Proceedings of the Second Workshop on AI in Production

Held at KI2025 in Potsdam, Germany

Judith Knoblach Martin Krockert Erik Voigt

Sep. 16, 2025

10.20378/irb-114337

Introduction

The 2nd Workshop on Artificial Intelligence in Production (AIP) was held as part of the 48th German Conference on Artificial Intelligence in Potsdam on September 16, 2025. The workshop was organised jointly by academic and industrial partners from the University of Bamberg, HTW Dresden, and the BMW Group.

The aim of the AIP workshop is to bring together researchers and practitioners from Artificial Intelligence (AI) and production to exchange ideas, experiences, and solutions for applying AI across production planning, control, and optimisation. By fostering interdisciplinary dialogue and sharing transferable approaches, the workshop seeks to advance effective, scalable, and sustainable AI applications in production and related domains, and strengthening the connection between academic research and industrial practice.

In the 2025 edition, seven researchers and practitioners presented their latest work in the field. The workshop received eight submissions, of which seven were accepted after a peer-review process involving two members of the programme committee per paper. We would like to thank all authors, reviewers, and participants for their valuable contributions, which made the workshop a success.

Contents

Fusion-Based Neural Generalization for Predicting Temperature Fields in Industrial PET Preform Heating

Ahmad Alsheikh[1,2] ahmad.alsheikh@krones.com
Andreas Fischer[2] andreas.fischer@th-deg.de

[1] *KRONES AG, Böhmerwaldstr. 5, 93073 Neutraubling, Germany*
[2] *Deggendorf Institute of Technology, Dieter-Görlitz-Platz 1, 94469 Deggendorf, Germany*

Accurate and efficient temperature prediction is critical for optimizing the preheating process of PET preforms in industrial microwave systems prior to blow molding. We propose a novel deep learning framework for generalized temperature prediction. Unlike traditional models that require extensive retraining for each material or design variation, our method introduces a data-efficient neural architecture that leverages transfer learning and model fusion to generalize across unseen scenarios. By pretraining specialized neural regressor on distinct conditions such as recycled PET heat capacities or varying preform geometries and integrating their representations into a unified global model, we create a system capable of learning shared thermal dynamics across heterogeneous inputs. The architecture incorporates skip connections to enhance stability and prediction accuracy. Our approach reduces the need for large simulation datasets while achieving superior performance compared to models trained from scratch. Experimental validation on two case studies material variability and geometric diversity demonstrates significant improvements in generalization, establishing a scalable ML-based solution for intelligent thermal control in manufacturing environments. Moreover, the approach highlights how data-efficient generalization strategies can extend to other industrial applications involving complex physical modeling with limited data.

Keywords
Industrial Microwave • Transfer Learning • Generalized Regression • Temperature Field Prediction • Model Fusion • Finite Element Simulation • Data-Driven Modeling • Intelligent Manufacturing

1 Introduction

Polyethylene-terephthalate (PET) preforms are small, injection-molded plastic parts that are used to form bottles and containers through a blow molding process [1]. Before injection molding, the PET preforms need to be heated to a specific temperature to make them easier to mold. Traditionally, infrared (IR) heating has been the industry standard, but its limitations including energy inefficiency and lack of precise spatial control—have spurred interest in alternative technologies.

Microwave (MW) heating has emerged as a promising alternative due to its volumetric heating capabilities, faster processing times, and potential for selective energy deposition [2]. Preheating PET preforms carefully and consistently is important for producing high-quality containers before the blow molding process. It provides advantages when the heating is distributed equally along the preform, resulting in better and higher quality bottles by ensuring that the material is uniformly heated and has consistent properties throughout.

This leads to more precise and predictable molding of the preform, resulting in bottles with consistent wall thickness and better clarity. In contrast, uneven heating can lead to defects such as variations in wall thickness, haze, or stress marks, which can compromise the quality and performance of the final product [3]. However, this becomes more challenging to achieve due to the huge variations available in preforms to suit different bottle and container sizes and shapes. Manufacturers can produce preforms in a range of weights and lengths, with different neck finishes and thread designs to match the requirements of various bottling applications [4]. The specific design of a PET preform will depend on the final container shape and size needed and will be determined by the requirements of the customer.

Recent advancements in deep learning, offer new ways for modeling complex physical phenomena like microwave heating. Yet, training deep neural networks typically requires large datasets, which can be prohibitively expensive and time-consuming to generate through high-fidelity simulations or experiments.

This work proposes a data-efficient, generalizable deep learning framework for predicting the 2D temperature distribution within PET preforms subjected to microwave heating. Although 3D modeling is theoretically more comprehensive, this study focuses on 2D temperature distribution due to the rotational symmetry of PET preforms during the heating process. Expanding to a full 3D model would not yield significantly different results but would substantially increase computational cost and data requirements. Therefore, a 2D approach offers a more efficient and equally accurate alternative for this specific application.

Our method incorporates two key innovations: (1) transfer learning through fine-tuning, which enables leveraging knowledge from one set of material or geometric conditions to another, and (2) model fusion, where multiple specialized models are combined into a single, robust predictor that generalizes well across unseen scenarios. Two practical case studies are examined:

- **Case Study 1:** Generalization across variations in the heat capacity of PET, relevant for incorporating recycled materials.
- **Case Study 2:** Generalization across different preform geometries, a common variability in manufacturing lines.

By using limited datasets (just 450 to 550 samples per category), we demonstrate that our approach significantly reduces data requirements while maintaining high prediction accuracy. The proposed methodology offers a scalable and intelligent alternative to traditional modeling, paving the way for smart, adaptive thermal control systems in plastic manufacturing.

Related Work

Recent advancements in transfer learning, fine-tuning, and model fusion have significantly enhanced neural network performance across various domains. However, their application to PET preform heating remains limited.

Most existing studies focus on image and signal classification tasks. Liu et al. [5] and Zhou et al. [6] used transfer learning in contexts like garbage sorting and medical imaging. Ghazi et al. [7] and Chakraborty et al. [8] demonstrated adaptability in plant identification and human action recognition, while Korzh et al. [9] and Whitney et al. [10] showed improved performance using ensemble models.

Emerging methods such as model merging [11] and AdapterFusion [12] further improve task generalization. Ge and Yu [13] and Zhai et al. [14] explored multi-fidelity and multi-channel fusion to enhance predictive accuracy.

(a) (b)

Figure 1: (a) Molded and defective PET bottles; (b) Preforms with different designs.

Machine learning use in industrial heating is still rare. Notable efforts include Hsieh [15], who applied deep reinforcement learning to control blow molding temperatures, and Zhai et al. [16], who used transfer learning in heating furnace prediction. Di Barba et al. [17] also introduced neural metamodels for adaptive induction heating control.

Traditionally, industrial heating systems have relied on physics-based models and heuristic control strategies. However, these often struggle with dynamic production environments. In PET blow molding specifically, infrared (IR) heating remains the dominant method for preheating preforms. Conventional infrared (IR) ovens are commonly used to heat PET preforms prior to blow molding. However, IR heating is limited to surface absorption, offers slow thermal response, and often struggles with achieving uniform radial temperature profiles [2]. In contrast, microwave (MW) heating penetrates deeply into the material, enabling volumetric absorption and significantly shorter heat-up times—up to 80 [18].

2 Methodology

2.1 Applicator Design and Simulation

The design and functionality of the applicator play a critical role in achieving precise heating patterns, particularly in applications requiring uniform temperature distribution. The proposed applicator consists of a rectangular cavity with internal dimensions of 250 mm in width, 190 mm in length, and 150 mm in height. It is equipped with dielectric slabs that fine-tune the electromagnetic field distribution.

The cavity is energized via a Type-N coaxial antenna, located at the center of the bottom wall and aligned along the z-axis. It is configured to generate the TE_{101} electromagnetic mode at 915 MHz [18], a standard frequency in industrial microwave applications.

The heating process is highly dependent on the geometry of the PET preform, including parameters such as wall thickness, neck dimensions, and overall shape of which directly affects the characteristics of the final container. To manipulate the field distribution, the applicator includes dielectric slabs composed of two stacks of 16 PTFE (polytetrafluoroethylene) sheets. These slabs are microwave-transparent and are oriented parallel to the y-axis.

The slabs are placed symmetrically on either side of the preform, with their positions adjustable along the x-axis to control the distance from the preform. Each slab measures 25 mm

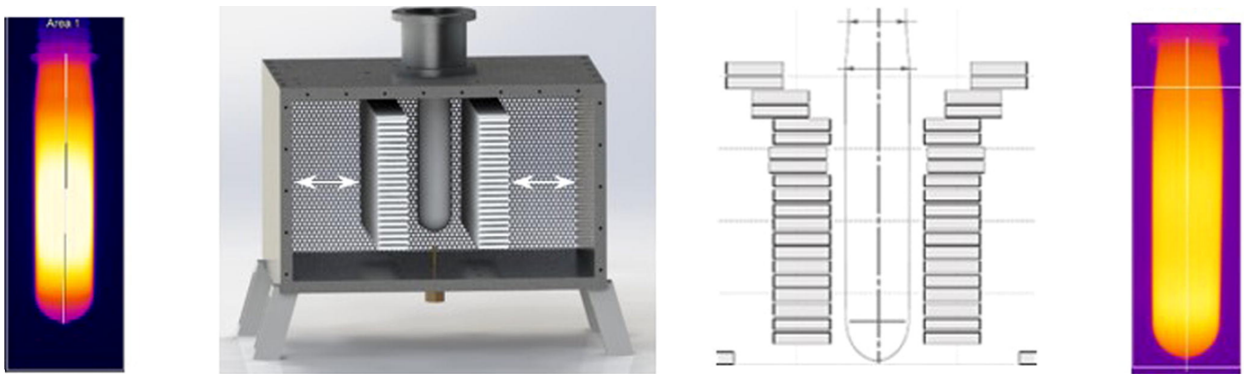


Figure 2: Schematic representation of the microwave heating optimization process for PET preforms. The default temperature profile (left) shows uneven heating typically observed without slabs modifications. A custom-designed microwave cavity with adjustable dielectric slabs (center-left) enables fine-tuning of the electromagnetic field distribution. The optimized slab con-figuration (center-right) is tailored to the preform geometry, leading to a significantly im-proved and more uniform temperature profile (right). . Adapted from García-Baños et al. [18].

in width, 190 mm in length, and 5 mm in height.

The complete system was modeled and simulated using Ansys HFSS, a high-frequency electromagnetic simulation platform, to optimize design parameters and validate heating effectiveness. A key innovation in the system is the use of dielectric slabs as near-field focusing lenses [19], which allow for precise manipulation of electromagnetic waves through reflection, refraction, and diffraction [18].

2.2 Generalization and Model Fusion Methodology

In machine learning, the challenge of model generalization extends beyond simply performing well on unseen data. A regression model that generalizes effectively can extrapolate to new, related datasets. A more advanced challenge involves merging multiple locally generalized models into a single, globally generalized model—a process known as model fusion. By combining diverse models, the overall predictive performance can be enhanced, as each model contributes unique strengths to the final output.

Several model fusion techniques exist [20], including:

- **Voting:** Typically used in classification tasks, this method aggregates predictions from multiple models and selects the majority outcome.
- **Averaging:** Applicable to both regression and classification, it smooths predictions and helps reduce overfitting.
- **Stacking:** This method involves training a second-level model on the outputs (predictions) of several base models to form a meta-learner.

We selected stacking over simple voting or averaging ensembles because stacking trains a meta-learner to combine the outputs of base models in a non-linear fashion, often outperforming fixed combination rules—especially in regression contexts [20]. Voting (or averaging) only computes the mean or majority output and cannot learn how to weight or combine predictions in task-specific ways.

2.3 Data Collection and Predictor Model Training

Our methodology for training and merging pretrained predictor models begins with selecting three distinct variations of the target variable, representing low, medium, and high values. Initial data collection is conducted using Design of Experiments (DOE) to reduce the dimensionality of the input space, particularly focusing on the slab positions. Among various DOE strategies, we employed Latin Hypercube Sampling (LHS) [21], which is known for effectively covering large parameter spaces with minimal experimental runs. As previously described, the microwave heating system was simulated using Ansys HFSS to generate the training data. The input features to the predictor model include:

- Slab positions along the x-axis (continuous variables),
- Preform geometrical attributes such as length, weight, and neck dimensions (when applicable),
- Material-specific properties such as heat capacity (used in Case Study 1).

Note: We utilized a 2D axisymmetric simulation to reduce computational complexity, while still capturing the full axial thermal behavior of PET preforms. Due to the nearly rotationally symmetric geometry, a full 3D simulation would yield essentially the same temperature results as the 2D model, but at much higher cost.

In both case studies, the output (target) is the spatial temperature field, represented as a set of continuous temperature values (in °C) at 32 discrete surface points along the PET preform. This setup defines a regression task in which each model predicts fine-grained temperature distributions based on configuration inputs.

For each case, an initial model was trained using data from the first variation and validated on a separate unseen dataset to verify accuracy. The same model architecture and pretrained weights were then fine-tuned for the remaining two variations, ensuring consistency across training phases.

Model performance was evaluated using standard regression metrics, including:

- Root Mean Squared Error (RMSE),
- Mean Absolute Error (MAE),
- Coefficient of Determination (R^2).

2.4 Fusion Implementation and Neural Network Architecture

Figure 3 illustrates the process of model fusion, beginning with the individual training of predictor models for each variation. The subsequent step involves *experience extraction*, wherein a new Design of Experiments (DOE) setup is constructed and each trained model is tasked with making predictions based on this fresh experimental design.

These predicted outputs are appropriately scaled, and corresponding target variables (associated with each preform variation) are integrated into the dataset. This fusion process merges information extracted from each pretrained model and incorporates variation-specific characteristics to build a more generalizable predictor. The goal is to extend the size and diversity of the initial dataset by synthesizing new samples using model predictions, effectively augmenting the training data.

A new, generalized predictor model is then trained on this merged dataset. By leveraging the fused information, the model gains a broader understanding of the input–output relationships

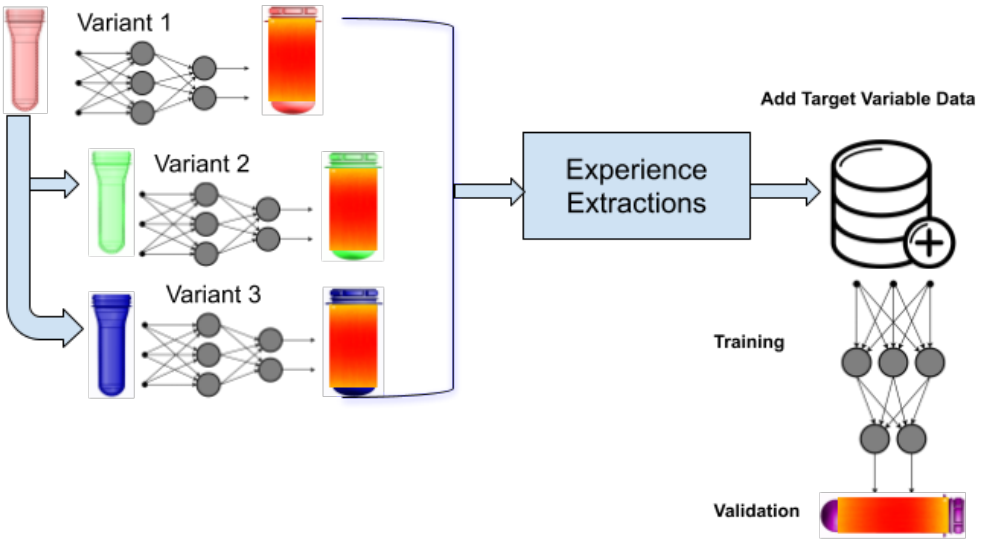


Figure 3: Model fusion workflow combining outputs from variant-specific models into a unified generalized predictor.

Table 1: Evaluation metrics for standard MLP and MLP with Skip Connection models

Metric	Standard MLP	MLP with Skip Connection	Improvement
RMSE	0.185	0.052	↓ 72%
MAE	0.148	0.039	↓ 74%
R^2	0.91	0.98	↑ 7.7%

and demonstrates improved prediction accuracy and generalization performance across unseen configurations.

To evaluate the effectiveness of this fusion-based learning strategy, the final model is tested on a preform variation that was not included in any prior training phase. In this context, "experience" refers to the predictive knowledge embedded within each pre-trained model. This knowledge is utilized to simulate outcomes for new scenarios, which are then compiled into an expanded training dataset. The outputs are rescaled and coupled with their corresponding target labels, forming a comprehensive dataset enriched with structural and material variability.

Input parameters used in all predictors—including slab positions, preform geometries, and material properties—were obtained from Ansys HFSS simulations and verified against known manufacturer specifications. All data generation and transformation steps were automated using scripting and batch processing to ensure reproducibility and scalability.

To develop the predictor model, we compared two neural network architectures, shown in Figures 4a and 4b, to determine the more efficient option. Both networks shared identical input and output configurations, the same number of hidden layers, and identical activation functions.

The first architecture is a standard Multilayer Perceptron (MLP), while the second incorporates *skip connections*. These connections, also known as *residual connections*, are designed to address challenges such as the vanishing gradient problem. They enable more efficient learning by allowing information to bypass intermediate layers and directly propagate forward [22].

Table 1 presents the evaluation metrics—Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2)—on the test dataset. The results show that the MLP with Skip Connection consistently outperforms the standard MLP, achieving lower prediction errors and higher predictive accuracy across all metrics.

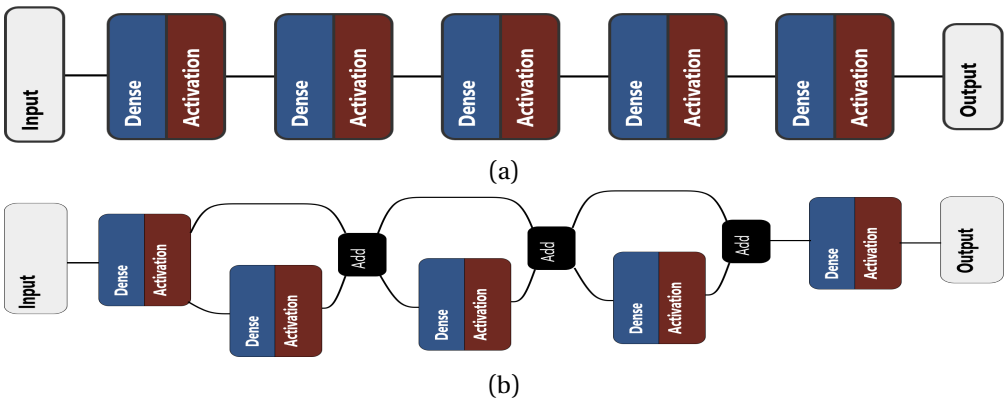


Figure 4: Comparison of MLP architectures: (a) standard and (b) with skip connections ("Add") for improved gradient flow and stability.

3 Case Studies for Model Validation

To evaluate the robustness and generalization capability of the proposed approach, we conducted two case studies: one focusing on variations in material characteristics specifically, heat capacity differences between virgin and recycled PET—and the other on geometrical variations in PET preforms. These case studies demonstrate how the model adapts to both material-related and shape-related differences commonly encountered in production environments.

3.1 Case Study 1: Material Characteristics of PET

This case examines model generalization with respect to PET material variations. While virgin PET is standard in preform production, environmental concerns have increased the use of recycled PET (rPET). Although rPET aims to mimic virgin PET properties, the recycling process can introduce impurities, poor sorting, and thermal degradation [23], leading to structural inconsistencies.

