCHTA[1],
Jürgen FLEISCHER[1]

REVIEW OF PROCESS MONITORING AND ANOMALY DETECTION APPLICATIONS FOR CNC MILLING MACHINES IN HIGHLY FLEXIBLE PRODUCTION ENVIRONMENTS

The rapid individualization of milling operations have introduced unprecedented complexity and variant diversity, necessitating adaptive process monitoring under scarce-data conditions. This systematic literature review (SLR) takes a comprehensive look at machine learning (ML)-based approaches for process monitoring and anomaly detection in highly flexible milling processes, focusing on single-part and rapidly changing production scenarios. The fourteen most relevant studies published since 2019 were identified by adhering to established SLR frameworks. The methods are evaluated in terms of their flexibility, data efficiency, model quality and cost-effectiveness. It is revealed by the SLR that transfer learning (TL), physics-informed ML and active learning (AL) are frequently used to address the issue of limited training data whilst improving the robustness of the model. However, there a shortcoming in the integration of multiple data-efficient training strategies within holistic frameworks. Additionally, focusing on internal machine signals could reduce the burden of monitoring systems on brownfield machines. Combining monitoring via internal machine signals with AL, TL, physics-informed ML and data augmentation offers promising research directions for scalable, low-cost process monitoring in flexible manufacturing environments.

1. INTRODUCTION

Ongoing individualization [1] has led to increased complexity and diversity in manufacturing. The continuous and early detection of anomalies has become a critical challenge in these highly flexible environments with frequently changing workpiece configurations. Only precise monitoring can effectively prevent quality losses, machine damage and unplanned downtime while meeting the rising demand for transparency, automation and adaptability [2].

[1] wbk Institute of Production Science, Karlsruhe Institute of Technology, Germany
[*] E-mail: marcus.mau@kit.edu

https://doi.org/10.36897/jme/211609

In order to meet the resulting requirements for greater efficiency, customisation, sustainability and resilience, data-driven and adaptive monitoring technologies are receiving increased attention in research and industry [3]. ML in particular has proven promising as it can extract complex patterns from historical process data and detect anomalies at an early stage. Figure 1 illustrates the increase in publications on ML-based process monitoring since 2000, highlighting the growing importance of the field. Collecting experimental data reduces overall equipment efficiency (OEE). Therefore, methods that focus on limited data availability, rapidly changing workpiece geometries, or non-standardised processes are especially promising.

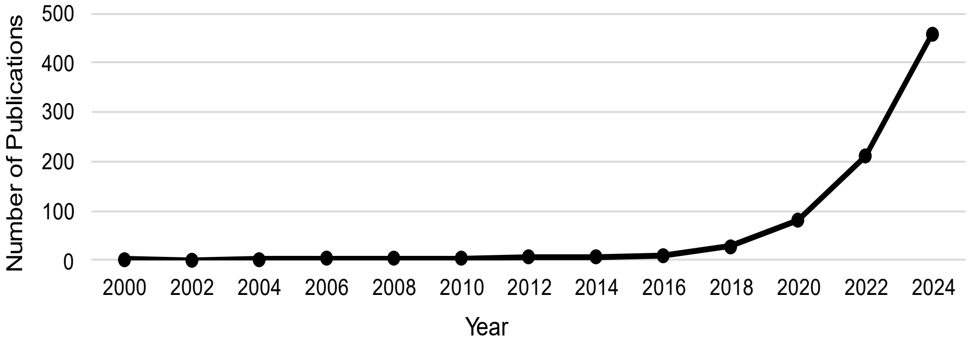

Fig. 1. ML-based anomaly detection publications in milling processes between 2000 and 2024 (scopus.com 2025).

To understand future research directions, a systematic and comprehensive overview of the current state of ML-based anomaly detection for highly flexible milling processes is necessary. Therefore, a structured literature review (SLR) will be conducted to identify and evaluate relevant studies according to the following criteria:

- Methodological approaches: Architectures and ML algorithms employed.
- Data requirements and sources: Scope, quality, and effort involved in data acquisition.
- Performance metrics: Accuracy, robustness to process changes, and adaptability.
- Practical feasibility: Additional sensors needed, computational and training overhead.

Building on this analysis, we will identify existing research gaps and potential research directions. The aim is to identify robust, data-efficient and application-oriented ML strategies that can reliably detect anomalies in highly flexible milling operations without requiring extensive data collection.

This paper is structured as follows: Section 2 describes the methodology used for the literature search and evaluation. Section 3 summarises and rates the resulting publications. Section 4 evaluates the current state of the art and presents two future research directions. Finally, Section 5 summarises the key insights and provides an outlook.

2. METHODOLOGY

The SLR in this paper is based on the guidelines developed by [4] for literature research in engineering sciences. These guidelines describe several steps, which are labelled with the letters A-E, with numbers next to the letters representing parallel sub-processes. A detailed overview of all process steps is shown in Fig. 2.

In process step A, a protocol is created at the beginning to document the procedure (A1), which documents all relevant information in parallel to processes B-D (A2). Process step B involves conceptualizing the research topic. In process step (C), the initial literature search is carried out. Process step (D) narrows down the literature relevant for the subsequent analysis and makes an initial selection. Process step (E) expands the literature search as necessary and supplements the literature consulted for further analysis.

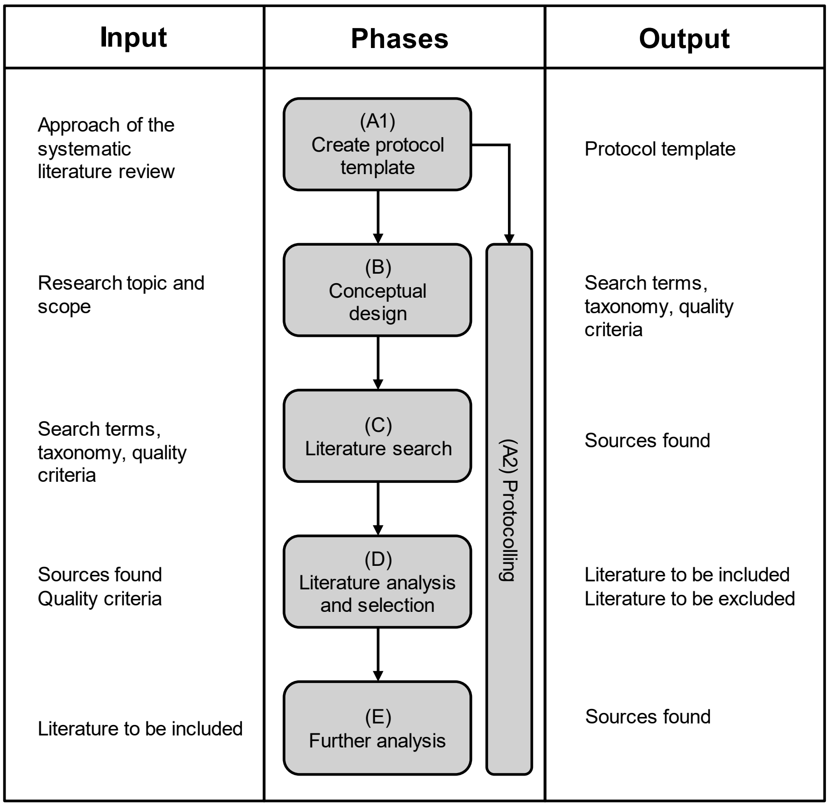
Input
Phases
Output
Approach of the systematic literature review
(A1) Create protocol template
Protocol template
Research topic and scope
(B) Conceptual design
Search terms, taxonomy, quality criteria
Search terms, taxonomy, quality criteria
(C) Literature search
Sources found
Sources found
Quality criteria
(D) Literature analysis and selection
Literature to be included
Literature to be excluded
Literature to be included
(E) Further analysis
Sources found
(A2) Protocolling

Fig. 2. Overall process of the systematic literature review in engineering (adapted from [4], p. 6).