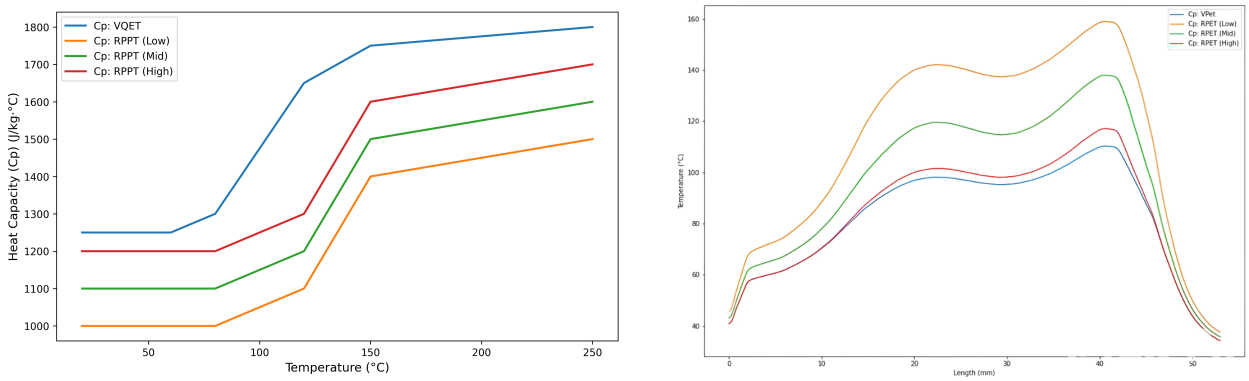
One key property affected is heat capacity—the ability to absorb and retain heat. Virgin PET typically has higher heat capacity due to fewer structural defects, while rPET often shows reduced values due to crystallization disruption and contamination.

Since rPET data are limited, we modeled three plausible heat capacity variations as temperature-dependent functions and compared them with a reference virgin PET curve. These models, shown in Figure 5, reveal that variations in heat capacity significantly impact thermal distribution, influencing heating model accuracy.

3.2 Case Study 2: Geometrical Variations of Preforms

This case investigates the model's generalization to varying preform geometries—critical in industrial settings where preforms differ in size, weight, and design.

Four representative geometries were selected, varying in length, wall thickness, weight, and curvature. These were used to test the generalized model's robustness, particularly its ability to handle unseen shapes without retraining. The results demonstrate the model's flexibility and accuracy across diverse preform designs.



(a) (b)

Figure 5: (a) Heat capacity vs. temperature for virgin and recycled PET. (b) Resulting temperature profiles along preform length under different heat capacity assumptions.

Table 2: Heat capacity and temperature array definitions for each dataset category used in Case Study 1

Heat Capacity Category	Heat Capacity Array J/kg°C	Temperature Array °C	Dataset Size # of signals
Low Cp	[1000, 1050, 1100, 1350, 1450]	[80, 100, 120, 150, 250]	550
Mid Cp	[1100, 1150, 1200, 1500, 1600]	[80, 100, 120, 150, 250]	450
High Cp	[1250, 1300, 1650, 1750, 1800]	[80, 100, 120, 150, 250]	450

3.3 Adaptation to Material and Geometry-Specific Characteristics

This section describes how the proposed approach was validated and fine-tuned for the material and geometrical differences introduced in the previous case studies.

All training and validation datasets used in this study were generated entirely through high-fidelity electromagnetic and thermal simulations in Ansys HFSS. Each simulation run was automated using parametric scripting, enabling efficient, reproducible data generation at scale. This simulation-driven approach allowed us to compile datasets of over 6,000 labeled samples while avoiding the time and cost of physical experiments.

Case Study 1: Adaptation to Material Characteristics. The first case addresses adaptation to variations in PET heat capacity. A base predictor model was initially trained using a mid-range heat capacity dataset comprising 550 samples. The model was subsequently fine-tuned using 450 additional samples from datasets representing low and high heat capacity conditions, respectively.

Each of these three models was trained using different slab position settings to predict temperature values at 32 predefined spatial locations along the PET preform surface.

To assess generalization performance, each model was evaluated on an unseen test dataset. The heat capacity values used to define the low, medium, and high material categories are summarized in Table 2.

After confirming high accuracy across the locally trained models, a new Design of Experiments (DOE) was constructed, generating 2,000 synthetic data points for each model. Each model was then used to predict outcomes on this DOE, resulting in a merged dataset of 6,000 samples. This unified dataset included predicted temperature distributions, associated heat capacity values, and relevant input features such as slab positions.

Figure 6: Training Loss and validation Accuracy for Case 1

A global predictor model was subsequently trained using this enriched dataset, as illustrated in Figure 6. For benchmarking purposes, the global model was compared to a baseline model trained from scratch using a combined real dataset of 1,950 samples—comprising 625 samples from the low, 700 from the mid, and 625 from the high heat capacity categories.

To validate performance, the models were tested on a new, previously unseen heat capacity profile not included in any training phase.

Case Study 2: Adaptation to Geometrical Variations. In the second case study, we evaluated the generalization capabilities of the model with respect to PET preform geometry. Three preform sizes—small, medium, and large—were used for model training and fine-tuning.

Following the same DOE-based experience extraction process described in Case Study 1, a new synthetic dataset comprising 6,000 samples was compiled. Each sample included inputs such as slab positions and critical geometrical attributes (e.g., weight, neck length, and wall thickness).

A global model was then trained on this dataset to generalize predictions across a wide range of preform geometries. For validation, the model was tested on a preform geometry not present in any of the training datasets, thereby assessing its ability to extrapolate across shape-based variability as shown in Figure 7.

4 Conclusions and Future Work

This paper presented a data-efficient generalization technique for regression tasks, applied to temperature prediction in microwave preheating of PET preforms before blow molding. By combining fine-tuning and model fusion, the approach achieved accurate predictions across diverse material and geometrical variations using significantly fewer samples than traditional methods. Results confirm strong generalization to unseen variants, offering a scalable solution for data-limited industrial applications. This method through integration of transfer learning and model fusion is well-suited for physical modeling tasks where simulation data are costly or system variability is high. Future work will focus on handling dynamic material and environmental variations, exploring adaptive, real-time updates, and advancing fusion strategies. A key limitation is the need to fine-tune multiple models; we aim to develop architectures capable of generalizing across variants using a single training pass on a unified dataset, improving scalability. While the dataset is proprietary, it was generated via Ansys HFSS simulations using Latin Hypercube Sampling. Each variant-specific dataset (450–550 samples) was fine-tuned independently, then

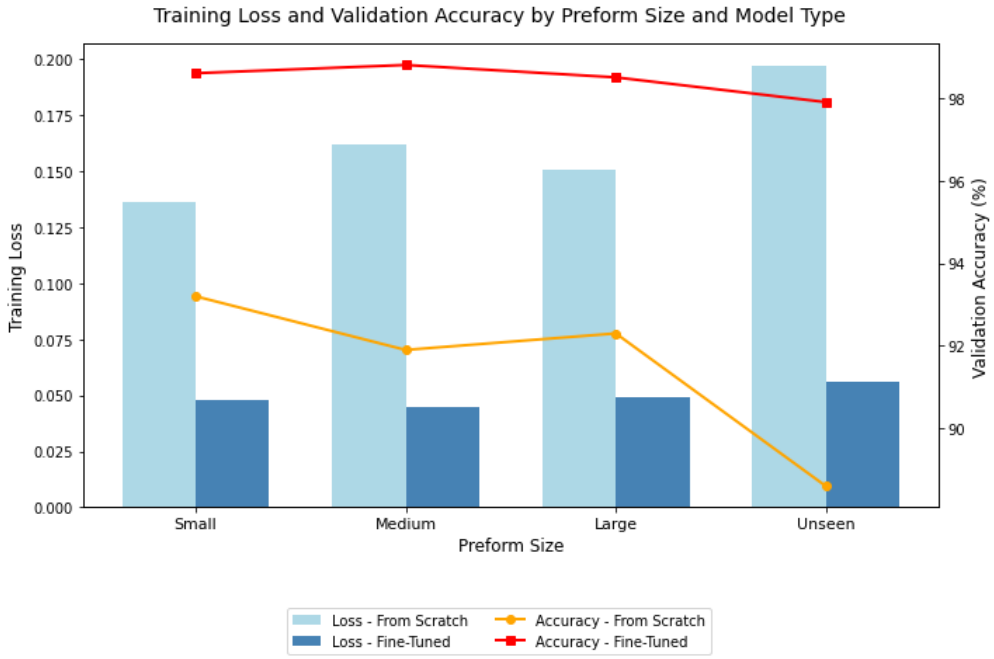


Figure 7: Training Loss and validation Accuracy for Case 2

merged via model outputs for fusion. Models were built in TensorFlow, trained on a workstation with an RTX 3080 GPU. Although the dataset cannot be shared, pseudo-code and synthetic data will be released in the future. Interested researchers may contact the corresponding author via email for further information.

Acknowledgments

The authors wish to thank Thomas Albrecht and Guenter Winkler for their support, fruitful discussions, and useful advices.

Disclosure of Interests

The first author is pursuing a PhD at Deggendorf Institute of Technology in collaboration with Krones AG, which funded the research presented in this article. The second author, a faculty member at Deggendorf Institute of Technology, contributed in an academic supervisory capacity. The authors declare that they have no other competing interests.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (GPT-4) in order to: assist with language editing and LaTeX formatting. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] P. Wawrzyniak, W. Karaszewski, A literature survey of the influence of preform reheating and stretch blow molding with hot mold process parameters on the properties of PET containers part ii, Polimery 65 (2020) 437–448.

[2] Z. Yang, W. Naeem, G. Menary, J. Deng, K. Li, Advanced modelling and optimization of infrared oven in injection stretch blow-moulding for energy saving, IFAC Proceedings Volumes 47 (2014) 766–771.

[3] Y.-M. Luo, L. Chevalier, T. Nguyen, Optimization of the temperature profile of PET preform via a 3d modelling of the infrared heating and ventilation, Materials Research Proceedings 41 (2024).

[4] S. Monteix, F. Schmidt, Y. Le Maoult, G. Denis, M. Vigny, Recent issues in preform radiative heating modelling, in: International Conference of Polymer Processing Society, 2001, pp. 1–6.

[5] W. Liu, H. Ouyang, Q. Liu, S. Cai, C. Wang, J. Xie, W. Hu, Image recognition for garbage classification based on transfer learning and model fusion, Mathematical Problems in Engineering (2022) 1–12.

[6] J. Zhou, Z. Li, W. Zhi, B. Liang, D. Moses, L. Dawes, Using convolutional neural networks and transfer learning for bone age classification, in: 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA), IEEE, Sydney, 2017, pp. 1–6.

[7] M. Ghazi, B. Yanikoglu, E. Aptoula, Plant identification using deep neural networks via optimization of transfer learning parameters, Neurocomputing 235 (2017) 228–235.

[8] S. Chakraborty, R. Mondal, P. Singh, R. Sarkar, D. Bhattacharjee, Transfer learning with fine tuning for human action recognition from still images, Multimedia Tools and Applications 80 (2021) 20547–20578.

[9] O. Korzh, M. Joaristi, E. Serra, Convolutional neural network ensemble fine-tuning for extended transfer learning, in: BigData 2018, 7th International Congress, Held as Part of the Services Conference Federation (SCF 2018), Springer, Seattle, 2018, pp. 110–123.

[10] H. Whitney, H. Li, Y. Ji, P. Liu, M. Giger, Comparison of breast MRI tumor classification using human-engineered radiomics, transfer learning from deep convolutional neural networks, and fusion methods, Proceedings of the IEEE 108 (2019) 163–177.

[11] R. Geyer, L. Corinzia, V. Wegmayr, Transfer learning by adaptive merging of multiple models, in: International Conference on Medical Imaging with Deep Learning, PMLR, 2019, pp. 185–196.

[12] J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, I. Gurevych, Adapterfusion: Non-destructive task composition for transfer learning, arXiv preprint arXiv:2005.00247 (2020).

[13] W. Ge, Y. Yu, Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1086–1095.

[14] Y. Zhai, Y. Huang, Y. Xu, J. Gan, H. Cao, W. Deng, R. Labati, V. Piuri, F. Scotti, Asian female facial beauty prediction using deep neural networks via transfer learning and multi-channel feature fusion, IEEE Access 8 (2020) 56892–56907.

[15] P. Hsieh, Intelligent temperature control of a stretch blow molding machine using deep reinforcement learning, Processes 11 (2023) 1872.

[16] N. Zhai, X. Zhou, Temperature prediction of heating furnace based on deep transfer learning, Sensors 20 (2020) 4676.

[17] P. Di Barba, F. Dughiero, M. Forzan, D. Lowther, A. Marconi, M. Mognaschi, J. Sykulski, Neural metamodels and transfer learning for induction heating processes (TEAM 36 problem), International Journal of Applied Electromagnetics and Mechanics 73 (2023) 389–398.

[18] B. García-Baños, P. Plaza-Gonzalez, J. Sánchez, S. Steger, A. Feigl, F. Penaranda-Foix,

J. Catalá-Civera, Focusing dielectric slabs for the optimization of heating patterns in single mode microwave applicators, Applied Thermal Engineering 201 (2022) 117845.
[19] J. Baker-Jarvis, S. Kim, The interaction of radio-frequency fields with dielectric materials at macroscopic to mesoscopic scales, Journal of Research of the National Institute of Standards and Technology 117 (2012) 1.
[20] X. Huang, L. Zhang, Comparison of vector stacking, multi-SVMs fuzzy output, and multi-SVMs voting methods for multiscale VHR urban mapping, IEEE Geoscience and Remote Sensing Letters 7 (2009) 261–265.
[21] D. Huntington, C. Lyrintzis, Improvements to and limitations of latin hypercube sampling, Probabilistic Engineering Mechanics 13 (1998) 245–253.
[22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, arXiv preprint arXiv:1512.03385 (2015).
[23] S. Venkatachalam, S. Nayak, J. Labde, P. Gharal, K. Rao, A. Kelkar, Degradation and recyclability of poly(ethylene terephthalate), in: InTech, Rijeka, Croatia, 2012, pp. 75–98.

Teacher-Student Guided Inverse Modeling for Steel Final Hardness Estimation

Ahmad Alsheikh[1,2] ahmad.alsheikh@krones.com
Andreas Fischer[2] andreas.fischer@th-deg.de

[1] KRONES AG, Böhmerwaldstr. 5, 93073 Neutraubling, Germany
[2] Deggendorf Institute of Technology, Dieter-Görlitz-Platz 1, 94469 Deggendorf, Germany

Predicting the final hardness of steel after heat treatment is a challenging regression task due to the many-to-one nature of the process—different combinations of input parameters (such as temperature, duration, and chemical composition) can result in the same hardness value. This ambiguity makes the inverse problem—estimating input parameters from a desired hardness—particularly difficult. In this work, we propose a novel solution using a Teacher-Student learning framework. First, a forward model (Teacher) is trained to predict final hardness from 13 metallurgical input features. Then, a backward model (Student) is trained to infer plausible input configurations from a target hardness value. The Student is optimized by leveraging feedback from the Teacher in an iterative, supervised loop. We evaluate our method on a publicly available tempered steel dataset and compare it against baseline regression and reinforcement learning models. Results show that our Teacher-Student framework not only achieves higher inverse prediction accuracy but also requires significantly less computational time, demonstrating its effectiveness and efficiency for inverse process modeling in materials science.

Keywords
Inverse Prediction, Teacher-Student Learning, Heat Treatment, Steel Hardness, Many-to-One Mapping

1 Introduction

Heat treatment is a cornerstone in materials engineering for modifying the mechanical properties of metals, with hardness being a critical target metric in structural and industrial applications. Processes such as tempering, quenching, and annealing involve multiple input parameters including temperature, treatment time, and chemical composition which interact in nonlinear and often material-specific ways. The resulting complexity makes it difficult to predict the process conditions required to achieve a desired final hardness.

A particularly challenging aspect of this problem is its many-to-one nature: Multiple combinations of input variables can yield the same hardness result. This poses a fundamental obstacle to the inverse prediction, where the goal is to infer the set of input conditions that will produce a specified output. Mathematically, many-to-one mappings are non-invertible. This ambiguity complicates the application of traditional regression techniques in inverse modeling. Addressing this challenge requires modeling strategies that can incorporate solution multiplicity, enforce output consistency, and remain robust to the inherent uncertainty of inverse mapping.

This ambiguity complicates the application of traditional regression techniques in inverse modeling, as such methods typically assume a one-to-one correspondence between inputs and outputs. Moreover, standard optimization methods often struggle to converge reliably when

faced with non-invertible mappings, as the error surface can contain multiple local minima or flat regions corresponding to different valid inputs. Addressing this challenge requires modeling strategies that can incorporate solution multiplicity, enforce output consistency, and remain robust to the inherent uncertainty of inverse mapping.

The purpose of this study is to explore and address the inverse problem of predicting process parameters for steel heat treatment under the constraints of many-to-one mappings. By examining this problem in depth, we aim to better understand the limitations of existing approaches and identify avenues for constructing more ro-bust, generalizable inverse models within materials informatics.

The remainder of this paper is organized as follows. Section 2 reviews prior research related to non-invertible mappings and inverse modeling techniques in machine learning. Section 3 presents our proposed framework. Section 4 outlines the experimental setup and results. Section 5 concludes with key findings and potential future directions.

2 Related Work

In this section, we review recent advancements addressing the challenges of One-to-Many and Many-to-One mappings in machine learning, specifically in regression tasks. These mappings often introduce ambiguity due to their non-invertible nature, which complicates prediction, generalization, and interpretation. To address these challenges, researchers have explored a variety of modeling techniques, optimization strategies, and auxiliary frameworks.

Grollman and Jenkins [1] proposed a multi-map regression approach to handle perceptual aliasing in robotic controllers. Their method employs sparse online learning to resolve ambiguities introduced by overlapping many-to-one mappings, demonstrating its effectiveness in robotics applications. Courts and Kvinge [2] introduced bundle networks, a framework that uses generative modeling and fiber bundles to disentangle ambiguities in classification and regression tasks by generating local trivializations.

In robotics, Singh et al. [3] used regression-based kinematic modeling to optimize gait trajectories for biped robots, incorporating auxiliary constraints to improve the handling of many-to-one mappings. Valdés and Tchagang [4] tackled inverse mappings in simulation-based models by combining deterministic surrogate models with machine learning approaches, highlighting the strength of mixed strategies in regression.

Yang et al. [5] used recurrent neural networks to model therapy decision making in metastatic breast cancer, addressing ambiguity through hierarchical regression and encoder-decoder architectures. Chen and Zhu [6] proposed a guided deep learning algorithm for structural surface design, showing that incorporating multiple loss functions can help disambiguate many-to-one regression outputs. Kreuzig et al. [7] developed DistanceNet, which combines recurrent convolutional neural networks with ordinal regression to estimate traveled distance from monocular images, effectively reducing mapping ambiguity in visual tasks.

In the medical imaging domain, Yurt et al. [8] introduced mustGAN, a multi-stream generative adversarial network for MR image synthesis. Their model addresses feature-level ambiguities through adversarial training. Wang et al. [9] presented M2ORT, a transformer-based framework for spatial transcriptomics prediction, demonstrating how auxiliary data and novel architectural designs can improve the modeling of non-invertible mappings.

Zhang et al. [10] proposed a local-to-global cost aggregation method for semantic correspondence using a Teacher-Student framework. This approach adapts learned representations to mitigate ambiguities in many-to-one feature matching tasks. Lastly, Neupane et al. [11] surveyed techniques for 3D human pose estimation, focusing on methods that apply auxiliary constraints and deep learning models to resolve depth ambiguities inherent in coordinate regression.

3 Methodology: Teacher-Student for Inverse Prediction

Solving inverse problems in many-to-one mappings is especially challenging due to the inherent ambiguity: a single output can correspond to multiple valid input configurations. Traditional regression methods struggle with this non-invertibility, often resulting in poor generalization and conflicting gradients during optimization. To overcome these limitations, we propose a Teacher-Student learning framework for robust inverse prediction.

As illustrated in Figure 1, the framework consists of two Multi-Layer Perceptron (MLP) models:

- A **Teacher model**, trained on the forward (many-to-one) task, maps a set of 13 input features—including tempering time, temperature, and elemental composition—to the final hardness value (HRC).
- A **Student model**, trained on the inverse (one-to-many) task, predicts plausible input configurations given a target hardness.
- Both models use three dense layers with ELU activations and residual connections.

The training process proceeds in two phases. First, the Teacher model is trained to accurately learn the forward mapping using supervised learning. Once trained, it remains fixed and acts as a reference model for supervising the Student.

The Student model receives randomly sampled target hardness values as inputs and predicts the corresponding input parameters. These predicted inputs are then passed through the Teacher model, which outputs a predicted hardness. This predicted value is compared to the original target, and a loss is computed. The loss is then back-propagated to update the Student model's weights. This process continues iteratively, allowing the Student to learn input patterns that are functionally consistent with the desired hardness values.

Importantly, because of the many-to-one nature of the problem, the Student is not trained to replicate original dataset entries. Instead, it learns to generate valid, functionally equivalent inputs—those that the Teacher model accepts as leading to the correct hardness. This design enables the system to embrace ambiguity while still producing accurate and interpretable predictions for inverse process modeling.

4 Experimental Setup and Results

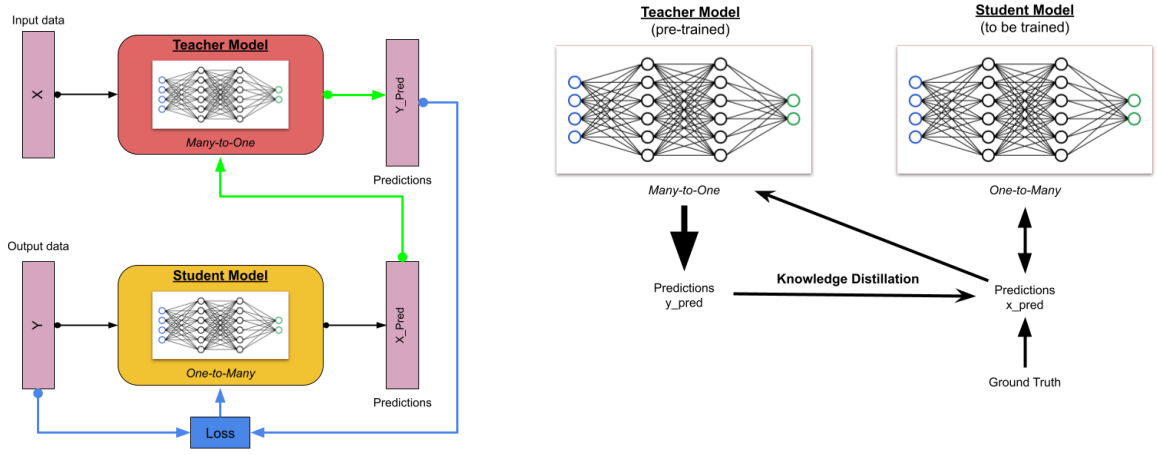
4.1 Dataset Overview

To evaluate our proposed framework, we used a publicly available dataset [12] containing steel samples subjected to various heat treatment conditions. Each data point consists of 13 input features:

- **Process parameters:** tempering time and temperature
- **Chemical composition:** elemental percentages of C, Mn, P, S, Si, Ni, Cr, Mo, V, Al, and Cu

The target variable is the final hardness of the steel sample, measured in Rockwell Hardness (HRC) after quenching and tempering.

Figure 2a illustrates the relationship between tempering temperature and final hardness (HRC) for varying tempering times. As temperature increases, final hardness consistently decreases across all time intervals, confirming a strong inverse correlation. Longer tempering durations (e.g., 100,000 s) result in lower hardness compared to shorter durations (e.g., 20,000 s), highlighting the combined effect of thermal exposure and time on material softening. Elements like Mn



(a) (b)

Figure 1: Overview of the Teacher-Student framework for inverse hardness prediction. (a) The Teacher model is trained to map process inputs (X) to final hardness (Y). The Student model learns to predict valid input configurations (X_{pred}) from target hardness values. Predictions are evaluated by the fixed Teacher model, and the loss is used to iteratively update the Student. (b) A schematic view highlighting the knowledge distillation process between Teacher and Student models.

and P show positive correlation, while others (e.g., C, Cr, Si) exhibit weaker or nonlinear relationships.

4.2 Observing the Many-to-One Mapping

Exploratory analysis revealed that the dataset exhibits a many-to-one mapping: multiple distinct combinations of input parameters lead to the same hardness value. This is visualized in Figure 2b, where overlapping input regions map to identical target values. Such non-invertibility complicates inverse prediction, as standard regression models struggle to differentiate between equally valid input solutions.

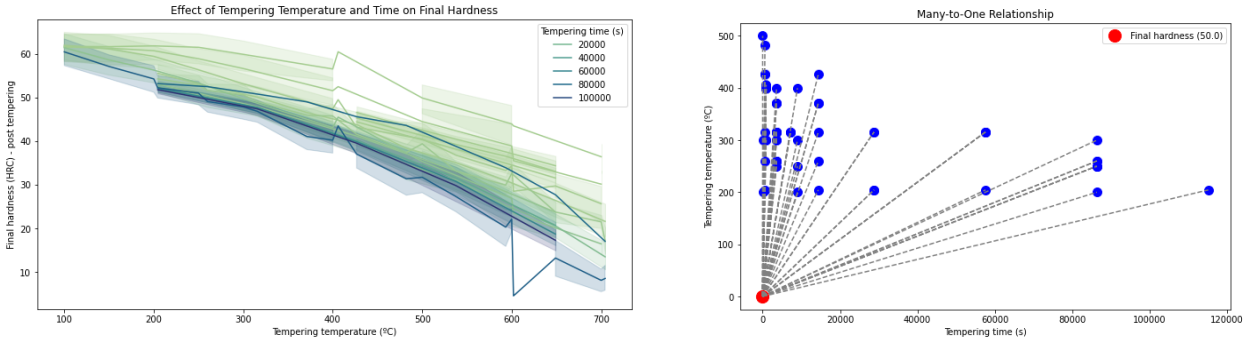
4.3 Baseline Inverse Modeling Attempts

To assess the complexity of the inverse prediction task, we first trained two conventional regression models to estimate process parameters from a given target hardness value:

- A Random Forest Regressor
- A Multi-Layer Perceptron (MLP)

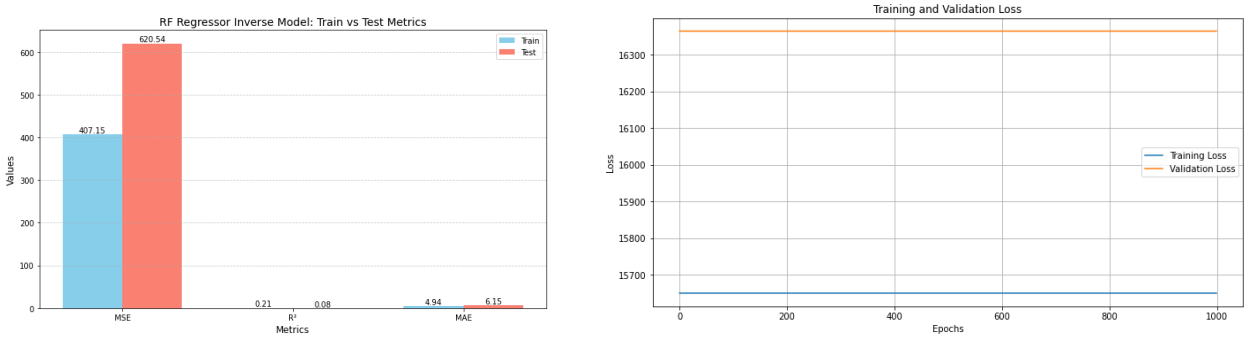
Both models were configured to take the final hardness (HRC) as input and predict the full set of 13 output variables, including tempering parameters and elemental composition. Hyperparameters were optimized using random search.

Despite optimization, the results demonstrate the limitations of using off-the-shelf regression models for one-to-many problems. The random forest regressor yielded high prediction error, with an MSE of 620.54 on the test set and a low R^2 score of 0.08, as shown in Figure 3a. This large discrepancy between predicted and true values highlights the model's inability to generalize in the presence of multiple valid input solutions for the same output.



(a) (b)

Figure 2: Exploratory data analysis. (a) Relationship between tempering temperature, time, and final hardness (HRC). Higher temperatures and longer durations result in lower hardness, illustrating a clear inverse trend. (b) Multiple input configurations leading to the same hardness value, highlighting the many-to-one nature of the inverse problem.



(a) (b)

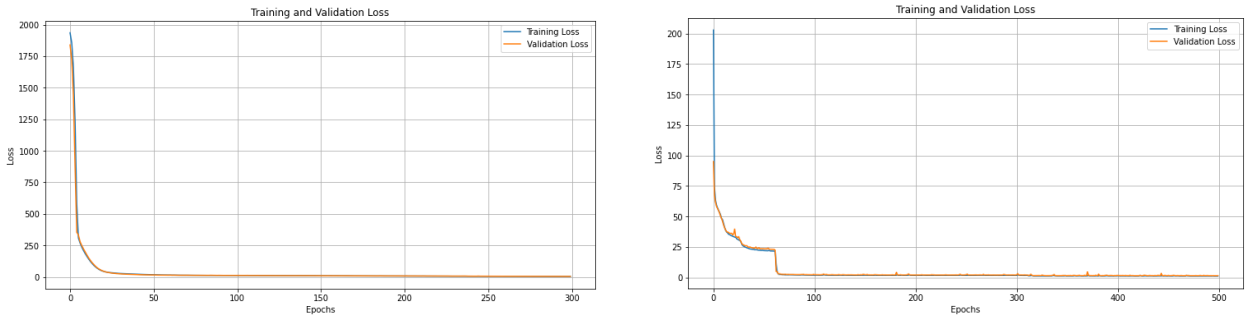
Figure 3: Baseline model performance. (a) The Random Forest inverse model shows high test error and low generalization. (b) The MLP baseline failed to converge, with persistent high loss across training and validation. These results highlight the difficulty of solving many-to-one inverse prediction problems with standard models.

The MLP model also performed poorly. As seen in Figure 3b, both training and validation loss remained flat across 1000 epochs, with no significant improvement. This indicates that the model failed to capture any meaningful inverse mapping structure. The results highlight that the MLP struggled to achieve meaningful improvements, with persistent high loss across both training and validation, further reinforcing the complexity of the one-to-many prediction problem.

Baseline models were evaluated by comparing predicted configurations to the corresponding ground truth inputs in the dataset. While this does not account for the one-to-many nature of the problem, it provides a standardized measure of model accuracy. We acknowledge that a more meaningful comparison would involve verifying whether the predicted inputs yield the correct output, which is how our Teacher-Student and reinforcement learning (RL) models are evaluated.

4.4 Teacher-Student Framework Performance

Building on the limitations observed in baseline inverse modeling, we next implemented the forward Teacher model, which serves as the foundation of our Teacher-Student framework. The Teacher is trained to learn the many-to-one mapping from the 13 input features—tempering



(a) (a)

Figure 4: Training and validation loss curves. (a) The forward (Teacher) MLP model converges quickly within the first 50 epochs, with both losses approaching zero, demonstrating strong generalization and suitability as a reference for inverse training. (b) The inverse (Student) model effectively learns the mapping over 500 epochs, with both training and validation losses sharply decreasing and stabilizing.

time, temperature, and chemical composition—to the target variable: final hardness (HRC).

We used a Multi-Layer Perceptron (MLP) architecture for the Teacher model. The training process was carried out using standard supervised learning, with mean squared error (MSE) as the loss function.

As shown in Figure 4a, the model converged rapidly and effectively. Both training and validation loss decreased sharply within the first few epochs and stabilized at low values, demonstrating excellent generalization and predictive performance. This strong forward model is critical, as it serves as a consistent, differentiable reference for training the inverse Student model in the next phase.

Once the Teacher model was trained and fixed, we proceeded to train the Student model to solve the inverse problem: predicting process parameters from a target hardness value. The Student model was trained in a collaborative loop with the Teacher, allowing it to learn valid inverse mappings despite the many-to-one nature of the problem. Each training iteration proceeds as follows:

- A batch of random target hardness values is sampled.
- The Student predicts the corresponding 13 input features, including tempering parameters and elemental composition.
- These predicted inputs are passed through the Teacher model, which outputs predicted hardness values.
- A loss is calculated by comparing the Teacher's output to the original target.
- The loss is backpropagated to update the Student model's weights.

This iterative feedback mechanism continues until convergence. As shown in Figure 4b, the training and validation loss both decrease rapidly and stabilize at low values, indicating that the Student successfully learns to produce inputs that yield accurate hardness predictions when passed through the Teacher.

Evaluation metrics, shown in Figure 5, further confirm the Student's effectiveness. The model achieved high R^2 scores of 0.98 on both training and test sets, with relatively low mean squared error (MSE) and mean absolute error (MAE) values. Importantly, the predicted input values do not need to match dataset entries exactly—due to the many-to-one nature of the mapping—but

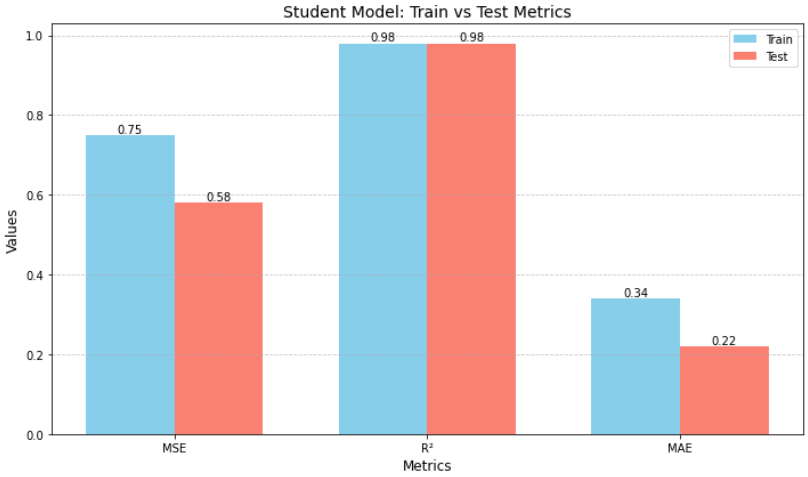


Figure 5: Evaluation metrics for the Student model on training and test data. The model achieved high R^2 values (0.98) and low mean squared and absolute errors, demonstrating accurate and generalizable inverse predictions when supervised by the Teacher model.

only need to generate the correct target output. This flexibility is a key strength of the Teacher-Student framework. It allows the Student model to explore the solution space beyond the training data, while still producing consistent, functionally accurate results.

In contrast to the approaches described previously—such as Bundle Networks [2] or mustGAN [8] our method is designed to be both simple and data-efficient. Rather than employing complex generative models, recurrent structures, or transformers, we adopt a purely feedforward architecture based on MLPs with residual connections and ELU activations. Unlike previous methods that require large datasets and intricate training regimes, our Teacher-Student framework can learn effective inverse mappings from limited data. Additionally, our method avoids probabilistic modeling, employing a deterministic Student model supervised by a fixed Teacher MLP. This architectural simplicity, combined with a training strategy focused on consistency rather than reconstruction, allows our model to produce multiple functionally valid input configurations without resolving all ambiguity. These distinctions make our approach especially well-suited for industrial applications.

4.5 Comparison with Reinforcement Learning Approach

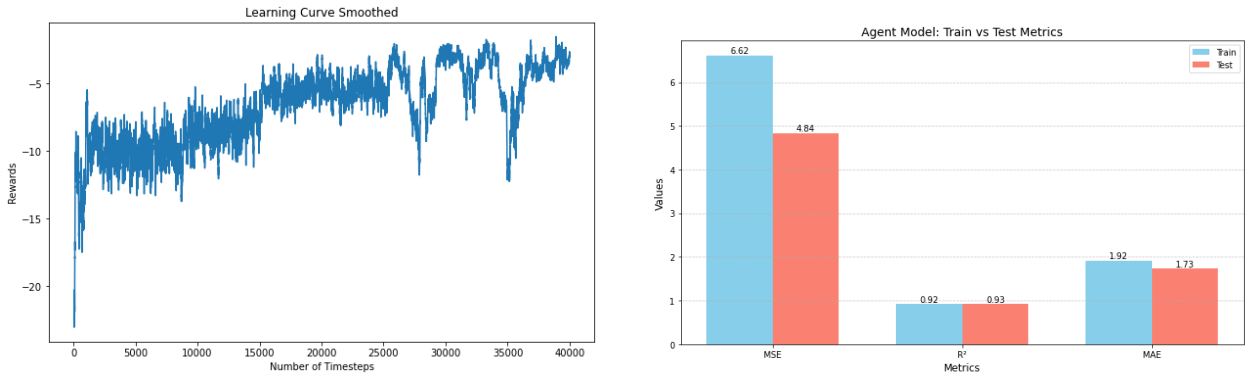
To benchmark the performance of our Teacher-Student approach, we implemented a model-free reinforcement learning (RL) agent using the TD3 (Twin Delayed Deep Deterministic Policy Gradient) algorithm. Like the Student model, the RL agent was tasked with solving the inverse problem: predicting 13 process parameters that would yield a specified target hardness value.

The training setup used the trained Teacher model as the environment. At each timestep:

- The agent receives a target hardness value.
- It predicts a set of 13 input variables (time, temperature, composition).
- These inputs are passed to the Teacher, which returns a predicted hardness.
- A reward is computed as the negative mean squared error (MSE) between the predicted and actual hardness.
- This reward is used to update the agent's policy via TD3.

Table 1: Performance comparison of inverse modeling approaches. The Teacher-Student framework achieves the lowest MSE and fastest convergence, outperforming both traditional models (Random Forest, baseline MLP) and the RL-based TD3 agent in accuracy and efficiency.

Model	MSE	Training Time	Notes
Teacher-Student	0.52	2.3 min	Best overall performance
RL (TD3)	4.84	27.7 min	Slower training, higher error
Random Forest	620.54	∼1.2 min	Poor generalization
MLP (baseline)	15730	∼3.5 min	Failed to converge



(a) (b)

Figure 6: Performance of the reinforcement learning (RL) approach. (a) Smoothed learning curve of the TD3 agent over 40,000 timesteps, showing gradual policy improvement with typical fluctuations. (b) Evaluation metrics indicate that while the RL model achieves R^2 scores above 0.9, its MSE and MAE are significantly higher than those of the Teacher-Student model, highlighting its lower accuracy and efficiency.

The RL agent was trained for 40,000 steps, compared to 15,000 steps for the Student model. As shown in Figure 6a, the smoothed reward curve demonstrates gradual performance improvement, indicating convergence. However, the learning process is noticeably more volatile and prolonged.

In terms of quantitative performance, Figure 6b shows that the RL agent achieved reasonable accuracy, with R^2 values around 0.92–0.93, and moderate error rates (MSE = 4.84, MAE = 1.73 on the test set). Despite these results, the RL agent was clearly outperformed by the Student model, which achieved significantly better accuracy with far less computational cost.

Notably, training the RL agent took 27.7 minutes, while the Teacher-Student system required only 2.3 minutes. These findings highlight the efficiency, stability, and predictive quality of our proposed framework over traditional reinforcement learning approaches for inverse process modeling.

5 Conclusions and Future Work

In this work, we tackled the inverse prediction problem in steel heat treatment, where multiple input configurations can lead to the same final hardness, creating a non-invertible, many-to-one mapping that poses a major challenge for conventional regression methods. To address this, we introduced a Teacher-Student learning framework in which a forward model (Teacher) predicts hardness from process parameters, and an inverse model (Student) learns to infer valid input configurations from target hardness values through iterative supervision.

Our approach was evaluated on a real-world dataset of tempered steel and demonstrated superior performance compared to both standard regression techniques and a reinforcement learning baseline, achieving higher accuracy and significantly lower computational cost. However, while our Student model is capable of generating functionally correct inverse configurations, we did not explicitly enforce or measure diversity in the predicted inputs. Because the loss is computed on the Teacher's output, the Student is not constrained to match any specific input sample from the dataset. In principle, this allows it to converge on multiple valid solutions per target value. However, we acknowledge that without explicit diversity-promoting mechanisms (e.g., entropy regularization, sampling strategies or loss enforcing), the model may still collapse toward a subset of the input space. Quantifying the variance of predicted configurations is an important direction for future work, especially in scenarios requiring solution diversity or process flexibility.

Looking forward, several directions can extend this work. Incorporating uncertainty modeling could allow the Student to express confidence in its predictions. Attention mechanisms or transformer-based architectures may improve both accuracy and interpretability. Applying the framework to other material properties or processes would test its generalization ability, while integration into real-time control systems could enable dynamic process optimization. Finally, combining this data-driven approach with physics-informed constraints offers a path toward more robust and explainable models for complex inverse problems in materials science.

Acknowledgments

The authors wish to thank Thomas Albrecht for his support, fruitful discussions, and useful advice.