2.1. DOCUMENTATION (PHASE A)

At the outset (A1), a detailed review protocol was drafted to capture every methodological decision, search string, filter, and data-extraction template. This protocol was maintained throughout Phases B to D (A2) to ensure full traceability and reproducibility. Key elements include: Research questions and derived sub-objectives, Review taxonomy, Quality criteria (general and topic-specific), Search strategy (databases, keywords, search fields), Screening & data-extraction templates, and Plan for synthesis of extracted findings. Documentation is done in parallel to all other phases, providing information about the workflow, from raw search results to the final analysis. The full protocol is available in the Appendix.

2.2. CONCEPTUALIZATION (PHASE B)

Conceptualization forms the methodological foundation of an SLR. It sharpens the research questions, derives specific sub-goals and requirements, and defines a clear research taxonomy. By establishing quality criteria and precise keywords, a targeted, comprehensible, and efficient literature search is made possible. This ensures that only relevant and high-quality sources are included in the analysis. Since the research questions have already been defined, the next step is to derive the sub-goals and their requirements. These are:

- Investigate whether the approaches presented address the challenges of reduced data availability.
- Investigate whether the approaches presented explicitly refer to machine-internal data and thus save hardware costs.

  The research taxonomy can now be defined from the research questions and sub-goals. The research taxonomy classifies the content and methodological orientation of the SLR. The quality criteria for publications and the search terms derived from them are significantly influenced by this. The research taxonomy was therefore configured from a selection of categories defined by [5].

- **Focus:** Methods. ML methods and concepts for process monitoring and anomaly detection are to be analyzed.
- **Objective:** Identification of key issues. The literature is to be examined with regard to the state of the art, focal points, and research gaps.
- **Perspective:** Critical position. Studies should not only be summarized, but also assessed in terms of transferability and effort.
- **Coverage:** Representative. Focus on high-quality, broadest possible spectrum of relevant work.
- **Organization:** Conceptual and methodological presentation. Structure based on method types and data sources.
- **Target audience:** Subject matter experts. In particular, engineers and researchers in the fields of mechanical engineering, manufacturing technology, and AI.

To ensure the quality and relevance of the selected publications, general and content-related quality criteria must be met. These are defined below:

General quality criteria: In general, the selected publications must be peer-reviewed and published in scientifically legitimate databases. Only publications with a defined scientific character, such as articles in journals or book chapters, are permitted. Furthermore, only publications in English are permitted. Fig. 1 shows that only a very small number of publications were produced in this field of research up to 2018. To ensure that the SLR is up-to-date, all publications before 2019 have been excluded based on their review date. If no information is available, the publication date was used.

Content quality criteria: In order to ensure a consistent and thematically focused database, specific content requirements were defined for selecting relevant literature sources. First, relevant terms must appear in the title, abstract, or keywords of the publication to ensure thematic relevance to the research question. Studies with non-specialist application relevance, particularly from areas such as the food industry or healthcare, are excluded. The studies examined must explicitly refer to milling processes: Publications on other manufacturing technologies (e.g., turning or grinding) are only permitted if they play a minor role in the study. Similarly, studies with a strong material-specific focus are not considered, as these would limit the transferability of the results to different application scenarios. Furthermore, only primary scientific studies are considered in order to ensure a consistent methodological basis. Accordingly, reviews and surveys are not considered.

Another key selection criterion concerns the technical field of the studies. Only publications that fall within the technical and scientific field, in particular the disciplines of computer science and engineering (hereinafter referred to as CSE), are considered. Studies outside this context are not considered in order to ensure consistency with the research topic.

Keywords and search terms: This SLR focuses on ML applications for anomaly detection in milling processes. Therefore, the terms “milling,” “anomaly detection,” and “machine learning” are defined as search terms. To also include recent developments in state-of-the-art methods, the terms “active learning,” “incremental learning,” and “physics-informed machine learning” are also defined as search terms. These approaches are suitable for anomaly detection in milling processes and are particularly characterized by good support in applications with reduced data availability.

To ensure that anomalies are detected not only when they have already occurred but also when there are increasing signs that a process anomaly or malfunction is imminent, anomaly prediction also plays an important role in intelligent process monitoring. Therefore, the term “Predict*” is also defined with the truncation symbol “*.” Truncation symbols (wildcards) such as asterisks (*) or question marks (?) can be used to include different terms from the same root word in the search to increase the amount of relevant literature. The search algorithms of databases that have a corresponding function for truncation characters return all publications that contain the root word “predict” in the predefined search area (e.g., prediction, predictive, predicting, etc.). In this way, the exclusion of potentially important publications due to differing formulations of content can be avoided. The SLR is further refined using the Boolean operator “AND”, resulting in the following search terms:

(1) Milling AND Anomaly Detection AND Machine Learning.
(2) Milling AND Active Learning.
(3) Milling AND Physics-Informed Machine Learning.
(4) Milling AND Incremental Learning.
(5) Milling AND Predict*.
(6) Milling AND Process Monitoring AND Predictive Maintenance.

2.3. LITERATURE SEARCH (PHASE C)

The research was conducted in the Scopus database (provided by Elsevier), IEEE Xplore Digital Library (operated by the Institute of Electrical and Electronics Engineers), and SpringerLink (platform of the scientific publisher Springer Nature). These databases were selected because of their size and relevance to the fields of CSE, as well as the great similarities in the design of their search algorithms. Therefore, we deliberately avoided using meta search engines such as Google Scholar, as their search functions do not always allow field-specific restrictions (e.g., only in the abstract or in keywords), which would have compromised the systematic nature of the research. In addition to the similarities in the basic search mechanisms, the selection of databases also took into account that they each cover different content areas and thus complement each other.

While Scopus, for example, offers particularly broad coverage of publications from international research institutions and a high degree of flexibility in filtering by subject area, IEEE Xplore is a central source for of high-quality articles on computational and electrical engineering. SpringerLink, on the other hand, is particularly relevant for mechanical engineering and related engineering disciplines, as it offers many application-oriented journals and reference books.

At the beginning of the literature search, the uniform use of "All metadata" was defined as search area. This served to methodically harmonize differences in standard searchability between the platforms (e.g., explicit search area selection options in Scopus and IEEE Xplore versus fixed standards in SpringerLink). This ensured that the search terms developed could be applied systematically and with the highest possible reproducibility in all databases. The hits were then collected and analysed separately according to the respective database.

2.4. LITERATURE SELECTION (PHASE D)

In the first step, duplicates were removed from the hits in the respective database. This was done by „OR" linking all search terms (1) to (6). In the next step, all publications that did not contain the selected keywords in the title, abstract, or separate keywords were removed. Subsequently, all publications that could be assigned to fields outside CSE and ML applications in milling processes were removed.

In particular, many publications from the field of food technology were returned, as the words "mill" and "milling" are also used in the grain-processing context. Once this was complete, the remaining hits from all databases were merged and checked again for duplicates, as many researchers use multiple databases and media to list their research. Up to this point, the filter functions of the individual databases were used to exclude irrelevant literature.