Disclosure of Interests

The first author is pursuing a PhD at Deggendorf Institute of Technology in collaboration with Krones AG, which funded the research presented in this article. The second author, a faculty member at Deggendorf Institute of Technology, contributed in an academic supervisory capacity. The authors declare that they have no other competing interests.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (GPT-4) in order to: assist with language editing, LaTeX formatting, and figure caption refinement. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] D. H. Grollman, O. C. Jenkins, Multimap regression for perceptual aliasing in learning finite state machine robot controllers, in: Proc. of the Robotics: Science and Systems Workshop on Regression in Robotics, 2009.

[2] N. Courts, H. Kvinge, Bundle networks: Fiber bundles, local trivializations, and a generative approach to exploring many-to-one maps, arXiv preprint arXiv:2102.11140 (2021).

[3] B. Singh, A. Vijayvargiya, R. Kumar, Kinematic modeling for biped robot gait trajectory using machine learning techniques, Journal of Bionic Engineering 19 (2022) 355–369.

[4] J. J. Valdés, A. B. Tchagang, Deterministic numeric simulation and surrogate models with white and black machine learning methods: A case study on inverse mappings, in: 2020 IEEE Symposium Series on Computational Intelligence (SSCI), 2020, pp. 1616–1623.
[5] Y. Yang, P. A. Fasching, V. Tresp, Predictive modeling of therapy decisions in metastatic breast cancer with recurrent neural networks, in: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2017, pp. 496–501.
[6] Y. Chen, J. Zhu, A novel guided deep learning algorithm for designing structural surfaces, arXiv preprint arXiv:1905.06244 (2019).
[7] R. Kreuzig, M. Ochs, R. Mester, Distancenet: Estimating traveled distance from monocular images using a recurrent convolutional neural network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.
[8] M. Yurt, S. U. H. Dar, K. K. Oguz, mustgan: Multi-stream generative adversarial networks for mr image synthesis, Medical Image Analysis 70 (2021) 102002.
[9] H. Wang, X. Du, S. Ouyang, M2ort: Many-to-one regression transformer for spatial transcriptomics prediction, arXiv preprint arXiv:2403.03209 (2024).
[10] X. Zhang, Z. Fu, Y. Guo, Z. Li, Q. Yu, Local-to-global cost aggregation for semantic correspondence, IEEE Transactions on Image Processing 31 (2022) 4898–4911.
[11] R. B. Neupane, K. Li, T. F. Boka, A survey on deep 3d human pose estimation: Recent advances and challenges, Artificial Intelligence Review (2024). doi:10.1007/s10462-023-10532-9.
[12] Raiipa Technologies, Tempering data for carbon and low alloy steels [data set], https://doi.org/10.34740/KAGGLE/DSV/8468279, 2024. Kaggle.

Model Context Protocol in Manufacturing: A Comprehensive Framework for AI-Driven Industry 4.0?

Michael Banf michael.banf@perelyn.com
Johannes Kuhn johannes.kuhn@perelyn.com

Perelyn GmbH, Reichenbachstraße 31, 80469 München, Germany

The integration of artificial intelligence in manufacturing environments has been hindered by the complexity of connecting disparate systems and the lack of standardized protocols for AI-manufacturing communication. The Model Context Protocol (MCP), introduced by Anthropic in late 2024, presents a transformative solution to these integration challenges. MCP provides a standardized framework for connecting AI applications with external data sources and tools, similar to how USB-C standardized device connectivity. This review examines the Model Context Protocol and its potential applications in manufacturing and industrial production. We provide a brief introduction to MCP's architecture and core concepts. We place a particular emphasis on three key use cases, i.e. predictive maintenance, production planning and control, and process automation, followed by a critical discussion of the literature regarding its application in manufacturing and industrial production contexts. Our analysis reveals that while MCP shows significant promise for addressing longstanding challenges in manufacturing AI integration, empirical evidence remains limited, and most applications are still in early implementation phases.

Keywords
Generative AI • Model Context Protocol • Predictive Maintenance • Planning and Control • Process Automation

1 Introduction

The fourth industrial revolution, commonly known as Industry 4.0, has fundamentally transformed manufacturing through the integration of cyber-physical systems, Internet of Things, and artificial intelligence. However, despite significant technological advances, manufacturing organizations continue to face substantial challenges in integrating AI systems with existing infrastructure. According to recent industry reports, unplanned downtime costs industrial manufacturers an estimated $50 billion annually, while quality issues and production inefficiencies contribute to billions more in losses [1]. The Model Context Protocol (MCP) represents a recent development in the standardization of AI system integration, introduced by Anthropic in November 2024 [2]. As an open standard designed to facilitate secure, bidirectional connections between AI models and external data sources, MCP addresses a critical challenge in the deployment of artificial intelligence within industrial contexts. This review examines the emerging literature on MCP's application in manufacturing, with particular focus on its potential to transform predictive maintenance, production planning and control, and process automation. Our motivation for this brief review stems from the observation that while Industry 4.0 technologies have promised seamless AI integration for years, practical implementation has been hampered by the need for custom integrations between AI models and industrial systems [4]. MCP emerges as a potential

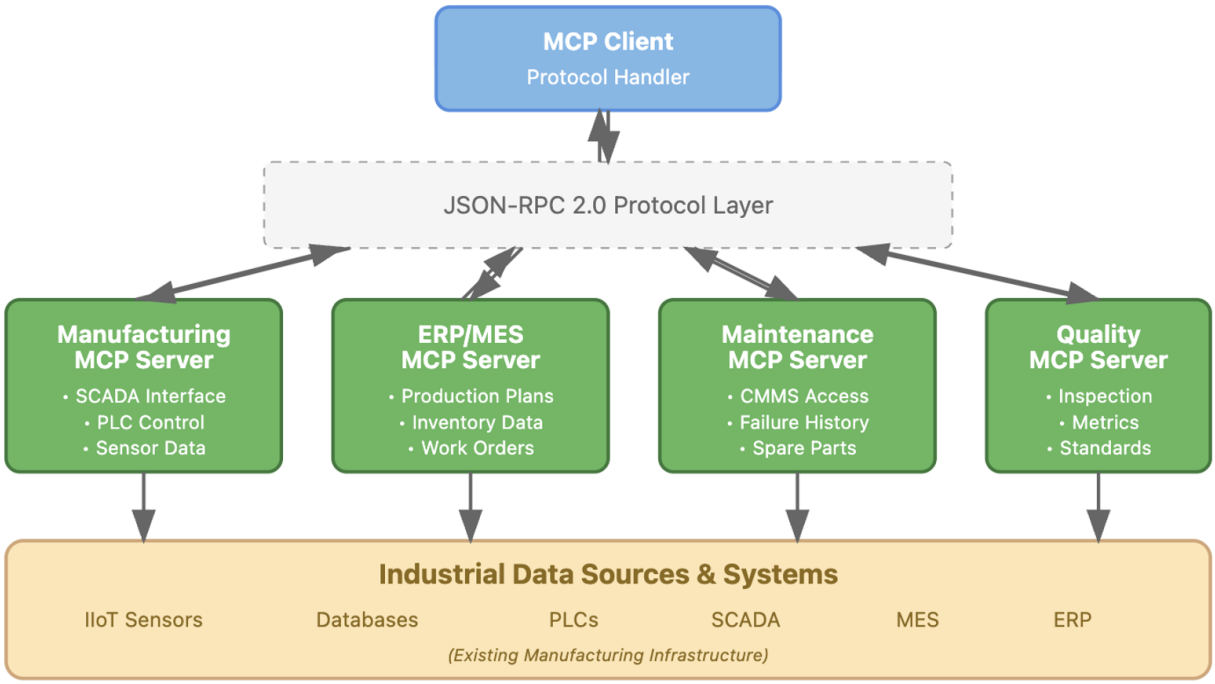


Figure 1: Illustration of an MCP architecture for manufacturing.

solution to this fragmentation, offering a standardized protocol analogous to how USB-C standardized device connectivity [6].

1.1 Architecture of the Model Context Protocol

The Model Context Protocol is a standardized interface that enables AI models, particularly large language models (LLMs), to interact dynamically with external tools, databases, and resources through a consistent protocol [2]. Thereby, it follows a client-server architecture consisting of three main components [7]:

- MCP Hosts/Clients: AI applications (such as Claude Desktop, Cursor, or custom industrial AI systems) that initiate connections
- Protocol Layer: Utilizes JSON-RPC 2.0 for message exchange and communication
- MCP Servers: Lightweight processes that expose specific capabilities and data sources

Further, MCP defines three fundamental primitives that enable AI-system interaction [8]: i) resources, i.e. static or dynamic datasets accessible to AI models; ii) tools, i.e. invokable functions that allow AI models to perform actions; and iii) prompts, i.e. predefined templates that guide AI interactions with resources and tools.

In terms of security, host-mediated security models have been proposed, where permissions are granularly controlled, isolation is maintained between different data sources, and all AI-initiated actions are auditable [5]. This security architecture is particularly relevant for manufacturing environments where unauthorized access to production systems could have severe consequences.

2 Current State of Research: MCP in Manufacturing

The academic literature on MCP in manufacturing is nascent, reflecting the protocol's recent introduction. Silva et al. [3] represent the first academic exploration of MCP specifically within manufacturing contexts. Silva et al. [3] argue that traditional approaches to modeling manufacturing capabilities require considerable manual effort and often result in representations that are not easily accessible to Large Language Models (LLM). Their research proposes MCP as an alternative that allows manufacturing systems to expose functionality through interfaces directly consumable by LLM-based agents. The authors conduct a prototypical evaluation on a laboratory-scale manufacturing system where resource functions were made available via MCP. A general-purpose LLM was then tasked with planning and executing multi-step processes, including constraint handling and resource function invocation. Their preliminary results suggest that MCP can enable flexible industrial automation without relying on explicit semantic models [3]. While specific MCP manufacturing literature is limited, several researchers have identified the need for standardized AI integration protocols in manufacturing. Frank et al. [9] note that traditional Production Planning and Control (PPC) systems struggle with the integration of emerging digital capabilities, particularly in connecting AI models to production data [9]. Rossit et al. [10] emphasize that smart PPC systems require the ability to utilize a wider range of data sources from the production system and to capture and utilize the experience of production planners. These requirements align with MCP's value proposition of replacing fragmented integrations with a unified protocol, and MCP's architecture, with its resources and tools primitives, directly addresses these requirements by providing standardized access to diverse data sources and enabling AI models to invoke production-relevant functions.

2.1 Predictive maintenance

The literature suggests that predictive maintenance represents one of the most promising applications for MCP in manufacturing, although direct empirical studies are yet lacking. The theoretical framework for MCP-enabled predictive maintenance builds on established Industry 4.0 concepts. Recent surveys on predictive maintenance emphasize the critical role of real-time data access from Industrial Internet of Things devices [11]. MCP's resource primitive could standardize how AI models access e.g. real-time sensor data, historical maintenance records and equipment operational parameters. Further, MCP's tool primitive addresses a key limitation identified in traditional predictive maintenance systems, i.e. the gap between prediction and action [12]. MCP could enable AI models to not only predict failures but also adjust production schedules to accommodate maintenance windows, order spare parts based on predicted failure modes, or notify maintenance personnel with context-specific information. However, the integration of MCP with predictive maintenance systems faces several challenges. Security concerns are paramount, as noted by Hou et al. [4]. Manufacturing environments require stringent access controls to prevent unauthorized manipulation of maintenance schedules or equipment parameters.

2.2 Production Planning and Control

Production Planning and Control (PPC) emerges as a critical application area for MCP, addressing longstanding challenges in manufacturing operations management [10]. Traditional PPC systems suffer from inflexibility and inability to adapt to real-time changes [9]. MCP could transform PPC by enabling AI models to e.g. access real-time production status through resource interfaces, or retrieve current work-in-progress levels, machine availability, and material stocks, as well as dynamically adjust production sequences based on emerging constraints. Silva et al.'s

experimental work demonstrates this capability in practice, showing how an LLM successfully planned and executed multi-step manufacturing processes through MCP interfaces [3].

One underexplored aspect in the current literature is how MCP might enhance human-AI collaboration in production planning. To this end, the prompt primitive could encode planning best practices and standard operating procedures, allowing AI systems to suggest actions that align with organizational knowledge and constraints.

2.3 Process Automation

Process automation might represent another major use case for MCP in manufacturing, although the literature remains scarce. Traditional process automation relies on rigid, pre-programmed control logic. MCP enables a paradigm shift towards adaptive automation [14] where AI models may e.g. monitor process variables through resource interfaces, analyze patterns and anomalies in real-time, and invoke control actions and strategies through tool interfaces based on changing conditions. MCP could facilitate more sophisticated optimization strategies in process automation by providing AI models with comprehensive access to energy consumption data, material flow information, and production constraints [13].

3 Discussion & Outlook

The Model Context Protocol emerges as a potentially transformative technology for manufacturing AI integration, offering a standardized approach to connecting AI systems with industrial environments. While the theoretical advantages are compelling, particularly for predictive maintenance, production planning, and process automation, the transition from concept to industrial reality faces substantial challenges. The pioneering work by Silva et al. [3] demonstrates MCP's capability to enable flexible automation without complex semantic modeling, yet this research remains confined to laboratory settings. This highlights a critical gap in our understanding: how MCP performs under the demanding conditions of real manufacturing environments. Industrial settings present unique challenges including extreme latency requirements for real-time control, massive scalability needs when interfacing with thousands of sensors and actuators, and stringent reliability standards where downtime costs can reach millions per hour.

Several crucial aspects remain unexplored in current literature. Performance metrics essential for manufacturing applications, such as response times for safety, critical decisions, bandwidth requirements for sensor streams, and fault tolerance mechanisms, lack comprehensive analysis. Additionally, practical implementation concerns receive insufficient attention. How do manufacturers migrate from established systems? What training investments are required for personnel accustomed to traditional automation paradigms? How does MCP interface with existing industrial protocols like OPC-UA that dominate current factory floors? Safety and security considerations, while partially addressed, require specific examination. Industrial environments demand fail-safe mechanisms that protect workers when AI systems make decisions affecting physical machinery. Compliance with industry standards isn't optional, it's mandatory for deployment. Yet the literature provides limited guidance on validating AI decisions within these regulatory frameworks.

The path forward requires coordinated effort across multiple fronts. Researchers must move beyond laboratory demonstrations to pilot programs in operating factories. Standards bodies need to develop manufacturing-specific guidelines for MCP implementation. Technology vendors should create tools that bridge the gap between MCP's capabilities and industrial requirements. Most critically, manufacturers themselves must engage in this evolution, sharing experiences and establishing best practices. The manufacturing industry stands at an inflection point.

MCP could accelerate AI adoption by removing technical barriers that have historically limited implementation. However, success depends on addressing the practical realities of industrial environments, where reliability, safety, and performance aren't just desirable features but fundamental requirements. As this technology evolves from promise to practice, careful attention to these factors will determine whether MCP becomes a cornerstone of Industry 4.0 or remains an interesting but impractical concept.

References

[1] IndustryWeek. (2022). *Unlocking performance*. Retrieved from https://partners.wsj.com/emerson/unlocking-performance/how-manufacturers-can-achieve-top-quartile-performance/

[2] Anthropic. (2024). *Introducing the Model Context Protocol*. Retrieved from https://www.anthropic.com/news/model-context-protocol

[3] Silva, L. M. V., et al. (2025). *Beyond Formal Semantics for Capabilities and Skills: Model Context Protocol in Manufacturing*. arXiv preprint arXiv:2506.11180.

[4] Hou, X., et al. (2025). *Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions*. arXiv preprint arXiv:2503.23278.

[5] Narajala, V. S., et al. (2025). *Enterprise-Grade Security for the Model Context Protocol (MCP): Frameworks and Mitigation Strategies*. arXiv preprint arXiv:2504.08623.

[6] Gradient Flow. (2025). *Model Context Protocol: What You Need To Know*. Retrieved from https://gradientflow.com/model-context-protocol-what-you-need-to-know/

[7] Model Context Protocol. (2025). *Introduction - Model Context Protocol*. Retrieved from https://modelcontextprotocol.io/introduction

[8] DataCamp. (2025). *Model Context Protocol (MCP): A Guide With Demo Project*. Retrieved from https://www.datacamp.com/tutorial/mcp-model-context-protocol

[9] Frank, A. G., Dalenogare, L. S., & Ayala, N. F. (2020). *Smart production planning and control in the Industry 4.0 context: A systematic literature review*. Computers & Industrial Engineering, 149, 106852.

[10] Rossit, D. A., Tohmé, F., & Frutos, M. (2021). *Designing and developing smart production planning and control systems in the industry 4.0 era: a methodology and case study*. Journal of Intelligent Manufacturing, 32(4), 1-24.

[11] Sharma, A., et al. (2024). *Predictive maintenance in Industry 4.0: a survey of planning models and machine learning techniques*. Peer Journal of Computer Science.

[12] IIoT World. (2023). *Elevating Predictive Maintenance: The Transformative Impact of IIoT*. Retrieved from https://www.iiot-world.com/predictive-analytics/predictive-maintenance/

[13] Shakudo. (2025). *What is MCP (Model Context Protocol) & How Does it Work? Use Cases + Examples*. Retrieved from https://www.shakudo.io/blog/mcp-model-context-protocol

[14] Boomi. (2025). *How to Use Model Context Protocol the Right Way*. Retrieved from https://boomi.com/blog/model-context-protocol-how-to-use/

Artificial Intelligence in SME Production: An Analysis of Adoption Barriers and Strategic Recommendations

Lars Fichtel[1] lars.fichtel@thws.de [a]
Dominik Seuß[1,2] dominik.seuss@thws.de

[a]Corresponding author.

[1] *Center for Artificial Intelligence (CAIRO), University of Applied Sciences Würzburg-Schweinfurt, Würzburg, Germany*
[2] *Fraunhofer Institute for Integrated Circuits IIS, Fraunhofer-Gesellschaft, Erlangen, Germany*

This paper explores why small and medium-sized enterprises (SMEs) in manufacturing struggle to adopt AI technologies at scale. Based on the experience of several industry projects and recent studies, we identify five interconnected barriers and illustrate how they affect production-specific AI applications. The paper concludes with practical strategies tailored to SME constraints, aiming to bridge the gap between AI potential and industrial reality.

Keywords
Artificial Intelligence (AI) • Small and Medium-Sized Enterprises (SMEs) • Manufacturing • Production • Data Management • Infrastructure Limitations • Regulatory Compliance

1 Introduction

The field of artificial intelligence (AI) is currently experiencing a period of rapid transformation, with significant implications for the industrial production landscape. This transformation is characterised by the introduction of new opportunities that promise to enhance efficiency, flexibility, and innovation in the manufacturing sector. While large enterprises have begun to harness the potential of AI to optimise processes and drive competitive advantage, small and medium-sized enterprises (SMEs) often face distinct and formidable barriers to adoption. These organisations, which form the backbone of many national economies, encounter unique challenges related to data management, infrastructure, regulatory compliance, workforce capabilities, and economic constraints.

The insights and analysis presented in this paper are grounded not only in a comprehensive review of recent literature and empirical studies but mainly in practical experience gained from multiple projects and workshops conducted in collaboration with SMEs. The range of these projects extends from use cases of Retrieval-Augmented Generation to individual solutions for 3D anomaly detection in quality control. The workshops for companies in production included usecase ideation workshop, strategy and ambition development workshops, maturity assessments and a simulation game for introducing AI into production. Through direct engagement with SME leaders, production managers and technical staff, the research team has been able to observe first-hand the specific obstacles and enablers that shape AI adoption in real-world production environments. These collaborative activities have provided valuable context, highlighted sector-specific nuances, and informed the development of actionable recommendations tailored to the needs of SMEs.

Despite the potential of AI, there's still a gap between potential and reality, especially for SMEs. Problems like data landscape fragmentation, old production systems, skills shortages and unclear return on investment can stop or even halt AI projects. Also, changing rules and certification requirements make things more complicated, so careful navigation is needed to stay within them and keep innovating.

The integration of artificial intelligence within small- and medium-sized enterprises can be approached from a number of theoretical perspectives. These include the technical, organisational, economic and regulatory dimensions. All problems can be categorised within a minimum of two distinct classifications, as seen in Figure 1. The decision has been taken to emphasise the data as a primary theme; consequently, the paper has been divided into five sections: Data, Systems, Compliance, Workforce and Economic.

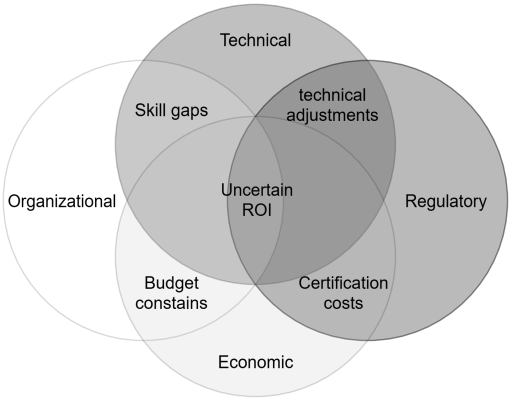


Figure 1: Overlapping Domains of Barriers to AI Adoption in SME Production

2 Related Work

2.1 Comprehensive Reviews and Frameworks

Recent literature offers a comprehensive overview of the challenges and strategic considerations for adopting AI in manufacturing. A comprehensive review by Masod et al. highlights organizational complexity, legacy systems, digital skills, company size, and R&D intensity as key determinants of successful AI implementation in the sector. This work emphasizes that both technical and organizational factors must be addressed to facilitate effective integration of AI technologies in production environments [20].

Another systematic review by Windmann et al. focuses on Industry 4.0, mapping the integration challenges, standards, and research trends associated with industrial AI. The study provides a detailed analysis of data management, system integration, workforce development, regulatory compliance, and quality assurance, offering a valuable reference for understanding the multifaceted barriers to AI adoption in the manufacturing sector [35].

2.2 Empirical and Case Study Research

Empirical research has examined the specific barriers faced by SMEs in adopting AI. Aarstad et al. conducted a multiple case study that identified a lack of AI competence, heavy dependency on external expertise, IT knowledge gaps, legacy systems, resistance to change, unclear business cases, and financial constraints as major obstacles for small and medium-sized enterprises. These findings underscore the importance of tailored support and capacity building for SMEs [1].

A case study of Swedish manufacturing firms by Mirzazadeh et al. explored both encouraging and discouraging factors in AI adoption. The research found that digital skills gaps, technical and cultural barriers, high initial costs, risk and uncertainty, and ethical concerns all play significant roles in shaping AI adoption decisions in the manufacturing sector [23].

2.3 Industry and Applied Research

Industry-driven research projects have further illuminated the practical challenges and opportunities of AI in manufacturing. The Bosch Research Group discusses ongoing research on data integration, hybrid modeling approaches, and the development of digital factories, emphasizing the potential of AI to enhance flexibility, quality, and efficiency in production processes [8].

In addition, Baljevic D. examines the challenges of scaling AI for manufacturing, with a particular focus on IT-OT (Information Technology–Operational Technology) convergence, data governance, workforce skills gaps, the contextualization of domain knowledge, and the physical realities of manufacturing environments [6].

2.4 Programs

Practical support programs play a crucial role in bridging the gap between research and real-world application. The KI-Transfer Plus program, as described by the Technical University of Applied Sciences Würzburg-Schweinfurt, provides hands-on assistance to SMEs in Bavaria, facilitating local AI knowledge transfer and showcasing use cases in areas such as computer vision, predictive maintenance, and chatbots. These initiatives demonstrate the value of targeted support in overcoming adoption barriers and fostering innovation in SME manufacturing [31].

While these programs and studies provide valuable insights, many remain limited in scalability, rely on external facilitation, or lack long-term evaluation of AI integration success. Furthermore, there is still a gap for SME stakeholders in understanding how to identify and solve AI integration within the SME ecosystem.

3 Data Management

AI systems require large volumes of high-quality data to function effectively. For SMEs, the ability to collect, store, process, and govern data is crucial for leveraging AI to improve production, optimize processes, and gain a competitive advantage.

3.1 Data Quality Issues

Data quality is a key concern for SMEs using AI in production. High-quality data is vital for training reliable AI models, making accurate predictions and supporting decision-making. However, SMEs often face data quality issues that can hinder AI success.

Inaccurate, Incomplete, or Inconsistent Data: Inaccurate, incomplete or inconsistent data Poor data quality can significantly reduce the reliability of AI outputs. When data is missing, mislabelled or recorded differently across systems, AI models may produce biased results or make costly errors. For SMEs, these issues often arise due to manual data entry, lack of standardised processes or disparate data sources. The consequences include reduced model performance, flawed insights and diminished trust in AI-driven recommendations. [28, 36].

Bias and Missing Data: Bias and missing data complicate the development of robust AI solutions. If datasets are not representative or contain omissions, AI models may reinforce existing

prejudices or make incorrect decisions. For example, predictive maintenance models may fail to identify critical issues if maintenance logs only capture certain types of failures or omit specific production scenarios. Addressing these risks requires careful data collection, validation, and ongoing monitoring to ensure completeness and fairness.s [34, 14].

Combining Data from Different Sources: Combining data from different sources presents additional challenges. Differences in data formats, standards and quality can make merging information into a coherent dataset suitable for AI applications difficult. SMEs may need to invest in transformation tools, establish common data standards and implement validation procedures to ensure integrated datasets are accurate and usable. [2, 29].

By proactively addressing these issues, SMEs can establish a stronger foundation for successful AI adoption, ensuring that their models are both reliable and effective in supporting production objectives.

3.2 Data Volume and Complexity

Managing the volume and complexity of data is a significant challenge for SMEs that adopt AI in production environments. Modern production systems generate vast amounts of data from sensors, machines, and business processes. For SMEs, the sheer scale and diversity of this information can quickly become overwhelming, particularly when IT resources are limited.

Handling Large, Complex Datasets: Data streams can include real-time sensor readings, logs and more. These datasets may contain many types of data, be high-frequency and unstructured and require significant storage and processing power, and specialised tools to manage them. SMEs may find this hard as they often lack the necessary infrastructure and expertise. Valuable insights may be missed and the potential benefits of AI unrealised. [9, 19].

Too Much or Too Little Data: Both data excess and scarcity pose unique problems for AI initiatives:

Too much data can cause information overload, making it difficult to identify relevant patterns or trends. It can also increase storage and processing costs and complicate model training. [19].

Conversely, **too little data** can undermine the reliability of AI models. Insufficient data may result in underfitting, where models fail to capture underlying relationships, or overfitting, where models become too tailored to limited examples and perform poorly on new data [18].

To address challenges, SMEs should collect quality data. This means prioritising relevant data sources, using reduction techniques, and ensuring datasets are both comprehensive and manageable. This optimises data use for AI applications, improving decision-making and efficiency.

3.3 Data Privacy and Security

Robust data privacy and cybersecurity are essential for SMEs adopting AI. Compliance with evolving regulations demands management of sensitive data and technical and organisational safeguards; yet, SMEs often lack resources to address these requirements effectively [32, 37]. Limited IT security infrastructure increases vulnerability to data breaches and cyberattacks, underscoring the need for continuous investment in privacy and security measures to protect information while maintaining compliance.

3.4 Data Silos and Foundations

SMEs encounter considerable challenges when seeking to implement artificial intelligence within production environments. These challenges can be attributed to the presence of data silos, fragmentation, and inadequate data management foundations. The concept of a data

silo refers to the isolation of information within a specific department, system, or application, thereby rendering it inaccessible to other parts of the organisation. Fragmentation can be defined as the distribution of data across multiple platforms or formats, which hinders efforts to gain a unified view of business operations. The result of these issues is the untapped potential of valuable data from different sources, which cannot be effectively combined. To illustrate this point, consider the challenge of integrating production data stored in manufacturing systems with sales or supply chain information. This integration issue can impede the potential for comprehensive analytics and effective AI model training. Fragmented data can also lead to inconsistencies, as different departments may maintain their standards, formats, or update schedules [2].

The consequences of data silos and fragmentation include reduced data quality and consistency, limited collaboration and knowledge sharing, increased integration costs, and slower decision-making processes [29]. The absence of clearly defined data management policies and governance frameworks is indicative of underlying issues. In the absence of standardized procedures, data is frequently collected, processed and stored in a non-uniform manner, resulting in elevated error rates and datasets that are unreliable, thereby compromising the efficacy of AI initiatives [28].

Additionally, many SMEs lack data governance and AI competencies. The absence of personnel with expertise in data integrity, metadata standards, and lifecycle management affects data quality and readiness, heightening the risk of non-compliance. Poor data preparedness often delays or fails AI projects [26].

In order to address these challenges, SMEs should prioritize the integration of their data systems, the establishment of robust data management policies, and the investment in skills development. The dismantling of silos, the reduction of fragmentation, and the implementation of governance frameworks are imperative steps in the process of unlocking the full value of data assets and establishing a robust foundation for the successful and scalable adoption of AI in production.

3.5 The Consequences of Weak Data Management in AI Projects

Data-related challenges are a leading cause of failure in AI initiatives, with studies indicating that as many as 30% of GenAI projects fail to progress beyond the pilot phase [13]. This high failure rate is largely attributable to substandard data quality, inconsistent formats, and a lack of centralized data governance. Poor-quality data not only diminishes model performance but also leads to biased outputs, unreliable predictions, and flawed decision-making. These outcomes erode stakeholder trust and can compromise critical business operations. Moreover, the absence of automated data integration tools forces SMEs to rely on labor-intensive manual processes for data cleaning and preparation. These tasks consume valuable time and human resources, diverting attention away from innovation and core business functions. For resource-constrained SMEs, this creates a bottleneck that prevents the scaling and operationalization of AI technologies [28].

3.6 Building Resilient Data Capabilities for AI Integration

To address data management challenges in AI adoption, several key strategies should be implemented. Organizations should begin by establishing clear data governance policies, which involve defining standards for data collection, validation, and maintenance [2]. Leveraging AI-driven data cleaning tools is also essential, as these tools can automate error detection, deduplication, and validation processes, thereby improving overall data quality [9].

Integrating data sources is another important step. By breaking down silos and connecting disparate systems, organizations can standardize data formats and ensure a unified data foundation [29]. Investing in staff training is equally critical; upskilling employees in data manage-

ment and AI fundamentals enhances the organization's ability to maintain high-quality, AI-ready data [26].

Regular data audits should be conducted to review and address any data quality issues that arise periodically. Finally, prioritizing data privacy and security is vital. This means adopting robust cybersecurity measures and ensuring compliance with relevant regulations to protect sensitive business and customer information [37]. Together, these strategies help build a strong foundation for successful AI initiatives in production environments.

4 Systems as a Barrier to AI Integration

Infrastructure limitations are a significant barrier to AI adoption, as many organizations struggle to integrate AI systems with existing IT environments that lack the necessary processing power, storage, and scalability to support AI workloads. Legacy systems, inadequate equipment, and unpredictable costs can result in delays, inefficiencies, or even failure of AI initiatives if businesses do not invest in modernizing their infrastructure and ensuring compatibility with advanced AI technologies [3].

4.1 Legacy System Integration

Integrating AI into production environments often exposes infrastructure limitations, particularly for SMEs with old assets. Many SMEs use 15+ year old equipment, causing compatibility issues with modern AI technologies. Legacy systems may use outdated communication protocols, lack digital interfaces, or be incompatible with contemporary data acquisition and control platforms, making integration challenging [4].

Another critical concern pertains to the latency of real-time inference. In 62% of AI deployments within industrial settings, latency exceeds the acceptable tolerance of $\pm 5\%$. AI-driven decisions or predictions are not delivered quickly enough to meet production process requirements, potentially leading to inefficiencies or safety risks. This is often exacerbated by legacy hardware's limited processing capabilities and the complexity of integrating AI with older control systems [4].

Proprietary control systems are another barrier. Legacy production environments rely on closed, vendor-specific solutions restricting external access via APIs. This hinders the ability to collect real-time data, automate workflows, or implement AI-driven optimizations, as external systems cannot communicate with proprietary controllers. SMEs face substantial costs and technical hurdles when modernizing their production infrastructure for AI readiness.

4.2 Scalability Demands

Scalability represents a pivotal challenge for SMEs seeking to broaden the implementation of AI beyond the confines of pilot projects. Edge devices, which facilitate the execution of AI processes in proximity to the data source, frequently exhibit inadequate computational capacity to support sophisticated models. Consequently, organisations are compelled to either streamline their models, a process that may entail a diminution in accuracy, or to allocate resources towards the acquisition of costly hardware enhancements [16]. Network bandwidth limitations further constrain scalability, as distributed AI systems require frequent data transfers between edge devices, local servers, and cloud platforms; insufficient bandwidth can slow data transmission, hinder real-time analytics, and limit the volume of data that can be processed across multiple sites [16]. As AI operations scale, cloud service costs become more unpredictable, with rapidly rising data

storage, processing, and transfer expenses making budget forecasting difficult. Sometimes this results in project scope reductions or overruns.

5 Compliance Barriers

A significant barrier for SMEs is the lack of universally accepted standards for AI system validation and certification. It is evident that initiatives such as ISO/TS 24028:2020 seek to establish guidelines for the trustworthiness of AI systems. However, the absence of mature, widely adopted protocols engenders uncertainty for organisations aiming to demonstrate compliance and reliability. This, in turn, complicates the process of evaluating solutions and gaining market trust. [15]. Formal certification can be expensive, especially for SMEs, and nearly half cannot afford it, which hinders market entry and innovation. AI systems create unique regulatory challenges. Certification processes are designed for static systems, so how do you validate, monitor, and certify evolving AI solutions? This uncertainty can delay deployment, increase risks, and deter investment in advanced adaptive AI technologies. [7].

6 Workforce: Bridging the Human–AI Divide

The successful adoption of AI in production environments by SMEs is heavily dependent on the capabilities of the workforce. However, significant gaps in skills and organizational acceptance often hinder the effective implementation and scaling of AI solutions.

6.1 Skills Mismatch

A critical challenge faced by SMEs is the shortage of "bilingual" experts who possess both operational technology and AI competencies. These professionals are essential for bridging the gap between traditional production systems and advanced AI technologies, yet they remain scarce in the labor market [17]. Furthermore, knowledge transfer bottlenecks frequently occur between data scientists and production operators, impeding collaboration and slowing down AI deployment [30]. Another key issue is the limited expertise in machine learning operations, which is crucial for the reliable deployment, monitoring, and maintenance of AI models in production settings. Lack of MLOps skills often leads to deployment failures and reduced system performance [12].

6.2 Organizational Resistance

Beyond technical skills, organizational resistance poses a significant barrier to AI adoption. Operators often distrust AI recommendations, primarily when models function as black boxes without transparent explanations [22]. Labor unions may express concerns about job displacement due to automation, creating social and political challenges for AI initiatives [25]. Management's reluctance to automate core processes further slows AI integration, as decision-makers may fear loss of control or disruption to established workflows [10].

Addressing these workforce capability gaps requires targeted training programs, improved communication between technical and operational teams, and transparent AI systems that build trust among users.

7 Economic and Operational Hurdles

The most significant financial impediments confronting SMEs contemplating or implementing AI in production are as follows: uncertainty with regard to return on investment (ROI) and the financial burden of maintenance overhead. SMEs frequently encounter elevated initial costs and protracted return periods, with initial investments often exceeding €250,000 and returns occasionally requiring over three years to materialise [33, 24]. The nature of many AI benefits, which are often intangible in nature, such as enhanced product quality and process flexibility, serves further to complicate the justification for such expenditures [33]. This uncertainty, in conjunction with budgetary constraints, compels many SMEs to prioritise immediate operational needs over long-term AI projects [24].

Once deployed, AI systems require ongoing maintenance to ensure continued effectiveness. It is important to note that model performance can degrade over time. In order to ensure continued accuracy and compliance, especially in regulated sectors, it is necessary to carry out regular retraining, often every few months. Maintenance entails rigorous version control and documentation, which introduce complexity and operational overhead [5]. Moreover, reliance on proprietary AI platforms can result in vendor lock-in, thereby reducing flexibility and increasing costs over the lifecycle of the solution [27].

SMEs should adopt thorough ROI assessments, communicate the benefits of AI to stakeholders, implement structured maintenance plans, and prioritise open, interoperable AI technologies to avoid lock-in.

8 Conclusion: A Roadmap for SME AI Transformation

The integration of artificial intelligence within SMEs has the potential to enhance efficiency, product quality, and adaptability to a considerable extent. However, as this paper has demonstrated, the path to successful AI integration is complex and shaped by interrelated barriers across technical, organisational, regulatory, and economic domains.

The training of reliable AI models continues to be hindered by poor data quality, fragmentation, and volume challenges. These issues often result in suboptimal outcomes or failed implementations. The deployment of AI is further complicated by legacy infrastructure, limited scalability, and unpredictable operating costs, particularly in resource-constrained environments. SMEs must also navigate strict compliance requirements, such as the General Data Protection Regulation, and evolving certification frameworks, often without dedicated legal or data protection personnel.

From a human perspective, the scarcity of professionals who possess dual expertise in both operations and AI—termed "bilingual experts"—in conjunction with organisational resistance and limited AI literacy, impedes the adoption and scaling of AI. The economic case continues to present significant challenges, characterised by an uncertain return on investment, maintenance burdens, and limited internal funding, which collectively contribute to the perception of AI as a risky proposition.

To overcome these challenges, SMEs should pursue a set of interconnected strategies:

- Strengthen Data Management: Build robust data governance, invest in data cleaning and integration tools, and prioritize staff training on data quality and AI basics.
- Modernize Infrastructure: Upgrade legacy systems gradually, adopt scalable cloud and edge computing solutions, and minimize vendor lock-in through open and interoperable platforms.

- Ensure Regulatory Readiness: Incorporate privacy-by-design and stay current with GDPR and certification standards, using industry partnerships and guidance where possible.
- Build Workforce Capabilities: Support cross-disciplinary training, hire or develop bilingual experts, and foster explainable AI systems to reduce resistance and build trust.
- Manage Economic Risk: Conduct detailed ROI analyses, plan for ongoing model maintenance and updates, and explore external funding or partnerships to spread cost and risk.

These approaches are consistent with the real-world conditions and constraints faced by SMEs. It is imperative to acknowledge that progress in this domain is contingent not solely on initiatives undertaken at the firm level, but also on the concerted efforts of a supportive ecosystem. This ecosystem encompasses policymakers, academia, technology providers, and industry bodies, who, through collaborative endeavors, must strive to diminish barriers and foster the adoption of scalable, trustworthy AI within the context of SME manufacturing.

Acknowledgments

We are grateful to the reviewers for their helpful suggestions, which led to significant improvements in the clarity and presentation of our findings. This research is supported by the Center for Artificial Intelligence (CAIRO) at the Technical University of Applied Sciences Würzburg-Schweinfurt (THWS), Würzburg, Germany and the Bavarian State Ministry for Digital Affairs.