Since errors on the part of the databases could not be ruled out when filtering the literature, the remaining publications were checked individually for compliance with the defined criteria. Analogous to the filtering by the databases, publications from other fields were removed first. In the next step, publications that prioritized other manufacturing processes (e.g., turning, grinding) were removed. Furthermore, publications that indicated in their titles or abstracts that they had used a specific material for their applications were excluded. Subsequently, all review articles and surveys were removed.

Finally, the full text of the remaining publications was thoroughly examined. This process disqualified further publications that did not meet the general or content-related quality criteria. The last available publications were then selected for the comprehensive analysis in Section 4. Figure 3 illustrates the entire selection process for the relevant literature with the respective (remaining) number of publications after each iteration step described in this chapter.

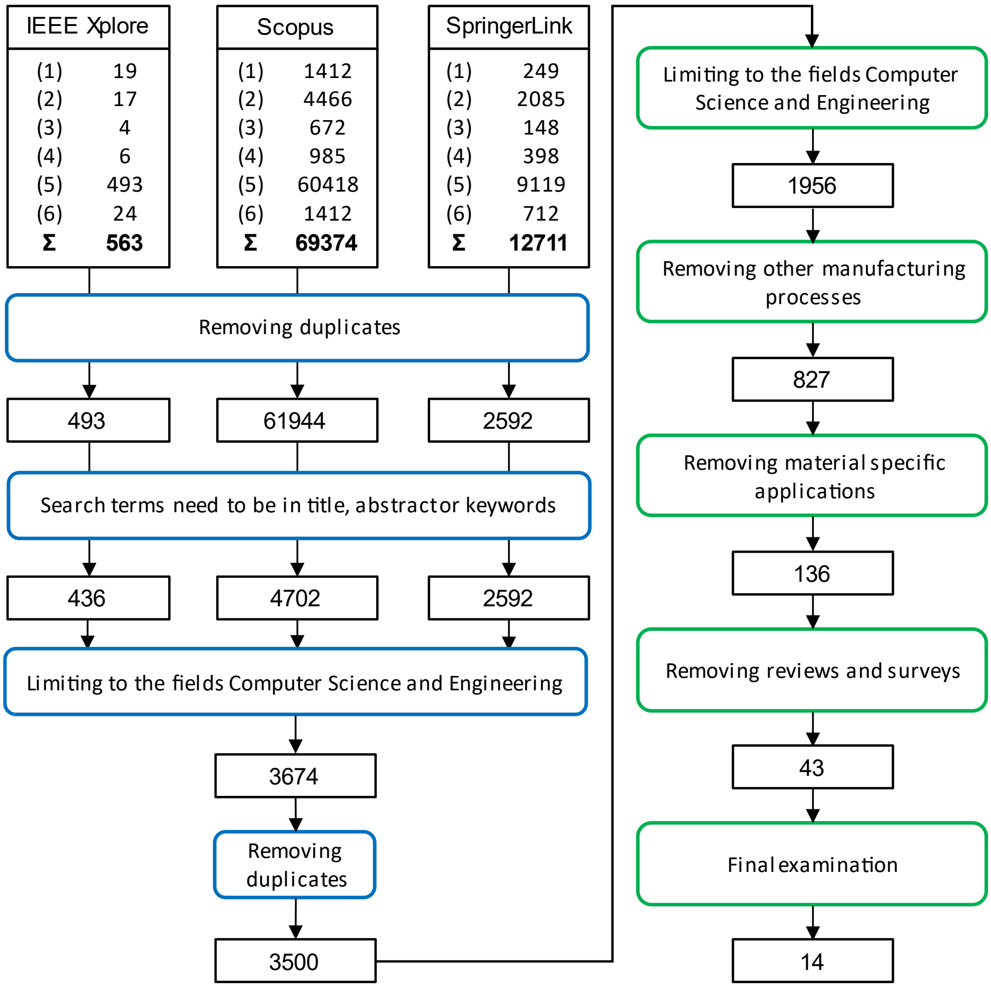

Fig. 3. Flow chart for the SLR. Filters marked in blue were performed by the databases, filters marked in green were performed manually. Numbers show how many publications are remaining after filtering

2.5. FURTHER ANALYSIS (PHASE E)

A semi-quantitative evaluation is carried out based on four defined criteria to structure the analysed approaches. Although the evaluation is based on a certain interpretation of the content, its primary purpose is to systematically visualize the relative strengths and weaknesses of the individual methods, thus enabling a comparative presentation in the following chapters. The selection and quantification of the evaluation criteria are based on the objective of identifying system and model configurations of research approaches with high potential for suitability for time- and cost-efficient single-piece production in industrial applications. The criterias are defined as:

Flexibility: The flexibility of the approach is reflected in the type and scope of the three configuration dimensions to be examined: Tool configuration (e.g. geometry, number of edges, material), workpiece configuration (e.g. material, contours, and dimensions) and machine configuration (e.g. manufacturing process, kinematics, and degree of automation) used for testing or validation in the individual publications.

Data efficiency: The amount of data required for reliable ML models varies depending on the complexity of the process. Processes with simple tool-workpiece combinations generally require less data than complex systems with many parameters. The model's performance with a reduced amount of training data is decisive for the evaluation. In addition, consideration is given to whether specific measures to reduce data requirements (e.g., through simulation, prior knowledge, or algorithmic strategies) were used.

Model quality: The quality of an ML model describes the deviation between the model prediction and reality. The smaller this deviation, the more accurate the model. In addition, the retraining effort is a decisive factor: An ML model with high quality requires less retraining effort and therefore fewer manual adjustments, so that it remains performant even when process conditions change. Both the reported model quality (e.g., based on R^2, RMSE, or F1 score) and the adjustment effort required in the event of process changes are used for the evaluation.

Cost efficiency: Low-cost approaches are a decisive advantage for implementation in industrial applications. As the exact costs of the necessary hardware and operating resources, as well as the implementation and testing of the system, are usually not disclosed, the evaluation is based on whether the necessary data can be provided by internal sources or if external sensors are required. The valuation is based on the ratio of the benefits to the costs of integrating sensors and acquiring data in the respective test setup. If existing, publicly available data sets are used, the costs of creating these are not taken into account in the assessment of cost efficiency. Only the costs directly incurred by the authors are included in the assessment.

The evaluation scales of each criteria can be seen in Tab. 1. It contains values between 1 and 5, with 1 representing the lowest score and five the highest score for the evaluation criterion. In special cases, additional scoring points may be added or deducted if certain strengths or weaknesses of the approaches presented justify this.

Table 1. Description for the scale for the criteria

Score	Flexibility	Data efficiency	Model quality	Cost efficiency
1	Simple process, no variation in the configuration dimensions	Very high data requirements, no measures for reduction	Poor model performance, no information, or inadequate metrics	High number of external sensors and very complex tests/trials
2	Variation in only one of the three configuration dimensions	Very high data requirements, minor measures for reduction	Moderate performance, extensive follow-up training required	Few additional sensors required, moderate effort for trials/tests
3	Variation in at least two of the three configuration dimensions	High data requirements with experimental measures for reduction	Good model quality for known conditions, limited generalizability	A combination of control-internal and a few additional data sources
4	Variation in all three configuration dimensions, limited in scope and depth	Use of data reduction methods (e.g., data augmentation)	High accuracy under multiple conditions, minimal adjustment required	Use of standard data from machine control, minimal expansion
5	Variation in all three dimensions, with at least two characteristics per category	Explicit strategies for data efficiency, such as AL, TL, PiML, or hybrid modeling	Very high quality (e.g., $R^2 > 0.95$, accuracy > 95%) with high robustness and no retraining	Full utilization of control-internal data, no additional sensors required

3. SUMMARY AND RATING OF SELECTED LITERATURE

This chapter provides an in-depth analysis and systematic evaluation of the technical articles identified in the SLR. The aim of this analysis is to compare the strengths and weaknesses of the approaches presented on the basis of the evaluation categories defined in Section 2 and to use them for the discussion of the research gaps identified in Section 4.
All statements made in the context of the methodological and content-related classification of the individual studies, in particular with regard to objectives, structure, algorithms used, and key results, are based directly on the original publications, unless indicated otherwise. To support traceability and as a basis for later comparative discussion, a tabular overview has been added to the appendix. This contains a condensed summary of all studies examined, including their key characteristics, content relevant to the assessment, and assigned scores. The table serves as a structured reference framework and enables transparent tracing of the classifications made in the main text. The following studies are sorted by publication date.