References

[1] Aarstad A. & Saidl M. (2019). Barriers to Adopting AI Technology in SMEs: A Multiple Case Study. MSc Thesis, Copenhagen Business School.
[2] Adams Z. & Nelson A. (2022). Breaking Down Silos: Integrating Big Data Across Organizational Functions International Journal of Network and Communication Research, Vol.7
[3] Akoh, E. (2024). Adoption of artificial intelligence for manufacturing SMEs' growth and survival in South Africa: A systematic literature review. International Journal of Research in Business and Social Science (2147- 4478). 13. 23-37. 10.20525/ijrbs.v13i6.3561.
[4] Alqoud, A., Schaefer, D., & Milisavljevic-Syed, J. (2022), Industry 4.0: a systematic review of legacy manufacturing system digital retrofitting, Manufacturing Rev. 9 32, DOI: 10.1051/mfreview/2022031
[5] Ashmore, R., Calinescu, R., & Paterson, C. (2021). Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges, ACM Computing Surveys, 54(5), 1-39
[6] Baljevic D. (2023). Challenges in Scaling AI for Manufacturing. *ISG Research Report*.
[7] Bigham T. , Gallo V. , Nair S., Lee M., Soral S., Mews T., Tua A. & Fouché M., AI and risk management
[8] Bosch Research. (2025). Research Projects on the Use of AI in Manufacturing.
[9] Dai, H. N., Wang, H., Xu, G., Wan, J., & Imran, M. (2019). Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies. Enterprise Information Systems, 14(9–10), 1279–1303. https://doi.org/10.1080/17517575.2019.1633689
[10] Dally, S., Wiewiora, A., & Hearn, G. (202). Shifting attitudes and trust in AI: Influences on organizational AI adoption,Technological Forecasting and Social Change, Volume 215
[11] European Commission, General Data Protection Regulation (GDPR). Available: https://gdpr.eu/
[12] Faubel, L., Schmid, K. & Eichelberger (2023), H. MLOps Challenges in Industry 4.0. SN COMPUT. SCI. 4, 828 https://doi.org/10.1007/s42979-023-02282-2

[13] Gartner, GGartner Predicts 30% of Generative AI Projects Will Be Abandoned After Proof of Concept By End of 2025, (2024).
[14] Hafsi M., Hamour N. & Ouchani S. (2023). Predictive Maintenance for Smart Industrial Systems: A Roadmap. The 6th International Conference on Emerging Data and Industry (EDI40), Leuven, Belgium. ⟨hal-04398272⟩
[15] International Organization for Standardization (ISO). *24028:2020 – Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence*. Available: https://www.iso.org/standard/77608.html
[16] Kemnitz J. et al. (2023), An Edge Deployment Framework to Scale AI in Industrial Applications, 2023 IEEE 7th International Conference on Fog and Edge Computing (ICFEC), Bangalore, India, 2023, pp. 24-32, doi: 10.1109/ICFEC57925.2023.00012.
[17] Keshireddy S. (2024). Bridging the Gap: The Synergy of AI, Data Integration, and Data Science in Driving Innovation. International Journal of Innovative Science and Research Technology (IJISRT). 3. 16-18. 10.38124/ijsrmt.v3i10.48.
[18] Kraljevski, I., Ju, Y. C., Ivanov, D., Tschöpe, C., & Wolff, M. (2023). How to Do Machine Learning with Small Data? A Review from an Industrial Perspective. arXiv preprint arXiv:2311.07126.
[19] Li, X., Yang, C., Møller, C., & Lee, J. (2024). Data Issues in Industrial AI System: A Meta-Review and Research Strategy. arXiv preprint arXiv:2406.15784.
[20] Masod M. & Zakaria S. (2025). Artificial Intelligence Adoption in the Manufacturing Sector: Challenges and Strategic Framework. *International Journal of Research and Innovation in Social Science*, 9(4), 123-135.
[21] McMillan, L. & Varga, L. A review of the use of artificial intelligence methods in infrastructure systems. *Engineering Applications Of Artificial Intelligence*. 116 pp. 105472 (2022), https://www.sciencedirect.com/science/article/pii/S0952197622004626
[22] Miller T. (2019), Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence, Volume 267
[23] Mirzazadeh E. & Rostami M. (2024). AI Adoption in Production: Encouraging and Discouraging Factors in Swedish Manufacturing Firms as Case Study. Master Thesis, Jönköping University.
[24] Müller, J. M., Kiel, D., & Voigt, K.-I. (2018). What Drives the Implementation of Industry 4.0? The Role of Opportunities and Challenges in the Context of Sustainability. Sustainability, 10(1), 247.
[25] Nissim G., Simon T. (2021). The future of labor unions in the age of automation and at the dawn of AI, Technology in Society, Volume 67
[26] Oldemeyer, L., Jede, A. & Teuteberg, F. Investigation of artificial intelligence in SMEs: a systematic review of the state of the art and the main implementation challenges. Manag Rev Q 75, 1185–1227 (2025). https://doi.org/10.1007/s11301-024-00405-4
[27] Opara-Martins, J., Sahandi, R. & Tian, F. (2016), Critical analysis of vendor lock-in and its impact on cloud computing migration: a business perspective. J Cloud Comp 5, 4. https://doi.org/10.1186/s13677-016-0054-z
[28] Peretz-Andersson E., Tabares S., Mikalef P., Parida V.(2024), Artificial intelligence implementation in manufacturing SMEs: A resource orchestration approach, International Journal of Information Management, Volume 77
[29] Plathottam S., Rzonca A., Lakhnori R. & Iloeje C. (2023). A review of artificial intelligence applications in manufacturing operations. Journal of Advanced Manufacturing and Processing. 5. 10.1002/amp2.10159.
[30] Subramaniyan M., Skoogh A.,Bokrantz J., Sheikh M., Thurer M. & Chang, Qing. (2021). Artificial intelligence for throughput bottleneck analysis – State-of-the-art and future directions. Journal of Manufacturing Systems. 60. 734-751. 10.1016/j.jmsy.2021.07.021.
[31] Technische Hochschule Würzburg-Schweinfurt & appliedAI Initiative. (2025). KI-Transfer

Plus Program Overview. *Program Report*.
[32] Voigt, P., & Von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR): A Practical Guide, Springer
[33] Wamba, S., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015), How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study, International Journal of Production Economics, Volume 165
[34] Wei X., Kumar N. & Zhang H. (2021). Addressing Bias in Generative AI: Challenges and Research Opportunities in Information Management, arXiv:2502.10407v1
[35] Windmann A., Wittenberg P., Schieseck M. & Niggemann O. (2024). Artificial Intelligence in Industry 4.0: A Review of Integration Challenges for Industrial Systems. *arXiv preprint arXiv:2405.18580*.
[36] Xie J., Sun L. & Zhao Y. (2025). On the Data Quality and Imbalance in Machine Learning-based Design and Manufacturing—A Systematic Review, Engineering, Volume 45
[37] Yuhan, N., & Hamilton, J. (2024). Strengthening SMEs through Cybersecurity and AI: A Path to Operational Excellence

Anomaly Detection on the Edge for Quality Inspection

Andreas Marchl[1] 0009-0003-8471-8983
Maxmimilan Kasper[1] 0009-0007-3160-5028
Simon Geis[1] 0009-0006-5173-4815
Mark Deutel[1] 0000-0001-8932-5212
Axel Plinge[1] 0000-0001-7757-2953 [a]
Dominik Seuss[1,2] 0000-0001-5302-2236

[a]Corresponding author.

[1] *Fraunhofer Institute for Integrated Circuits, Division Positioning and Networks, 90411 Nürnberg, Germany*
[2] *Technical University of Applied Sciences Würzburg-Schweinfurt, Center for Artificial Intelligence and Robotics, 97082 Würzburg, Germany*

Quality inspection via computer vision is a growing use-case for artificial intelligence (AI). However, two obstacles are hindering its wide-spread adaptation. First, integrating AI often requires major process changes, such as cloud access and large installations. Second, to train the AI, many defective samples are needed. The first hurdle can be overcome by using edge AI, the second one by using anomaly detection. We demonstrate how the combination of the two can offer a lightweight yet reliable solution for quality inspection. This is implemented and demonstrated on a Raspberry Pi 4.

Keywords
Anomaly Detection • Edge AI • Demonstrator

1 Introduction

Artificial intelligence (AI) on the edge is the deployment of AI functionality at the end of the cloud-edge continuum, directly at the sensor and actuator level [1, 2]. This has a number of advantages, such as low latencies, fast response times, and reduced energy consumption, while ensuring maximum privacy and data governance, as the data do not have to leave the edge. Edge AI is deployed on embedded systems with limited computational resources [3, 4]. In our demonstrator, we combine recent scientific developments in edge AI and anomaly detection for visual quality inspection.

Visual quality inspection is a critical component of modern manufacturing. While machine learning offers powerful automation capabilities for this task, a fundamental challenge arises from the data itself. In functioning manufacturing processes, defects are inherently rare, but can manifest in countless, often unforeseen, ways. This severe data imbalance renders traditional supervised approaches impractical, because they require extensive labeled examples for every defect type.

Unsupervised anomaly detection (AD) resolves this. AD is the identification of rare events or observations which deviate significantly from the majority of the data. Unsupervised methods operate by learning a robust model of “normality”, purely from images of non-defective parts or products. During inference, any part that deviates significantly from this learned normality is automatically flagged as anomalous.

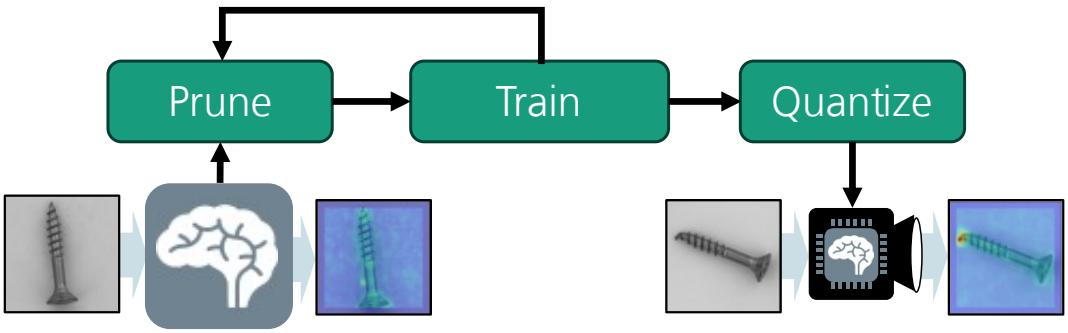


Figure 1: Deep compression workflow for developing an edge AI demonstrator for industrial visual quality inspection. Model pruning, retraining, and quantization allow an efficient deployment on resource-constrained edge devices.

Current state-of-the-art (SOTA) models for AD can be categorized into *generative* approaches, that learn to reconstruct normal images in the training phase, often utilizing autoencoder-based architectures, such as EfficientAD [5], *discriminative* approaches that encode image-level or patch-level extracted features into distributional statistics, such as PaDiM [6] and FastFlow [7], as well as *hybrid* methods, such as PatchCore [8] and SuperSimpleNet [9]. Anomalies are then identified either by measuring the reconstruction error, or the distance to the learned normal distributional statistics.

"Deep compression" is a suite of techniques that focus on optimizing deep neural networks (DNNs) by structural or implementation changes [4, 10]. The techniques try to detect and remove redundant information in trained DNNs in different ways. Pruning detects neural structures that encode similar features and then filters and removes redundancies [11, 12]. Quantization reduces the resolution of the weights of DNNs, typically from 32-bit floats to 8-bit integers, reducing the memory requirements of DNNs significantly [11, 13]. Especially for computer vision applications, implementation at the edge can be challenging due to constrained computational power and restricted memory [14].

2 Method

Our framework compresses large SOTA DNN models used for anomaly detection to a size that will work within the constraints of embedded systems, in the case of our demonstrator on a Raspberry Pi 4. To achieve this, we augment the training process of the anomaly detection models with iterative structured pruning and post-training quantization, see Figure 1. Our approach is implemented with PyTorch and a visual anomaly detection library "anomalib" [15] to implement the training, as well as our own proprietary pruning library [14], and PyTorch's built-in post-training quantization capabilities to design and train compressed DNN models. These models can then be converted and exported to the ONNX format, allowing us to deploy them on the Raspberry Pi 4 with the help of onnxruntime [16].

3 Results

To find a suitable anomaly detection model for the demonstrator, we first performed a comparative analysis of several current SOTA unsupervised anomaly detection models. We evaluated the performance of each model on the MVTecAD benchmark dataset [17], a standard benchmark for industrial visual inspection tasks. As shown in Table 1, the models were compared in terms

of their image-level AU-ROC score, inference latency on a desktop PC, number of trainable parameters, and resulting file size. The comparison shows that the "EfficientAD-S" architecture [5] outperformed all other models in performance and efficiency. Thus, we selected it for the additional pruning and quantization procedure: To further compress we iteratively re-trained it while applying increasing prunings, starting with $0\,\%$ pruned (baseline) and ending with approximately $60\,\%$ of the trainable parameters removed. We show the results of pruning in Table 2. For each pruning, we trained the "EfficientAD-S" model for 85 epochs and on input images of size 256×256 pixels. We show quantized (UInt8) and unquantized (Float32) results. The goal was to find a trade-off between the AU-ROC performance (col. 2, maximize) and latency on the Raspberry Pi 4 (col. 3, minimize).

The results show that when training on the MVTecAD screw subset, $20\,\%$ of the neural structures can be safely removed from the model without notably affecting its AU-ROC score. For stronger prunings, increasingly significant performance degradation is observed. This observation can also be validated qualitatively, cf. Figure 2. Even though the predicted anomaly mask of the $20\,\%$ pruned model (c) is still very similar to the unpruned baseline (b) and the anomaly is clearly detected compared to the ground truth (a), the $40\,\%$ pruned model (d) already produces an anomaly mask that is worse and also does not detect the anomalous defect of the screw as clearly. Nevertheless, even a pruning of $20\,\%$ speeds up the inference of the "EfficientAD-S" model on the Raspberry Pi 4, particularly when using the quantized version, as evidenced by the values in the fourth column of the table. Compared to the floating-point baseline, the $20\,\%$-pruned and quantized version already achieves a speedup of 1.9, making it, together with the still very good AU-ROC score, the optimal compromise we chose for final deployment on our demonstrator.

4 Conclusion

In this work, we presented our approach to finding an efficient DNN model for anomaly detection at the edge for quality inspection. After evaluating several SOTA models, we selected the "EfficientAD-S" architecture, which we pruned and quantized to enable deployment on a Raspberry Pi 4. We achieved about $2\times$ faster inference and $80\,\%$ reduction in model size through the compression.

References

[1] S. S. Gill, M. Golec, J. Hu, M. Xu, J. Du, H. Wu, G. K. Walia, S. S. Murugesan, B. Ali, M. Kumar, K. Ye, P. Verma, S. Kumar, F. Cuadrado, S. Uhlig, Edge AI: A taxonomy, systematic review and

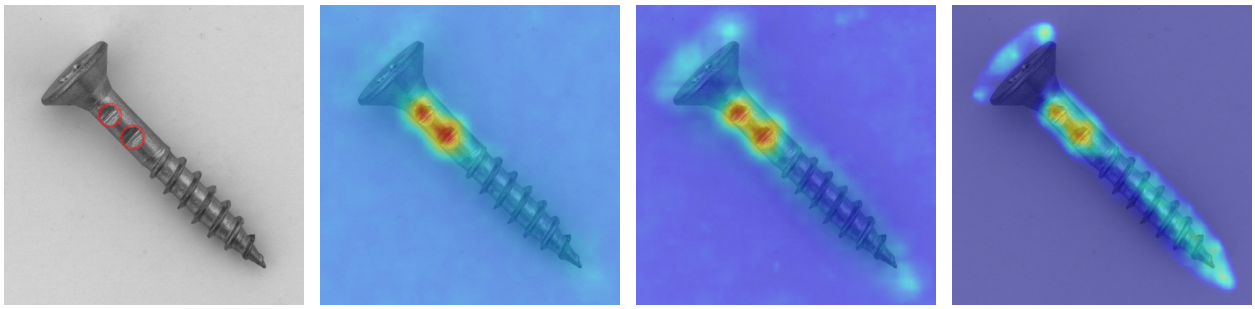
Table 1: Comparison of current state-of-the-art unsupervised anomaly detection models. Latency measured on AMD Ryzen7 5800H & NVIDIA RTX 3080. The reported "AU-ROC" score corresponds to the combined average image-level area under the receiver operator curve (ROC) curve over all 15 sub-categories of the MVTecAD dataset.

Model	**↑AU-ROC**	**↓Latency [ms]**	**↓Params. [$\times 10^6$]**	**↓File Size [MB]**
EfficientAD-S [5]	**0.987**	**8.9**	8.1	31
PatchCore [8]	0.980	24.7	24.9	291
SuperSimpleNet [9]	0.980	20.7	33.7	129
FastFlow [7]	0.907	16.2	7.7	**27**
PaDiM [6]	0.891	17.2	**2.9**	169

Table 2: Performance of the "EfficientAD-S" model under different pruning and quantization levels. The AU-ROC metric corresponds to the image-level score for the MVTecAD Screw category. Latency is evaluated on a Raspberry Pi 4.

Pruning	Quantization	↑AU-ROC	↓Latency [s]	↓File Size [MB]
0 %	Float32	0.975	4.03	30.8
	UInt8	0.977	2.65	7.8
20 %	Float32	0.928	3.30	25.6
	UInt8	0.905	2.13	6.5
40 %	Float32	0.730	2.46	20.8
	UInt8	0.730	1.60	5.3
60 %	Float32	0.544	2.31	16.7
	UInt8	0.528	1.21	4.3

future directions, Cluster Computing 28 (2025). doi:10.1007/s10586-024-04686-y.
[2] N. Witt, M. Deutel, J. Schubert, C. Sobel, P. Woller, Energy-efficient AI on the edge, in: Unlocking Artificial Intelligence: From Theory to Applications, Springer, 2024, pp. 359–380. doi:10.1007/978-3-031-64832-8_19.
[3] K. Morik, J. Rahnenführer, C. Wietfeld (Eds.), Machine Learning under Resource Constraints - Applications, volume 3, De Gruyter, Berlin, Boston, 2023. doi:10.1515/9783110785982.
[4] A. Plinge, A. Mishra, Getting AI in your pocket with deep compression, in: Embedded World Conference, Nuremberg, Germany, 2020. doi:10.13140/RG.2.2.27791.29603.
[5] K. Batzner, L. Heckler, R. König, Efficientad: Accurate visual anomaly detection at millisecond-level latencies, in: 2024 IEEE/CVF Winter Conf. App. Computer Vision, 2024, pp. 127–137.
[6] T. Defard, A. Setkov, A. Loesch, R. Audigier, PaDiM: A patch distribution modeling framework for anomaly detection and localization, in: A. Del Bimbo, R. Cucchiara, al. (Eds.), Pattern Recognition. ICPR Int. Workshops and Challenges, 2021, pp. 475–489.
[7] J. Yu, Y. Zheng, X. Wang, W. Li, Y. Wu, R. Zhao, L. Wu, Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows, arXiv preprint arXiv:2111.07677 (2021).
[8] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, P. Gehler, Towards total recall in industrial anomaly detection, in: IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 14318–14328.
[9] B. Rolih, M. Fučka, D. Skočaj, Supersimplenet: Unifying unsupervised and supervised learning for fast and reliable surface defect detection, in: International Conference on Pattern



(a) ground truth **(b)** unpruned **(c)** 20 % pruned **(d)** 40 % pruned

Figure 2: An example of the MVTecAD Screw sub-dataset with ground truth segmentation mask (a) and overlaid predictions of compressed models (b-d).

Recognition, Springer, 2024, pp. 47–65.
[10] M. Deutel, A. Plinge, D. Seuß, C. Mutschler, F. Hannig, J. Teich, Unsupervised learning of variational autoencoders on Cortex-M microcontrollers, in: IEEE Int. Sym. Embedded Multicore/Many-core Systems-on-Chip, 2025.
[11] S. Han, H. Mao, W. J. Dally, Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding, arXiv preprint arXiv:1510.00149 (2015).
[12] Y. LeCun, J. S. Denker, S. A. Solla, Optimal brain damage, in: Advances in neural information processing systems, 1990, pp. 598–605.
[13] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, D. Kalenichenko, Quantization and training of neural networks for efficient integer-arithmetic-only inference, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2704–2713.
[14] M. Deutel, P. Woller, C. Mutschler, J. Teich, Energy-efficient deployment of deep learning applications on Cortex-M based microcontrollers using deep compression, in: Workshop on Methods and Description Languages for Modelling and Verification of Circuits and Systems, 2023.
[15] S. Akcay, D. Ameln, A. Vaidya, B. Lakshmanan, N. Ahuja, U. Genc, Anomalib: A deep learning library for anomaly detection, 2022. arXiv:2202.08341.
[16] ONNX runtime developers, ONNX runtime, https://onnxruntime.ai/, 2021.
[17] P. Bergmann, M. Fauser, D. Sattlegger, C. Steger, MVTec AD – a comprehensive real-world dataset for unsupervised anomaly detection, in: IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2019, pp. 9592–9600.

Cross-Attentive Bipartite Graph Reinforcement Learning for Prize-Collecting Job Shop Scheduling

Alexander Mattick[1] alexander.mattick@iis.fraunhofer.de
Louis Sturm[2]
Felix Müller[2]
Sebastian Loftus[1] sebastian.loftus@iis.fraunhofer.de
Haris Asif[1] haris.asif@iis.fraunhofer.de
Axel Plinge[1] axel.plinge@iis.fraunhofer.de
Christopher Mutschler[1] christopher.mutschler@iis.fraunhofer.de
Jaspar Pahl[1] jaspar.pahl@iis.fraunhofer.de
Dominik Seuß[1,3] dominik.seuss@iis.fraunhofer.de
Quirin Göttl[1] quirin.göttl@iis.fraunhofer.de

[1] *Department Machine Intelligence, Fraunhofer Institute of Integrated Circuits (IIS), Nürnberg, Germany.*
[2] *DATEV eG, Nürnberg, Germany.*
[3] *Technical University of Applied Sciences Würzburg-Schweinfurt, Center for Artificial Intelligence and Robotics, Würzburg, Germany.*

In this work, the Prize-Collecting Job Shop Scheduling Problem is addressed under stochastic conditions using a Reinforcement Learning (RL) approach. The scheduling environment is modeled as a Markov Decision Process, where a bipartite graph representation is employed to capture the structure between workers and operations. A cross-attention mechanism is used to compute assignments based on worker availability, skill compatibility, and task readiness. Stochastic events, including worker unavailability, operation delays, and dynamically arriving jobs, are incorporated into the simulation. The proposed method is evaluated on realistic scheduling scenarios (approximately 15 workers and 250 jobs) derived from tax office operations. Performance is compared against a set of benchmark algorithms based on heuristic and evolutionary strategies, adapted to stochastic settings via iterative re-planning. Experimental results indicate that the RL-based approach achieves robust and efficient scheduling performance in the presence of uncertainty.

Keywords
Reinforcement Learning • Job Shop Scheduling • Machine Learning

1 Introduction

Reinforcement Learning (RL) has emerged as a powerful framework for sequential decision-making and control, with successful applications spanning a wide range of domains such as robotics, game playing, autonomous driving, and recommendation systems [1, 2, 3]. In these settings, RL agents learn to make decisions by interacting with an environment, receiving feedback in the form of rewards, and using this feedback to improve future behavior.

A key strength of RL is its natural formulation for planning under uncertainty. By modeling the environment as a Markov Decision Process (MDP), RL algorithms can reason about long-term consequences of actions and optimize behavior even when the dynamics of the environment are unknown or stochastic.

Compared to black-box optimization methods such as genetic algorithms or random search, RL methods explicitly leverage the structure of sequential decision problems. Rather than treating each action or policy evaluation independently, RL exploits temporal correlations and learns value functions or policy gradients to guide exploration and improve sample efficiency. This often results in fast convergence and more principled learning in high-dimensional or continuous control tasks [4, 5, 6, 3].

In this work, we investigate the application of RL to a challenging variant of the classical Job Shop Scheduling Problem (JSSP), which we refer to as the Prize-Collecting Job Shop Scheduling Problem (PC-JSSP). The name follows the convention of other combinatorial optimization problem variants, such as the Prize-Collecting Traveling Salesman Problem [7] and the Prize-Collecting Steiner Tree [8]. Unlike traditional JSSPs, which aim to sequence all given jobs to minimize objectives such as makespan or tardiness, the PC-JSSP considers a fixed and limited scheduling horizon. The goal is to select and schedule a subset of available jobs in order to maximize the total collected reward (e.g., profit or utility), subject to resource constraints. In the literature, such problems are sometimes also referred to as flexible JSSPs with throughput as the primary optimization objective [9, 10].

A distinctive aspect of our setting is the incorporation of stochastic events that can disrupt the planned schedule. For instance, workers may become unavailable due to sudden vacations, or unexpected delays may affect job durations. These uncertainties make it essential to develop adaptive strategies that can learn to plan effectively and re-plan dynamically in response to disruptions.

This problem arises in real-world scenarios where tasks must be assigned to limited human or machine resources, each with different and often overlapping skill sets. The problem of stochastic PC-JSSPs is particularly applicable in emerging Smart Manufacturing environments, e.g. [11, 12], where low-volume, high-mix orders are produced on a high number of specialised machines, increasing the likelihood of failure and allowing for more dynamic selection of individual jobs, rather than large bulk contracts. Beyond traditional job scheduling, PC-JSSP formulations are also important in healthcare, emergency or machine maintenance scenarios where specialized personnel has to be assigned to the most in-need patients or machines. Since these problems are usually constrained by resources (e.g., doctors or firefighters) rather than time, one has to resort to PC-JSSP formulations rather than traditional JSSPs.

In this work, we focus on a specific use case: the PC-JSSP for tax offices. Tax offices often face the challenge of having too few workers available to manage a surplus of jobs with tight and varying deadlines. It becomes sensible to commit to a job only if its operations can be fully completed within the given timeframe, making PC-JSSP a suitable formulation for this scenario. We established a simulation of tax office work scheduling that incorporates stochastic events such as delays, workers falling ill, new tasks arising, and vacations. We conceptualize this problem as a Markov Decision Process and explore the application of RL to develop scheduling policies that are robust to uncertainties and scalable to realistic problem sizes. Our experimental setup considers problem instances with approximately 15 workers and 250 jobs, reflecting the complexity and uncertainty found in real-world operations. Our approach illustrates how RL can be leveraged for not only online decision-making but also robust planning in uncertain, high-dimensional scheduling environments. Our formulation of the PC-JSSP is derived from real-world requirements, making it highly practical but also quite specialized. As a result, it is difficult to find existing methods, whether based on RL, heuristics, or other paradigms, that are directly applicable for comparison. Therefore, we adapt several well-established approaches originally developed for the classical JSSP to our setting, and benchmark our method against these adapted baselines.

2 Related Work

2.1 RL for JSSP variants

There are many JSSP variants with different constraints that reflect a wide range of real-world scenarios. These often require tailored solution methods, complicating direct comparisons between approaches. As a result, methods are frequently benchmarked against heuristic rule systems. We highlight representative RL-based approaches for JSSP variants and refer to [13] for a comprehensive review.

Waschneck et al. [11] apply Deep Q-Networks (DQN) to production scheduling in complex, dynamic job shop environments, using a semiconductor manufacturing case study. Their approach uses multiple cooperative agents responsible for dispatching at individual workcenters, accounting for constraints like setup times, re-entrant flows, batching, and tool dedication. The simulation includes stochastic elements to reflect real-world variability, and their DQN system performs comparably to expert-designed heuristics.

Zhang et al. [14] present a deep RL framework that learns scheduling policies using graph-based representations. A graph neural network encodes scheduling information and produces generalizable policies applicable to varying instance sizes without retraining. Their method achieves lower makespans than rule-based heuristics while maintaining computational efficiency.

In [12], a double-DQN approach trains an agent for resource allocation in business processes, structurally similar to JSSPs but with randomized transitions after each operation assignment. The method outperforms simple heuristics such as First-In-First-Out and Shortest Processing Time (SPT) on small instances.

Tassel et al. [15] propose a RL agent and environment for the classical JSSP using Proximal Policy Optimization (PPO) [5], similar to our work. PPO enables stable policy updates and robust learning. However, their setup excludes stochastic events, uses a static representation that cannot adapt to varying job or worker numbers, and does not model workers explicitly, enforcing feasibility by masking invalid actions.

Echeverria et al. [16] apply offline RL with heterogeneous graph neural networks to capture dependencies between operations, jobs, and workers. Like Tassel et al., their focus is on the classical JSSP rather than more complex variants like the prize-collecting JSSP. Their method explicitly models intra-job dependencies, resulting in a more complex graph structure than ours.

Bonetta et al. [17] adopt a sequence-to-sequence decoding approach based on Pointer Networks [18] to model the classical JSSP. Their method linearizes the JSSP instance before decoding, similar to earlier Pointer Network applications in combinatorial problems like the Knapsack and Travelling Salesman Problems [19].

2.2 Heuristics and Genetic Algorithms for JSSP Variants

Beyond learning-based methods, a large body of research addresses the JSSP and its variants using classical techniques, mainly mathematical optimization and heuristics. Mathematical optimization provides exact solutions but is generally limited to small instances, while heuristics offer more scalable alternatives that deliver high-quality solutions in reasonable time. For an overview, see Momenikorbekandi and Kalganova [20], who highlight the practical success of hybrid strategies, especially in flexible and non-standard JSSPs. Türkyılmaz et al. [10] survey heuristic methods for multi-objective flexible JSSPs, covering algorithms like PSO, tabu search, and evolutionary techniques. They emphasize the strength of hybrid algorithms but also point out the dominance of synthetic benchmarks and limited attention to real-world uncertainty.

Jamili [21] focuses on robust scheduling under stochastic disruptions, introducing buffer-based methods combined with exact and heuristic algorithms, including Branch and Bound, Beam Search, and PSO. These approaches work well for small cases (e.g., 10 jobs and 10 machines), but their scalability to larger, more complex scenarios—like our setting with about 15 machines, 250 jobs, and stochastic events—is unclear.

Additional work refines genetic and evolutionary approaches. Ripon et al. [22] apply evolutionary techniques to multi-objective JSSPs, while Bhatt and Panchal [9] review genetic algorithm configurations. Differential Evolution has been adapted to discrete scheduling via constraint relaxation [23].

To evaluate our RL method, we adapt several of these heuristic and evolutionary algorithms as benchmarks, modifying them to handle prize-collecting objectives and stochastic events in our problem setting.

3 Method

3.1 Simulation Environment

We implemented a custom simulation in Python to model staff scheduling within a tax office. The simulation environment is designed to reflect realistic operational conditions, including job structures, worker capabilities, departmental responsibilities, and a variety of stochastic disruptions.

Each incoming job consists of a sequence of operations that must be completed in a predefined order across different departments. The environment comprises four departments, and a given job may require only a subset of them. The duration of each operation is known in advance; however, stochastic delays can occur during execution. The probability of such delays is configurable. Every job has a predefined deadline, and a **reward** is assigned only if the entire job is completed before this deadline. The **reward** is calculated based on the total processing time allocated to each department, with each department having a fixed per-time-unit payment rate. This structure results in a linear scaling of the **reward** with the work contributed by each department. The tax office is staffed by a number of workers, each associated with at least one department, though multi-departmental assignments are possible. The simulation includes several types of stochastic events, such as workers falling ill, taking vacation, delays during task execution, and the arrival of new jobs.

The simulation operates in discrete time steps. At each step, it provides a complete state snapshot, including information on which workers are idle or occupied, which operations are currently in progress, which jobs have operations ready for assignment, which workers are unavailable due to illness or vacation, and how much time remains within the scheduling horizon. At every timestep, the user or RL agent can assign idle workers to eligible operations. Operations not ready for assignment are masked and thus not selectable. Alternatively, the agent may choose to advance to the next timestep without making any new assignments. In this case, idle workers remain unoccupied until the environment progresses, potentially unlocking new operations that match their skill sets. This decision allows the agent to strategically wait for better assignment opportunities rather than forcing suboptimal ones.

The objective of the agent is to maximize the total **reward** by efficiently scheduling operations and completing as many jobs as possible within a fixed planning horizon, while dynamically adapting to uncertain conditions.

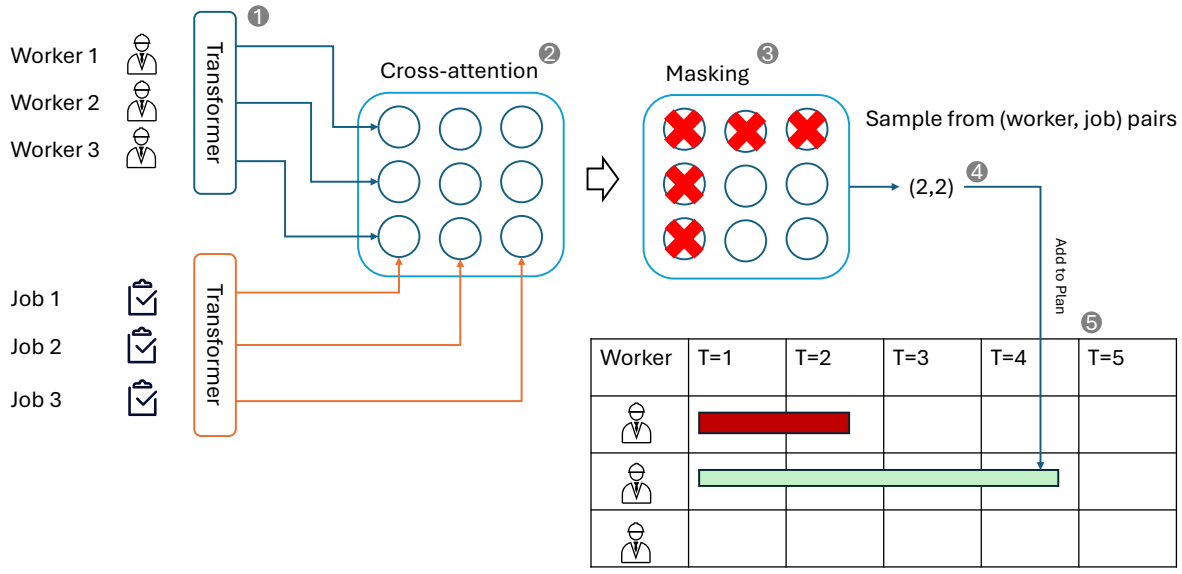


Figure 1: Our method relies on a bipartite graph representation of the per-step assignment process. For every timestep, for every worker assignment, we encode the worker information and job information using two separate transformers (1). The resulting embeddings are used in a cross-attention like scheme (2) to build a bipartite graph. We then mask invalid assignments (3) and sample a new pair from the bipartite graph (4). The resulting worker-job assignment is then placed on the plan at the current timestep (5). Not shown: A second output can issue a "continue" signal that directly steps to the next timestep without assigning all workers.

3.2 RL using Bipartite Graph Representation

We formulate the PC-JSSP as a sequential decision-making task over discrete time steps. To bridge the gap between real-world time slots and our decision process, we subdivide each time slot into smaller "scheduling steps". At each scheduling step, the agent assigns one available worker to one available operation within the current time slot. This process repeats until either all workers or all operations have been scheduled, or until the agent chooses a special "continue" action. A worker remains assigned to an operation until that operation either completes, or a random event occurs, e.g., the worker gets sick. This gives us an MDP with states being available workers and jobs, actions being worker-job assignments, and rewards being defined by the total value of jobs fully completed.

The state at time step is represented by two sets - workers and jobs - each encoded with their own feature vectors. A worker's features include their salary (which can vary over departments) and the subset of operations they can perform. A job's features comprise of its current operation, the list of remaining operations required for completion, and the expected reward (as described in Section 3.1) for finishing all operations on time. Since every job has at most four operations (the number of departments in our setting), we one-hot encode the pending operations into a fixed-length vector.

Because the numbers of available workers and operations vary at each time step, our model must flexibly accommodate variable-sized input sets while remaining computationally efficient. To achieve this, we first embed all workers and jobs into a shared latent space using a Trans-

former encoder [24]. We then model affinities between workers and jobs via a cross-attention-style mechanism (see Figure 1), which effectively constructs the adjacency matrix of a weighted bipartite graph between the two sets.

Concretely, let $W \in \mathbb{R}^{d\times|W|}$ be the worker embeddings, $J \in \mathbb{R}^{d\times|J|}$ be the job embeddings, and $M \in \{-\infty, 0\}^{|W|\times|J|}$ the mask for available selections. We compute the assignment probabilities as

$$p((w,j)) = \text{softmax}\left(W^T J + M\right)_{w,j}, \tag{1}$$

where the softmax is computed over the entire matrix, rather than row-wise as is done in ordinary cross-attention.

Because Transformers are permutation-equivariant, our entire pipeline - embeddings, cross-attention, and the resulting bipartite graph - is invariant to the ordering of workers or jobs. In other words, both workers and jobs are treated as unordered sets rather than as order-aware sequences.

RL for combinatorial optimization is often hampered by high inter-instance variance. Even with a fixed number of workers and jobs and no stochastic events, randomly sampling different makespans produces a wide spectrum of instances, from trivial to nearly infeasible. Such variance undermines RL's reliance on accurately estimating the advantage, making credit assignment intractable when instances themselves vary greatly.

We solve this limitation by using a new algorithm originally proposed for tuning large language models known as GRPO [25]. GRPO exploits environments where the same instance can be re-executed deterministically at low cost - a feature uncommon in most RL benchmarks but readily available in our scheduling setting. GRPO replaces the existing advantage estimation techniques (such as [4]) with empirical per-instance standardization. For each problem instance, we perform n independent rollouts to collect rewards $r_1, \ldots, r_n$ and then standardize them via

$$\hat{r}_i = \frac{r_i - \mathbb{E}[r]}{\sigma(r)}$$

These per-instance normalized rewards replace traditional advantage estimators (e.g., [4]), and, when aggregated across many instances, yield a good, unbiased estimate of the overall advantage.

Intuitively, this estimation method downweights uncertain (high variance) instances contributions, compared to certain (low variance) instances. This appears to be quite sensible as one has to be a lot more careful to draw inferences in very risky instances, compared to easy ones. Just like [25] , we then use these advantage estimations within PPO [5] to optimize our selection policy.

3.3 Benchmark Algorithms

To evaluate the performance of our RL approach, we implement several benchmark algorithms based on heuristics and evolutionary strategies. These algorithms are adapted to our formulation of the PC-JSSP under the assumption of a deterministic setting without stochastic events or disruptions.

For evaluation in the stochastic setting, each algorithm first generates a complete schedule assuming no stochasticity. This schedule is then executed within our simulation environment. When a stochastic event occurs at a given timestep, the current schedule is invalidated, and the algorithm replans the schedule from that point forward - again under the assumption of no further stochasticity. The updated schedule is then simulated until the next disruption, and the process repeats.

In the following, we describe the rationale and implementation of each benchmark algorithm and how it has been adapted to suit our PC-JSSP formulation.

Modified Shortest Processing Time (MSPT)

SPT is one of the simplest and most computationally efficient heuristics for the classical JSSP [26]. It schedules the operation with the shortest processing time next.

Inspired by SPT, we design a heuristic tailored to our setting. At each decision point, we first rank the currently available operations by their earliest deadline. In the case of a tie, we prioritize the operation belonging to the job with the smallest total remaining processing time (sum of all processing times within this job). If a tie persists, we select the operation with the shortest individual processing time.

Next, we assign workers to these operations. For each operation, we sort all eligible workers by the earliest possible start time. If multiple workers are available at the same time, we prefer the one with the lowest workload, measured as the sum of durations of all previously completed operations.

This process is repeated iteratively: as soon as all currently available operations have been assigned, we re-evaluate the list of available operations and assign them in the same manner. The procedure continues until the schedule is complete, i.e., until the predefined scheduling horizon of the PC-JSSP has been reached.

Deadline-Aware Greedy Heuristic (DAGH)

While MSPT aims to balance workload across workers, which can be beneficial in settings with human resources, it is not always optimal for pure scheduling performance. Moreover, MSPT assigns operations in batches, first ordering all available operations, assigning them to workers, and then proceeding to the next set, without dynamically reordering after each assignment.

To address these limitations, we propose the Deadline-Aware Greedy Heuristic (DAGH). DAGH operates as follows: At each step, all currently available operations are placed in a priority queue, sorted by the earliest possible starting time. The algorithm selects the first operation from the queue and assigns it to a suitable worker. Among eligible workers, it chooses the one that can start the task at the earliest time while ensuring that the job can still be completed within its deadline.

Once an operation is assigned, the next operation in the corresponding job is added to the queue. The queue is then re-sorted by earliest starting time, and the process repeats until the scheduling horizon has been reached.

Genetic Algorithm (GA)

We propose a genetic algorithm (GA) tailored to our variant of the PC-JSSP, inspired by concepts from evolutionary algorithms such as Differential Evolution [23]. The algorithm operates on a population of candidate schedules, where each individual represents a complete schedule, including both the operation sequence and worker assignments.

The process begins by initializing a random population. It is ensured that each randomly generated individual corresponds to a legal schedule, respecting the required operation order within each job and assigning only workers with the appropriate skills.

In each generation, for every individual, we generate a mutant schedule through reordering operations and apply crossover between the mutant and the original individual to produce an offspring. Both mutation and crossover modify only the sequence of operations, not the worker assignments. Diversity in worker allocation is maintained through the size of the population.

To guarantee feasibility during mutation and crossover, we apply a correction mechanism that identifies and reorders any operations that violate job precedence constraints. This ensures that all offspring represent legal schedules. While we considered extending crossover to include a

randomly chosen individual in addition to the original and mutant (as is common in some differential evolution strategies), we found that the high degree of constraint renders such additional crossover ineffective, as the correction mechanism typically neutralizes any benefit.