(1) Netzer, Palenga, & Fleischer (2022): Machine Tool Process Monitoring by Segmented Time Series Anomaly Detection Using Subprocess-Specific Thresholds [6]**.** Netzer et al. present a Mean Shift clustering-based method for identifying recurring patterns in milling processes. Time series data is segmented based on reference values (e.g., position), clustered, and grouped across all channels (e.g., current or torque). These clusters can then be used for anomaly detection. Therefore, reference values are calculated in an offline training phase. Monitoring is performed using rolling, variance-based tolerance bands, which adapt to steep signal gradients. When a live signal exceeds its tolerance band, the anomaly is flagged for expert validation. Once classified, it is fed back into the ML model to improve precision. Ten milling trials were performed on three workpiece geometries (circular, rectangular and horseshoe), where the system reliably detected instances of increased tool wear and blowhole anomalies.

Flexibility (3/5): Validation is done on a milling machine with three workpiece geometries, but only with one material and unspecified tooling, limiting its applicability to other setups.

Data Efficiency (3/5): Segmenting machining processes in recurring subprocesses leads to improved data usage, yet no explicit reduction techniques (e.g., TL or simulations) are employed.

Model Quality (3/5): Detects artificial pores and wear reliably, but is prone to overfitting if wear is present in training data and relies on position signals that may not reflect the true process.

Cost Efficiency (4/5): Uses only internal control signals, minimizing hardware costs, but still depends on manual expert validation, which hampers full automation and scalability.

(2) Deng et al. (2022): Efficient stability prediction of milling process with arbitrary tool-holder combinations based on transfer learning [7]**.** Deng et al. introduce an approach to domain adaptation based on transfer learning for efficient stability prediction when swapping tool-holder combinations in milling. They first built a large reference dataset by measuring the frequency response at different overhang lengths using impact hammer tests and accelerometers. For new tool-holder setups, only a handful of key measurements are needed for TL based adoption to the target domain. Comparing backpropagation neural networks (BPNN), random forest, and gradient-boosting decision trees on the reference data, they chose the BPNN for its lowest mean relative error. When validated on three different tool configurations (varying diameter, material, and tooth arrangement) under varied spindle speed, feed rate, and radial depth of cut, the model achieves mean relative errors between roughly 3.7% and 5.5%, markedly better than a non-TL approach. They further show that as few as three overhang measurements suffice for accurate stability forecasts, with predicted stability regions closely matching experimental results.

Flexibility (3/5): Covers multiple tool-holder combinations with varied diameters, materials, and tooth arrangements, but its applicability to different machines, process types, or workpiece geometries remains untested.

Data Efficiency (4/5): Leverages a rich reference dataset to enable accurate predictions with only a few new measurements, demonstrated to work with as few as three overhang lengths, though the initial data collection is still extensive.

Model Quality (4/5): Achieves significantly lower errors and well-matched stability diagrams for target tools, yet lacks evaluation of retraining effort required under fundamentally different configurations.

Cost Efficiency (2/5): Improves efficiency via TL but imposes high experimental costs and time for impact hammer and accelerometer measurements, especially for building the reference dataset.

(3) Fertig, Preis & Weigold (2022): Quality prediction for milling processes: automated parametrization of an end-to-end machine learning pipeline [8]**.** Fertig, Preis, and Weigold propose an end-to-end automated ML pipeline for predicting workpiece quality in milling by extracting signal components via discrete wavelet and Hilbert-Huang transform. The relevant features are identified and the best-suited model from a suite of classifiers and ensemble methods is selected and tuned automatically. Applied to four real-world datasets captured directly via edge computing on both 3-axis and 5-axis machines and covering up to 392 milled workpieces, the pipeline customizes a model for each quality-relevant geometric feature. The accuracy of the optimised models increased by between 1.87% and 15.89%, while reducing reliance on domain experts.

Flexibility (3/5): Validated on both 3-axis and 5-axis machines, with datasets collected at different times to account for machine aging and varied workpiece geometries, indicating adaptability. Unspecified quality features create an information gap for full generalization.

Data Efficiency (2/5): Employs four real-world datasets and automated feature selection to streamline training data, but lacks a clear analysis of data efficiency.

Model Quality (3/5): Accuracies of between 92% and 98% are achieved on complex geometries and reliable predictions are demonstrated. However, the effort required for retraining on new materials or tools is not addressed.

Cost Efficiency (4/5): Only internal machine signals are used and the process is automated to reduce the need for expert involvement. However, it requires a significant initial investment for pipeline setup, model selection and optimization for a specific problem.

(4) Brecher et al. (2023): Clustering of Milling Strategies using Machine Internal Data [9]**.** Brecher et al. introduce an automated method for recognizing and classifying milling strategies. Therefore, they segment the internal signals, extract features using the Time Series Feature Extraction Library, cluster recurring patterns with Hierarchical Density-

Based Spatial Clustering of Applications with Noise (HDBSCAN). The HDBSCAN model is compared with a neural network (NN) with two hidden layers for classification of unseen data. Evaluated using three workpieces with varied geometries, HDBSCAN achieved 96% cluster uniformity but struggled to distinguish similar strategies, while the NN classifier reached 80% precision, 87% recall, and an F1-score of 0.83.

Flexibility (3/5): Works on three different workpiece geometries and uses readily available control signals, but manual segmentation and untested transfer to other machines or process types limit adaptability.

Data Efficiency (3/5): Leverages high-resolution current data and avoids overfitting by training on 20% of extracted features, though no study defines minimal data requirements.

Model Quality (3/5): HDBSCAN provides good clustering performance. However, NN outperforms it for classification, and the effort required for retraining in new conditions remains unexamined.

Cost Efficiency (3/5): It requires no additional sensors, which reduces hardware costs. However, expert-driven segmentation identification adds manual overhead.

(5) Sun et al. (2023): Efficient Prediction of Stability Boundaries in Milling Considering the Variation of Tool Features and Workpiece Materials [10]**.** Sun et al. extend Deng et al.'s TL approach by incorporating spindle speed, radial cutting width, overhang length and workpiece material properties into stability predictor model. They trained a multi-layer perceptron (MLP) on a comprehensive source dataset of 18150 samples from one tool, and then retrained it on a target dataset of 915 samples consisting of four overhang lengths and two materials for new tools, through a global finetuning strategy. This strategy achieved a MAPE of between 5.6% and 6.2%, a RMSE of between 0.631 and 0.892, and an R^2 of 0.987, demonstrating clear advantages over a model trained only on the source or target dataset. Additionally, a data-volume study indicated that acceptable accuracy could be reached at 10% of the target data.

Flexibility (3/5): Tests three tools (varying diameter, material, tooth count) and three workpiece materials (Aluminum, cast iron, steel), but uses a single machine setup without exploring other kinematics or processes.

Data Efficiency (3/5): Reduces the amount of data required to 915 samples and demonstrates that 10% would lead to acceptable accuracy. However, experimental measures for reduction are required in the targeted domain.

Model Quality (4/5): The fine-tuned MLP achieves high R^2 (0.987) and low MAPE/RMSE (< 6.2%), especially at > 6000 rpm, though its robustness under new machine or process changes is untested.

Cost Efficiency (3/5): Cuts costs dramatically on target-tool measurements but still requires extensive initial impact tests and full reference datasets, limiting full scalability.

(6) Jourdan & Metternich (2023): A Nearest Neighbor-Based Concept Drift Detection Strategy for Reliable Tool Condition Monitoring [11]**.** Jourdan and Metternich propose Localized Reference Drift Detection (LRDD), which enhances concept-drift monitoring in tool-wear ML systems by locally adapting the reference dataset via k-NN before applying two-sample tests. When evaluated on a milling dataset comprising seven sensor signals (cutting forces, acceleration and acoustic emission at 50 kHz) and three labelled wear classes, LRDD achieved 100% recall while boosting precision to 97%, compared to selected baseline methods with precisions ranging from 54% to 70%.

Flexibility (2/5): Tailored to one standardized milling dataset and sensor setup. While k-NN-based localization could adapt to varied operating states, its transfer to other machines or processes is untested.

Data Efficiency (2/5): Leverages a large, multi-sensor dataset but does not explore how reducing data volume affects detection robustness, nor employ explicit data-reduction strategies.

Model Quality (3/5): Demonstrates high precision (97%) and perfect recall in drift detection, validating the localized reference approach. However, only one dataset is tested and overfitting risks remain.

Cost Efficiency (3/5): This method requires no additional sensors or hardware, and uses computationally inexpensive k-NN and statistical tests. This makes it a cost-effective enhancement to existing monitoring systems. However, only external sensors are present in the validation data.

(7) Li et al. (2023): Incremental Learning of LSTM-Autoencoder Anomaly Detection in Three-Axis CNC Machines [12]**.** Li et al. propose an Incremental Ensemble LSTM-Autoencoder (TL-IE) that combines TL for initialization on a new machine with an ensemble of “weak” LSTM learners added over time to handle concept drift. Reconstruction losses are thresholded via a three-sigma rule to flag anomalies. A directly trained LSTM, a directly trained IE LSTM, a TL LSTM and a TL-IE LSTM were compared. The TL-IE LSTM was shown to achieve 93.57% accuracy in detecting unstable cutting conditions (compared to 95.62% for directly trained IE LSTM and 75.23% for pure TL), demonstrating comparable performance to high-data methods but requiring much less data.

Flexibility (2/5): The incremental design adapts to evolving machine behavior over long runs, but validation is limited to a single machine. The dataset consists of varied spindle speed and depth of cut, leaving approaches transferability to other machines, tools, or materials untested.

Data Efficiency (3/5): Although TL reduces the amount of data required for new machines (30 000 data points in the setup), the base model still requires a large number of samples. Furthermore, no systematic analysis of minimum data requirements is provided.

Model Quality (4/5): The TL-IE approach boosts accuracy from 75.23% (TL alone) to 93.57% without manual retraining matching large dataset baselines.

Cost Efficiency (2/5): Using the ensemble and large LSTM-AE well as the use of external accelerometers. This makes practical deployment resource-intensive.

(8) Zhu et al. (2024): Physics-Informed Deep Learning for Tool Wear Monitoring [13]**.** Zhu et al. introduce a PiML framework for tool wear monitoring that incorporates physical knowledge into a data-driven model at multiple stages. Representative physical information is therefore selected, and four practical PiML methods are proposed. The periodic nature of the cutting signal is integrated into the design of the structure, and attention-based multilayer feature extraction, a multitask loss function and residual learning are used within attention-based dual-scale hierarchical LSTM and BiLSTM architectures. The approach is validated under eight different high-speed milling conditions (varying spindle speed, cutting depth and feedrate) using cutting-force sensors and microscopic wear measurements.

Compared to a purely data-driven baseline, the approach reduces the mean absolute percentage error by 42%, the mean final prediction error by 43% and the maximum final prediction error by 63%, yielding smoother predictions. Additionally, they identified that the periodic-based structure, attention-based feature extraction and multitask loss function improved the model's accuracy. However, residual learning did not benefit the PiML model.

Flexibility (2/5): The approach was tested with two-flute end mills on a 3-axis machine using one material. Transferability to other tools, machine types or complex geometries has not been proven. While the general ideas are transferable, the solution is tailored to tool wear.

Data Efficiency (3/5): Embeds physics to lower data demands and shows large error reductions, but lacks any study on minimal data requirements or the impact of data reduction.

Model Quality (3/5): Achieves improved performance and smoother prognoses under new conditions, yet the need for retraining on different tools or machines remains unspecified.

Cost Efficiency (3/5): Relies on expensive high frequency force sensors and microscopic wear measurements, which heighten hardware and measurement.

(9) Araghizad et al. (2024): Milling process monitoring based on intelligent real-time parameter identification for unmanned manufacturing [14]**.** Araghizad et al. used PiML to enhance a linear cutting force model to generate a cutting-force dataset. The simulation was validated using milling force measurements under varying process parameters using a piezoelectric dynamometer and two end mills with different diameters and tooth counts. This data was subsequently used by different ML-models (Support Vector Machine, RF and LSBoost) to identify varying process parameters. It is claimed that force prediction improved to an R^2 value of 97% for unseen datasets, with LSBoost achieving the lowest RMSE. The accuracy of the parameter identification models exceeds 95% for all machining parameters. LSBoost achieved the best results with R^2 values of 97.89% for the axial depth of cut, 98.65% for the radial depth of cut, and 97.32% for the feedrate. On a curved aluminum test part, the model was capable of predicting changed cutting depth due to displacement of the test part.

Flexibility (2/5): The initial training dataset covers two cutting tools, various spindle speeds, feed rates, and depths of cut. However, it lacks investigations into machine type and material and general transferability to yet unseen conditions.

Data Efficiency (3/5): Soft sensors are used to reduce the number of experiments, yet a comprehensive dataset was used. The extent to which the approach is beneficial and the amount of data required for reliable performance remain unclear.

Quality (4/5): The approach delivers high R^2 (>95%) and low RMSE across all targets, demonstrating strong predictive accuracy. However, its robustness when using new tools, materials or machines without retraining has not been evaluated.

Cost Efficiency (3/5): It reduces the experimental workload through soft-sensor data, but this depends on external dynamometer measurements. However, the optional use of internal sensors is also mentioned.

(10) Hassan, Sadek & Attia (2024): In-process self-configuring approach to develop intelligent Tool Condition Monitoring Systems [15]**.** Hassan, Sadek and Attia have introduced a self-configuring tool wear monitoring system that combines a wavelet-scattering convolutional neural network (WSCNN) with an LSTM prediction model. The system synthesizes WSCNN features based on accelerometer signals at an early stage of the tool's life, and uses the LSTM to predict the values of the features in a worn state. A subsequent SVM-based classifier is used to define the tool wear state by separating the two tool classes "usable" and "worn". The system was validated using 558 milling tests on two 5-axis machines with different tools, materials, sensors, and process parameters. It achieved a classification accuracy of over 94% in unseen conditions.

Flexibility (5/5): It has been validated on two 5-axis machines under varying conditions, demonstrating exceptional transferability without the need for manual adjustments.

Data Efficiency (4/5): It uses 1–3 s of data per tool for model transfer, yet no study for minimal data needs was provided.

Model Quality (4/5): It delivers ≥94% accuracy across all unseen combinations and uses a voting scheme to filter out inconsistent predictions.

Cost Efficiency (3/5): It reduces retraining effort, but it relies on high-resolution accelerometers, which limits its cost advantage.

(11) Wiederkehr, Finkeldey & Siebrecht (2024): Reduction of experimental efforts for predicting milling stability affected by concept drift using transfer learning on multiple machine tools [16]**.** Wiederkehr, Finkeldey and Siebrecht propose a two-stage, process-informed TL strategy for predicting the stability of milling processes on new target machines. A gradient-boosting ensemble was trained on 1037 source machine experiments (80% for training and 20% for testing), employing acoustic emissions (microphone) and spindle-integrated current sensors. Chatter vibrations were characterized by milling a thin workpiece with linearly increasing radial depth of cut at varying spindle speeds and tool wear conditions. A normalized RMSE of 10.78% ± 7.11% was achieved on the source machine. To reach comparable results with the retrained model on a new target machine, 173 additional experiments were required. Therefore, the number of cutting experiments required was reduced by 83.32%. However, 1210 tests must be conducted in total across both machines.

Flexibility (2/5): It demonstrates the transfer between two machines, but its applicability to other tool types, materials or part geometries remains untested.

Data Efficiency (3/5): Although the number of experiments required is reduced from 1037 to 173 by TL, a total of 1210 experiments are still needed. It is unclear if all of the 1037 source elements were needed or if a smaller number would have been sufficient.

Model Quality (4/5): It achieves reliable stability predictions (normalized RMSE of 10.78%) through stepwise ensemble updates, without the need for full retraining.

Cost Efficiency (2/5): Its practical usage is limited by the requirement for external microphones and the number of experiments needed.

(12) Hyun et al. (2024): Encoding Time Series as Images for Anomaly Detection in Manufacturing Processes Using Convolutional Neural Networks and Grad-CAM [17]**.** Hyun et al. convert multivariate sensor time series into four image representations (Gramian Summation/Difference Angular Fields, Markov Transition Field and Recurrence Plot). 10 different CNN network models are used to classify normal and anomalous machining states based on these images, together with Grad-CAM to provide insights into the detection process. Validation was performed using the IEEE PHM 2010 dataset, which includes vibration, force and acoustic emission signals (DS1), as well as a real-world CNC dataset consisting of force and acoustic emission data (DS2). ResNet50 achieves 99.6% accuracy on

DS1 (Recurrence Plots), and VGG16 achieves 91.8% accuracy on DS2 (Gramian Angular Summation Fields).

Flexibility (2/5): The image-based transformation and choice of architectures suggest broad applicability. However, the lack of information on the real-world dataset (DS2) makes it difficult to evaluate the transferability to other machines, tools, tollpaths or materials.

Data Efficiency (2/5): Requires tens of thousands of generated images (36000 on DS1) and lacks transparency on DS2's size or diversity, so minimal data requirements remain unknown.

Model Quality (2/5): This approach achieves an accuracy rate of up to 99.67% for DS1, but this drops to 91.82% for DS2. Without clear details of DS2, this is due to missing sensors, increased process complexity, or insufficient data.

Cost Efficiency (2/5): Both datasets consist of external sensors, especially force and acoustic emission, which are expensive and require high data processing capacities. Although moderate computing power is required for CNN training and image generation, insufficient information on DS2's preprocessing makes it difficult to assess the implementation effort.

(13) Denkena et al. (2024): Active learning for the prediction of shape errors in milling [18]**.** Denkena et al. introduce an AL framework that iteratively selects high-uncertainty, dissimilar measurement points via the machine's built-in probe to predict shape errors in milled parts. Starting with one part, the model retrains as new labeled data arrives. Validation was performed using a 5-axis and a 3-axis dataset for repetitive pocket milling. The approach is compared against passive learning (PL) under data budgets of 10%, 25%, and 50%. AL reduces the RMSE by up to 5% and cuts the standard deviation by 85% compared to PL. When evaluated on unseen holdout data partitions, AL does not always lead to better generalization than the PL strategy. While AL produced similar results on one dataset, it led to reduced RMSE on the second.

Flexibility (2/5): It has been validated on both 5-axis and 3-axis processes involving different tool-wear states and materials, demonstrating its adaptability within milling contexts. However, it has not been tested on other machines or complex geometries.

Data Efficiency (4/5): The usefulness of AL in scenarios with limited data was demonstrated, showcasing AL's capacity to maximize information per measurement. However, the results were mixed when it came to unseen holdout data partitions.

Model Quality (3/5): Although AL reduces RMSE by up to 10% in some cases, there is no significant difference in performance of AL and PL on the first dataset.

Cost Efficiency (3/5): The approach requires an external tool probe. Using targeted measurements reduces inspection time but retraining costs are not detailed.

(14) Lin et al. (2024): A self-adaptive machining parameters adjustment Method for stabilizing the machining-induced surface roughness [19]**.** Lin et al. propose a self-adaptive control loop that stabilizes surface roughness by combining a physics-driven roughness model with a CNN-LSTM feature extractor and a TL step to align source and target domains. A stochastically gradient-descent optimizer is used to adjusts federate and spindle speed during machining. Validated on a 3-axis milling machine under five parameter sets, the system predicts roughness with an RMSE of 0.0703 under varied process parameters. The computation time for the control procedure is 14.72 seconds on the test system, maintaining roughness within the target tolerance band of 1.6 μm over 11 time periods.

Flexibility (2/5): The model adapts dynamically to varied feed, speed and depth combinations on a 3-axis machine, but has not been tested with different tools, materials or toolpaths, or with different machine configurations.

Data Efficiency (4/5): Fuses a physics model with TL-enhanced CNN-LSTM to reduce data requirements and achieve high accuracy from limited experiments. However, no study was performed to investigate minimal data requirements.

Model Quality (3/5): The prediction model reaches an RMSE of 0.0703, yet robustness to new workpiece materials or geometries without retraining remains unverified.

Cost Efficiency (3/5): The system requires force sensors with a sampling rate of 50 kHz and integration into the CNC control system, which incurs application and hardware costs.

4. EVALUATION

The key insights of the SLR emerge with the evaluation of the four criteria in Fig 4. Most approaches result in sufficient model quality at moderate flexibility, data efficiency and cost efficiency. Figure 5 compares the flexibility of the approaches with their data efficiency and model quality. The green area highlights the target zone: Approaches with high generalizability and low training data requirements. It is striking that only one approached meets both requirements to a high degree. This underscores that flexibility requires high data volumes.

Publication	Flexibility	Data Efficiency	Model Quality	Cost Efficiency
(1) Netzer, Palenga & Fleischer (2022)	◑	◑	◑	◕
(2) Deng et al. (2022)	◑	◕	◕	◔
(3) Fertig, Preis & Weigold (2022)	◑	◔	◑	◕
(4) Brecher et al. (2023)	◑	◑	◑	◑
(5) Sun et al. (2023)	◑	◑	◕	◑
(6) Jourdan & Metternich (2023)	◔	◔	◑	◑
(7) Li et al. (2023)	◔	◑	◕	◔
(8) Zhu et al. (2024)	◔	◑	◑	◑
(9) Araghizad et al. (2024)	◔	◑	◕	◑
(10) Hassan, Sadek & Attia (2024)	●	◕	◕	◑
(11) Wiederkehr, Finkeldey & Siebrecht (2024)	◔	◑	◕	◔
(12) Hyun et al. (2024)	◔	◔	◔	◔
(13) Denkena et al. (2024)	◔	◕	◑	◑
(14) Lin et al. (2024)	◔	◕	◑	◑

Fig. 4. Fulfillment of the criteria of the identified literature. An empty ball corresponds to a score of 1, and a full ball corresponds to a score of 5

Since low-cost approaches facilitate implementation, cost efficiency was compared with data efficiency and flexibility. Based on the cluster in the centre of Figure 6 (left), missing data efficiency can be addressed through additional sensors, complex measurement technology or complex implementation. This discrepancy highlights the conflicting objectives of ensuring economically viable implementation while also meeting the demand for data-efficient models. Figure 6 (right) provides an additional perspective by comparing cost efficiency with flexibility.

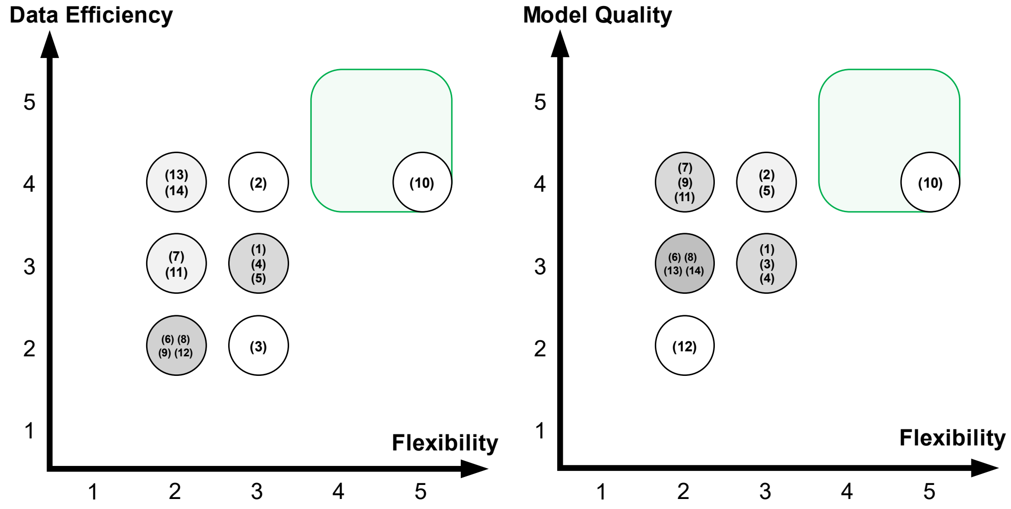

Fig. 5. Comparison of data efficiency (left) and model quality (right) with flexibility of the identified approaches

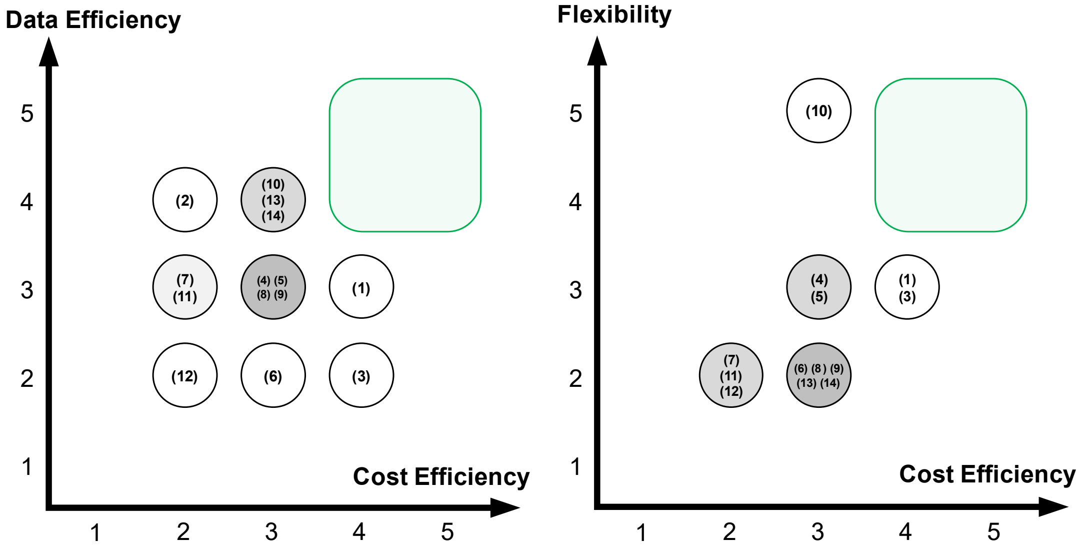

Fig. 6. Comparison of data efficiency (left) and flexibility (right) with cost efficiency of the identified approaches

The visualizations show that high flexibility alone is not enough to classify a system as suitable for single-piece production. Promising approaches for industrial applications are characterized by the simultaneous fulfilment of several evaluation criteria, especially data and cost efficiency. Therefore, existing research gaps cannot be categorized in a one-dimensional manner, but rather along two distinct lines of enquiry arising from the practical perspective of industrial single-part manufacturing.

First research focus: Efficiency-oriented ML strategies

There is a need for the design of ML-based monitoring systems that enable reliable predictions even under restrictive constraints, particularly with regard to limited data availability and low implementation resources. Particularly great potential lies in the targeted combination of data efficient methods, such as linking AL, PiML, TL, and data augmentation. While individual studies have integrated one or two of these approaches, systematic analysis shows that no work to date has combined all four methods in a comprehensive, practical approach. The development of such a hybrid, data-efficient monitoring system could make a decisive contribution to the industrial applicability of ML based monitoring in data-poor and dynamic manufacturing environments, especially if additional sensor technology can be avoided. In conjunction with the exclusive use of machine-internal signals such as spindle and axis currents, loads, or other control system data, this opens up the potential for data- and cost-efficient systems. Despite these prospects, the methodologically sound integration of these strategies into a real-time-capable and application-validated solution remains unexplored.

Second research focus: Flexibility-oriented and scalable monitoring systems

Parallel to the question of efficiency, future research should focus on the development of systems that offer a high degree of structural transferability and expandability. While many state-of-the-art models are already transferable in terms of their learning architecture, there is a lack of systematic data management. This would allow long-term adaptability, reproducibility, and integration into existing production systems. Database-supported architectures could make a significant contribution by enabling the consistent linking of workpiece, tool, and machine parameters, the structured storage of model instances and their training history, and the semantic classification of model quality and scope of validity. This structural foundation is necessary to make ML-supported systems maintainable, transferable, and modular in the long term, for example, in the sense of a digital knowledge base for future machining processes. However, storing large amounts of data can lead to redundancies, unnecessary storage requirements, and increasing complexity in model maintenance if the data is not selected carefully. For the industrial implementation of flexible and efficient monitoring systems, it is therefore crucial to only store data that offers concrete value for model quality. A promising solution may be the implementation of learning strategies that select and store data with high predictive quality incrementally. In this way, the predictive quality can be improved through selectively enriched training data without unnecessarily increasing the amount of data. This creates the basis for ML systems that continuously evolve while remaining lean and adaptable. This approach could further be expanded by using modern approaches that make use of flexible model structures, e.g., a mixture of experts, where sub-models can be added or adapted when data is added to the database.

Overall, the research gaps identified cannot be reduced to a single technical challenge, but rather highlight the need for a cross-method perspective that aims at both efficiency and transferability. Existing solutions often have to balance these two objectives, as simultaneous optimization of data efficiency and flexibility has so far only been achieved in isolated cases.

5. SUMMARY AND DISCUSSION

This review on ML-based monitoring of milling processes reveals a dilemma between model quality and practical deployability. Most approaches achieve sufficient prediction accuracy. Yet those gains frequently depend on extensive sensor suites and large training datasets, which constrain industrial scalability. Conversely, data-efficient methods reduce experimental burden but tend to sacrifice robustness or generalizability when confronted with new tools, materials, or machine kinematics. Only a few works approach an equilibrium between low data demands, broad transferability, and sustained accuracy. The SLR highlights two critical research fronts. First, there is a need for unified frameworks that systematically integrate complementary data-reduction strategies such as AL, PiML, TL, and data augmentation into an integrated pipeline. These promise to deliver reliable monitoring under scarce-data regimes while avoiding added hardware complexity. Second, industrial adoption hinges on scalable data-management infrastructures that support context-aware model evolution. Relational or semantic databases, augmented with smart selection criteria for long-term data retention, would enable traceable versioning, efficient reuse of prior process knowledge, and lean maintenance of model validity domains.

Looking ahead, future work should emphasize balanced optimization across model quality, data efficiency, cost efficiency and flexibility, rather than chasing single-metric performance. It will be essential to develop training strategies and ML architectures that can dynamically adjust to domain shifts in order to embed intelligent, adaptive monitoring in high-variety, resource-limited manufacturing environments.

ACKNOWLEDGEMENTS

This work was supported by the Federal Ministry for Economic Affairs and Climate Action (BMWK), based on a decision of the German Bundestag via Gesellschaft zur Förderung angewandter Informatik e.V.—GFaI under Grant 22849 BG/2.

REFERENCES

[1] KOREN Y., 2010, *The Global Manufacturing Revolution: Product ‐ Process ‐ Business Integration and Reconfigurable Systems*, 1. Aufl. Wiley, https://doi.org/10.1002/9780470618813.

[2] KAMAT P., SUGANDHI R., 2020, *Anomaly Detection for Predictive Maintenance In Industry 4.0- A Survey*, E3S Web Conf., Bd. 170, 02007, https://doi.org/10.1051/e3sconf/202017002007.

[3] LASI H., FETTKE P., KEMPER H.-G., FELD T., HOFFMANN M., 2014, *Industrie 4.0*, Wirtschaftsinf, Bd. 56/4, 261–264, Aug. https://doi.org/10.1007/s11576-014-0424-4.

[4] PRIELIPP R., EMANUEL C., WILSKY P., 2022, *How to Conduct a Systematic Literature Review in the Field of Engineering: A Practical Guide*, Technische Universität Chemnitz, [Online]. Verfügbar unter: https://nbn-resolving.org/urn:nbn:de:bsz:ch1-qucosa2-794654.

[5] COOPER H.M., 1988, *Organizing Knowledge Syntheses: A Taxonomy of Literature Reviews*, Knowledge in Society, Bd. 1/1, 104–126, https://doi.org/10.1007/BF03177550.

[6] NETZER M., PALENGA Y., FLEISCHER J., 2022, *Machine Tool Process Monitoring by Segmented Timeseries Anomaly Detection Using Subprocess-Specific Thresholds*, Prod. Eng. Res. Devel., Bd. 16/5, 597–606, https://doi.org/10.1007/s11740-022-01120-3.

[7] DENG C., TANG J., MIAO J., ZHAO Y., CHEN X., LU S., 2023, *Efficient Stability Prediction of Milling Process with Arbitrary Tool-Holder Combinations Based on Transfer Learning*, J. Intell. Manuf., Bd. 34/5, 2263–2279, https://doi.org/10.1007/s10845-022-01912-5.

[8] FERTIG A., PREIS C., WEIGOLD M., 2023, *Quality Prediction for Milling Processes: Automated Parametrization of an End-To-End Machine Learning Pipeline*, Prod. Eng. Res. Devel., Bd. 17/2, 237–245, https://doi.org/10.1007/s11740-022-01173-4.

[9] BRECHER C., OCHEL J., AHMED M.M., FEY M., 2023, *Clustering of Milling Strategies Using Machine Internal Data*, 2023 IEEE, 27th International Conference on Intelligent Engineering Systems (INES), Nairobi, Kenya: IEEE, 000077–000082. https://doi.org/10.1109/INES59282.2023.10297786.

[10] SUN H., DING H., DENG C., XIONG K., 2023, Efficient Prediction of Stability Boundaries in Milling Considering the Variation of Tool Features and Workpiece Materials, Sensors, Bd. 23/21, 8954, Nov. https://doi.org/10.3390/s23218954.

[11] JOURDAN N., METTERNICH J., 2023, *A Nearest Neighbor-Based Concept Drift Detection Strategy for Reliable Tool Condition Monitoring*, Proceedings of the 3rd International Workshop on Software Engineering and AI for Data Quality in Cyber-Physical Systems/Internet of Things, San Francisco CA USA: ACM, 1–4. https://doi.org/10.1145/3617573.3618027.

[12] LI E., LI Y., BEDI S., MELEK W., GRAY P., 2023, *Incremental Learning of LSTM-Autoencoder Anomaly Detection in Three-Axis CNC Machines*, Int. J. Adv. Manuf. Technol., 130/3-4, 1265–1277, https://doi.org/10.1007/s00170-023-12713-2.

[13] ZHU K., GUO H., LI S., LIN X., 2024, *Physics-Informed Deep Learning for Tool Wear Monitoring*, IEEE Trans. Ind. Inf., 20/1, 524–533, https://doi.org/10.1109/TII.2023.3268407.

[14] ARAGHIZAD E.A., TEHRANIZADEH F., PASHMFOROUSH F., BUDAK E., 2024, *Milling Process Monitoring Based on Intelligent Real-Time Parameter Identification for Unmanned Manufacturing*, CIRP Annals, 73/1, 325–328, https://doi.org/10.1016/j.cirp.2024.04.083.

[15] HASSAN M., SADEK A., ATTIA H., 2024, *In-Process Self-Configuring Approach to Develop Intelligent Tool Condition Monitoring Systems*, CIRP Annals, 73/1, 81–84, https://doi.org/10.1016/j.cirp.2024.04.049.

[16] WIEDERKEHR P., FINKELDEY F., SIEBRECHT T., 2024, *Reduction of Experimental Efforts for Predicting Milling Stability Affected by Concept Drift Using Transfer Learning on Multiple Machine Tools*, CIRP Annals, 73/1, 301–304, https://doi.org/10.1016/j.cirp.2024.04.084.

[17] HYUN Y.-J., YOO Y., KIM Y., LEE T., KIM W., 2024, *Encoding Time Series As Images for Anomaly Detection in Manufacturing Processes Using Convolutional Neural Networks and Grad-CAM*, Int. J. Precis. Eng. Manuf., 25/12, 2583–2598, https://doi.org/10.1007/s12541-024-01069-6.

[18] DENKENA B., WICHMANNA M., ROKICKIB M., STÜRENBURGA L., 2024, *Active Learning for the Prediction of Shape Errors in Milling*, Procedia CIRP, 126, 324–329, https://doi.org/10.1016/j.procir.2024.08.364.

[19] LIN Y., ZHOU S., SHU L., WU P., 2024, *A Self-Adaptive Machining Parameters Adjustment Method for Stabilizing the Machining-Induced Surface Roughness*, Int. J. Adv. Manuf. Technol., 135/5–6, 2019–2035, https://doi.org/10.1007/s00170-024-14631-3