4 Results

We evaluate two scheduling scenarios. First, we evaluate on a fully deterministic setting, where the planner (RL, heuristic or evolutionary algorithm) generates a full plan which is then evaluated against the environment's reward function. Second, we compare in a stochastic simulation where workers randomly leave for vacation[1]. Once a worker is on vacation, the environment stops, all jobs are unassigned and the planner is asked to re-plan for the remaining time.

We found a small fixed testset of instances to be unsufficient to evaluate the true performance. This is especially true for the stochastic events, as it is not possible to fully fix random events equally between runs. For instance, imagine an instance with two workers with identical skills: Method A assigning to the first worker and Method B assigning to the second would have different stochastic outcomes, despite the assignment itself being arbitrary. Such random effects compound throughout the plan, yielding to entirely differnt stochatic scenarios. Due to the high inter-instance variability and large search space, we evaluate all methods on 1000 randomly generated instances and report both the mean and standard deviation over the set. In the case of our RL method, we also evaluate an agent trained on deterministic settings (and evaluated on both) and an agent trained on stochastic events (and evaluated on both). All experiments (including the Neural Network) were run on CPU.

We set up the simulation reward (see Section 3.1) such that the maximum reward, i.e., the best possible reward, achievable for deterministic simulations is 500. A reward of 500 corresponds to a fully dense plan where every worker is busy at every timestep. Depending on which exact jobs and worker profiles exist - as well as which stochastic effects are present - the realizable reward may be lower, even if the plan is optimal.

We compare our model against the benchmark algorithms described in Section Section 3.3. We use a 4 layer transformer with 2 heads and a hidden dimension of 128. Following prior work, we use an inverted bottleneck of size 256 for the feedforward [24]. We set the learning rate to $3 \cdot 10^{-4}$, batchsize to 2048, and train for a total number of 8000 epochs, each with 16 instances and 4 rollouts per instance. We found 4 rollouts to yield sufficient stabilization to allow for training, but increasing the number of rollouts does improve the advantage estimation. For the simulation we consider, we found the additional stabilization not worth cost of running more rollouts. Increasing the degree of stochasticity may change this balance.

We report our results in Table 1. The deterministically trained RL agent achieves higher rewards than the heuristics in all cases, except against DAGH in the stochastic setting. The RL agent trained on stochastic simulations obtains the highest rewards overall, significantly outperforming all other methods. Interestingly, it also surpasses the agent trained on deterministic simulations, both in stochastic and deterministic evaluations. While this may seem counterintuitive, it can be attributed to overfitting effects commonly observed in RL, where agents exploit simulator-specific artifacts rather than solving the underlying problem. We hypothesize that training on stochastic simulations acts as a regularizer, improving generalization.

Moreover, both MSPT and DAGH are considerably faster than the RL-based methods (and the GA). This suggests potential for hybrid approaches, where fast heuristics like MSPT or DAGH are used to warm-start or guide RL training, or to generate high-quality initial solutions.

[1] Our modelling of vacations as unforseeable means that "vacation" and "sick leave" can be seen as equivalent

Table 1: Comparison of RL methods and heuristics based on final reward. We report the mean and standard deviation for each algorithm, along with the average runtime in brackets. When stochastic events are enabled, the overall scheduling load decreases, leading to shorter runtimes across all methods.

Method	Deterministic Simulation	Stochastic Simulation
RL Deterministic (ours)	390.04 ± 20.76 (2.52s)	300.59 ± 20.59 (2.50s)
RL Stochastic (ours)	477.57 ± 28.65 (2.64s)	470.03 ± 24.82 (2.51s)
MSPT	282.40 ± 60.56 (0.15s)	254.90 ± 52.22 (0.15s)
DAGH	366.00 ± 46.42 (0.21s)	318.35 ± 42.71 (0.22s)
GA	258.70 ± 51.81 (41.64s)	226.97 ± 47.84 (42.42s)

5 Discussion

Our results show that the RL approach, especially the model trained under stochastic conditions, consistently outperforms the benchmarks in terms of reward. The stochastic-trained agent not only excels in uncertain environments but also generalizes better to deterministic cases, suggesting that exposure to variability acts as a regularizer and reduces overfitting to simulator artifacts. While MSPT and DAGH offer solid results and produce solutions much faster, they lack the adaptability of RL when disruptions occur. The GA, although effective in small cases, suffers from poor scalability and impractically long runtimes for real-time use. The RL framework, though designed for a specialized problem class, could apply to other domains involving dynamic resource allocation and stochastic disturbances, such as healthcare or emergency response. Future work could explore hybrid models that combine the speed of DAGH with the adaptability of RL.

6 Conclusion

This work introduces a novel RL framework for the PC-JSSP under stochastic conditions, using a cross-attentive bipartite graph model for dynamic worker-job assignments. Experiments on realistic tax office scenarios show that the RL approach outperforms heuristics and evolutionary algorithms in terms of reward, particularly in uncertain settings. The strong generalization of the stochastic-trained model to deterministic cases highlights its robustness. These findings encourage applying similar architectures to other scheduling domains and exploring hybrid models that integrate heuristic rules for greater efficiency and interpretability.

Acknowledgments

This study was conducted within the project 'ROLF - Reinforcement Learning für Betriebswirtschaftliche Prozesse', funded by the Bavarian Ministry of Economic Affairs, Regional Development and Energy (program 'BayVFP Förderlinie Digitalisierung').

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] R. S. Sutton, A. G. Barto, Reinforcement Learning: An Introduction, A Bradford Book, Cambridge, MA, USA, 2018.

[2] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis, Human-level control through deep reinforcement learning, Nature 518 (2015) 529–533. doi:10.1038/nature14236.

[3] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Mastering the game of go with deep neural networks and tree search, Nature 529 (2016) 484–489. doi:10.1038/nature16961.

[4] R. Munos, T. Stepleton, A. Harutyunyan, M. G. Bellemare, Safe and efficient off-policy reinforcement learning, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, Curran Associates Inc., Red Hook, NY, USA, 2016, p. 1054–1062.

[5] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, ArXiv abs/1707.06347 (2017). URL: https://api.semanticscholar.org/CorpusID:28695052.

[6] H. F. Song, A. Abdolmaleki, J. T. Springenberg, A. Clark, H. Soyer, J. W. Rae, S. Noury, A. Ahuja, S. Liu, D. Tirumala, N. M. O. Heess, D. Belov, M. A. Riedmiller, M. M. Botvinick, V-mpo: On-policy maximum a posteriori policy optimization for discrete and continuous control, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, 2020.

[7] J. Blauth, N. Klein, M. Nägele, A better-than-1.6-approximation for prize-collecting tsp, in: J. Vygen, J. Byrka (Eds.), Integer Programming and Combinatorial Optimization, Springer Nature Switzerland, Cham, 2024, pp. 28–42.

[8] A. Ahmadi, I. Gholami, M. Hajiaghayi, P. Jabbarzade, M. Mahdavi, Prize-collecting steiner tree: A 1.79 approximation, in: Proceedings of the 56th Annual ACM Symposium on Theory of Computing, STOC 2024, Association for Computing Machinery, New York, NY, USA, 2024, pp. 1641–1652.

[9] N. Bhatt, N. R. Chauhan, Genetic algorithm applications on job shop scheduling problem: A review, in: 2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI), 2015, pp. 7–14. doi:10.1109/ICSCTI.2015.7489556.

[10] A. Türkyılmaz, Özlem Şenvar, İrem Ünal, S. Bulkan, A research survey: heuristic approaches for solving multi objective flexible job shop problems, Journal of Intelligent Manufacturing 31 (2020) 1949–1983. doi:10.1007/s10845-020-01547-4.

[11] B. Waschneck, A. Reichstaller, L. Belzner, T. Altenmüller, T. Bauernhansl, A. Knapp, A. Kyek, Optimization of global production scheduling with deep reinforcement learning, Procedia CIRP 72 (2018) 1264–1269. doi:https://doi.org/10.1016/j.procir.2018.03.212, 51st CIRP Conference on Manufacturing Systems.

[12] K. Żbikowski, M. Ostapowicz, P. Gawrysiak, Deep reinforcement learning for resource allocation in business processes, in: M. Montali, A. Senderovich, M. Weidlich (Eds.), Process Mining Workshops, Springer Nature Switzerland, Cham, 2023, pp. 177–189.

[13] L. Lv, C. Zhang, J. Fan, W. Shen, Deep reinforcement learning for job shop scheduling problems: A comprehensive literature review, Knowledge-Based Systems 321 (2025) 113633. doi:https://doi.org/10.1016/j.knosys.2025.113633.

[14] C. Zhang, W. Song, Z. Cao, J. Zhang, P. S. Tan, X. Chi, Learning to dispatch for job shop scheduling via deep reinforcement learning, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F.

Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1621–1632.
[15] P. Tassel, M. Gebser, K. Schekotihin, A reinforcement learning environment for job-shop scheduling, ArXiv abs/2104.03760 (2021). URL: https://api.semanticscholar.org/CorpusID:233181518.
[16] I. Echeverria, M. Murua, R. Santana, Offline reinforcement learning for job-shop scheduling problems, ArXiv abs/2410.15714 (2024). URL: https://api.semanticscholar.org/CorpusID:273501756.
[17] G. Bonetta, D. Zago, R. Cancelliere, A. Grosso, Job shop scheduling via deep reinforcement learning: A sequence to sequence approach, in: M. Sellmann, K. Tierney (Eds.), Learning and Intelligent Optimization, Springer International Publishing, Cham, 2023, pp. 475–490.
[18] O. Vinyals, M. Fortunato, N. Jaitly, Pointer networks, in: C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 28, Curran Associates, Inc., 2015.
[19] I. Bello, H. Pham, Q. V. Le, M. Norouzi, S. Bengio, Neural combinatorial optimization with reinforcement learning, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings, OpenReview.net, 2017.
[20] A. Momenikorbekandi, T. Kalganova, Intelligent scheduling methods for optimisation of job shop scheduling problems in the manufacturing sector: A systematic review, Electronics 14 (2025). doi:10.3390/electronics14081663.
[21] A. Jamili, Robust job shop scheduling problem: Mathematical models, exact and heuristic algorithms, Expert Systems with Applications 55 (2016) 341–350. doi:https://doi.org/10.1016/j.eswa.2016.01.054.
[22] K. S. N. Ripon, C.-H. Tsang, S. Kwong, An Evolutionary Approach for Solving the Multi-Objective Job-Shop Scheduling Problem, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 165–195.
[23] S. Das, P. Suganthan, Differential evolution: A survey of the state-of-the-art, IEEE Transactions on Evolutionary Computation 15 (2011) 4–31. doi:10.1109/TEVC.2010.2059031.
[24] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Neural Information Processing Systems, 2017. URL: https://api.semanticscholar.org/CorpusID:13756489.
[25] Z. Shao, P. Wang, Q. Zhu, R. Xu, J.-M. Song, M. Zhang, Y. K. Li, Y. Wu, D. Guo, Deepseekmath: Pushing the limits of mathematical reasoning in open language models, ArXiv abs/2402.03300 (2024). URL: https://api.semanticscholar.org/CorpusID:267412607.
[26] M. Pinedo, Scheduling: Theory, Algorithms, Systems, 3rd ed., Springer, Berlin/Heidelberg, 2008.

Data Driven Risk Estimation for FMEA: A Systematic Literature Review

Nils Mayat[1] nils.mayat@bmw.de [a] *
Chris Schönekehs[2] chris.schoenekehs@wzl-iqs.rwth-aachen.de *
Marcos Padrón Hinrichs[2] marcos.padron@wzl-iqs.rwth-aachen.de *
Tim Weisser[1] tim.weisser@bmw.de
Robert H. Schmitt[2,3] 0000-0002-0011-5962

[a]Corresponding author.
*These authors contributed equally.

[1] *BMW AG, 80809 München, Germany*
[2] *Laboratory for Machine Tools and Production Engineering WZL, RWTH Aachen University, 52074 Aachen, Germany*
[3] *Fraunhofer Institute for Production Technology (IPT), 52074 Aachen, Germany*

Failure Mode and Effect Analysis (FMEA) is a widely used methodology across industries for risk identification, evaluation, and prioritization. Despite its utility, traditional FMEA heavily relies on subjective expert judgement and is often time intensive. Recent advancements in data-driven methods present an opportunity to address these limitations by automating and enhancing key aspects of the FMEA process. This systematic literature review aims to identify and evaluate existing data-driven approaches for improving FMEA, specifically in estimating risk factors such as Severity (S), Occurrence (O), Detection (D), and the Risk Priority Number (RPN) using process data. A total of 20 relevant studies were identified and analyzed. The findings reveal that the integration of data-driven methods into the FMEA framework has gained significant traction in recent years. Proposed techniques range from traditional linear mapping models to advanced methods, such as Large Language Models (LLMs). The reviewed studies span a variety of industries, with the automotive and manufacturing sectors emerging as predominant contributors. Among the risk factors, Occurrence (O) has been most frequently addressed using data-driven techniques. The review highlights several avenues for future research, including the integration of diverse data sources and the application of cutting-edge technologies, such as multi-agent systems, to further enhance risk estimation within the FMEA framework.

Keywords
FMEA • Data Driven • AI • ML • SLR

1 Introduction

The ongoing advances in the development of methods of artificial intelligence (AI) open new opportunities for efficiency improvement in production environments. In particular, significant progress has been made in methods for processing text data to handle large amounts of information, summarize them, and to respond to complex queries [1]. In the field of quality and risk management, the potential for data-driven methods to improve traditional methods such as Failure Mode and Effect Analysis (FMEA) has been identified [2]. FMEA is a method for identifying failure modes and their effects, evaluating their risks, and prioritizing them. It has received major attention in recent years due to its widespread usage across industries. This method has been widely

adopted for quality management in automotive production to ensure a product of high quality and reliability, but has been established as a cross-industry method for preventive risk management [3, 4, 5]. Today, the approach is defined by different standards such as the cross-industrial DIN EN 60812 [6]. Due to its simplicity and effectiveness FMEA remains a widely adopted method for risk assessments in the industry, which is also recommended in industrial guidelines [2]. A common guideline consisting of seven steps used predominantly in the European and American automotive industry was established by the Automotive Industry Action Group (AIAG) and the Verband der Automobilindustrie (VDA) in the FMEA Handbook [4]. The guideline is separated in Design FMEA, where potential failure modes of the product are identified and mitigated and the Process FMEA where potential failure modes in production, assembly and logistics are identified and mitigated. The seven steps are the same in Design and Process FMEA. The method relies on experts identifying failure modes and their effects. Therefore, experts from different domains are needed to discuss these possible failure modes, evaluate their severity (S), likelihood (O) and chance of detection (D) in order to rank them according to their risk priority and derive risk mitigation actions [3]. Their S, O and D factors are evaluated on a scale from 1-10 and multiplied to a risk priority number (RPN) which subsequently ranges from 1-1000. The goal is to identify and analyze problems early in the product lifecycle in order to derive and implement measures to minimize failure costs during the life cycle [5]. Due to this dependence of expert knowledge the method is prone to subjectivity, complex, time consuming and expensive. Furthermore, the content of an FMEA is static and the method must be performed regularly to be up to date. For this reason, data-driven approaches promise a beneficial solution to conduct an FMEA, especially in deriving necessary failure modes and determining their severity, number of opportunity and detection likelihood. The recent advances in data-driven methods such as AI promise to supplement the FMEA process to make it more objective and faster. In their systematic literature review (SLR) [2] identified different shortcomings of the FMEA. Among them are the difficulty to accurately assess the risk parameters as well as the transferability of scores varies. Additionally, with methods being able to handle big data and form statements out of it, a continuous risk evaluation in the framework of FMEAs is possible. To tackle these identified short comings different data-driven methods have been proposed.

Building on the contrubition of [2] this SLR tries to answer thew research questions “What data-driven methods are used to determine S, O and D values for FMEAs?”. In order to gain even more insight the research questions “What factors of the RPN are determined based on process data?” as well as “What data is used in data-driven risk assessment for FMEAs?” are being tackled by this SLR.

2 Methods

To address the research questions, a systematic literature review (SLR) was conducted in May 2025. The PRISMA process for SLRs was followed to adhere to scientific standards for literature reviews [7]. After defining a search string, publications were title, abstract and full text screened. The screening process was documented in 1. The following search string was developed to capture a broad range of relevant studies: (“artificial intelligence” OR “AI” OR “machine learning” OR “ML” OR “deep learning” OR “data driven”) AND (“Failure Mode and Effect Analysis” OR “FMEA”). This formulation was intended to encompass recent research at the intersection of data-driven methodologies and FMEA. Only publications from 2015 onwards were considered to ensure the inclusion of recent developments in the application of data-driven methods to FMEA. The search was carried out in four electronic databases: Scopus, IEEE Xplore, Web of Science (WoS), and PubMed. The retrieved citations were imported into Pico Portal (https://picoportal.org/) for screening and further processing. Pico Portal facilitates collaborative screening for SLRs. Du-

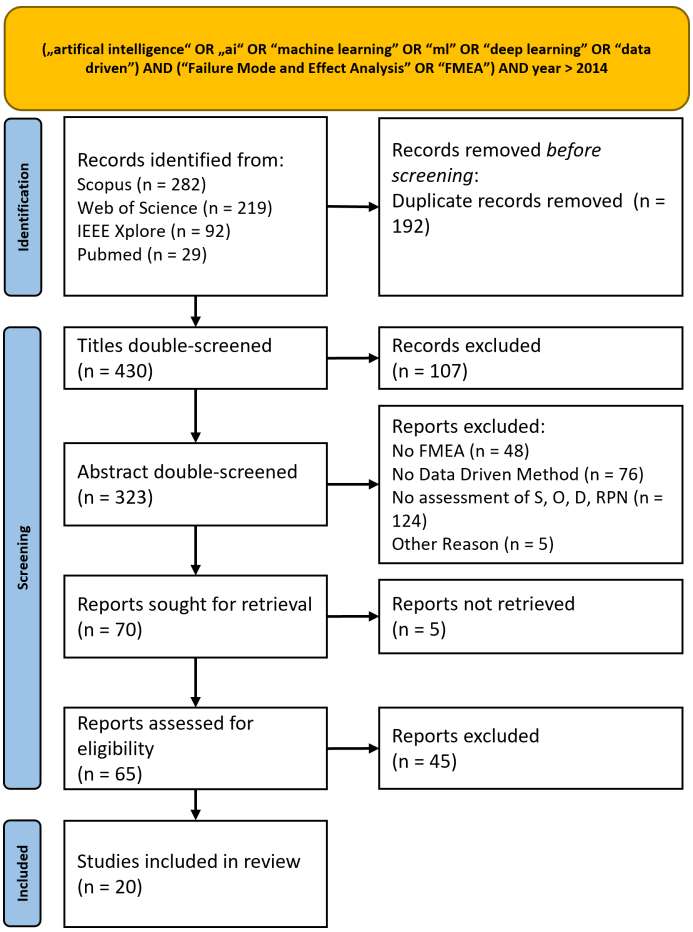


Figure 1: Prisma diagram of the SLR carried out

plicate records were removed prior to a two-stage screening process. In the first stage, titles and abstracts were independently screened by two reviewers. If both reviewers excluded a record but assigned different exclusion reasons, the reason provided by NM, who served as the primary reviewer, was retained. In cases of disagreement regarding inclusion or exclusion decisions, discussions were held until consensus was reached. Full texts of the remaining publications were subsequently retrieved and subjected to single-reviewer full-text screening. Publications meeting the inclusion criteria were then used for data extraction. For each included study, the following information was extracted: publication year, author(s), application domain, type of FMEA (e.g., process FMEA), and the specific risk factor addressed. Additionally, the data-driven method used to determine the risk factor, the source of data (e.g., real-world or simulated), and whether expert knowledge was incorporated were documented. Evaluation methods or metrics and corresponding results were also recorded. Data-driven methods were defined as methods which develop a model to form a statement based on data. This can be AI and ML methods, but we also included simpler approaches such as linear mapping or rule-based decision based on process data.

3 Results

The SLR identified a total of 622 records from four electronic databases: Scopus, IEEE Xplore, Web of Science, and PubMed. The search strategy was designed to capture a broad range of studies

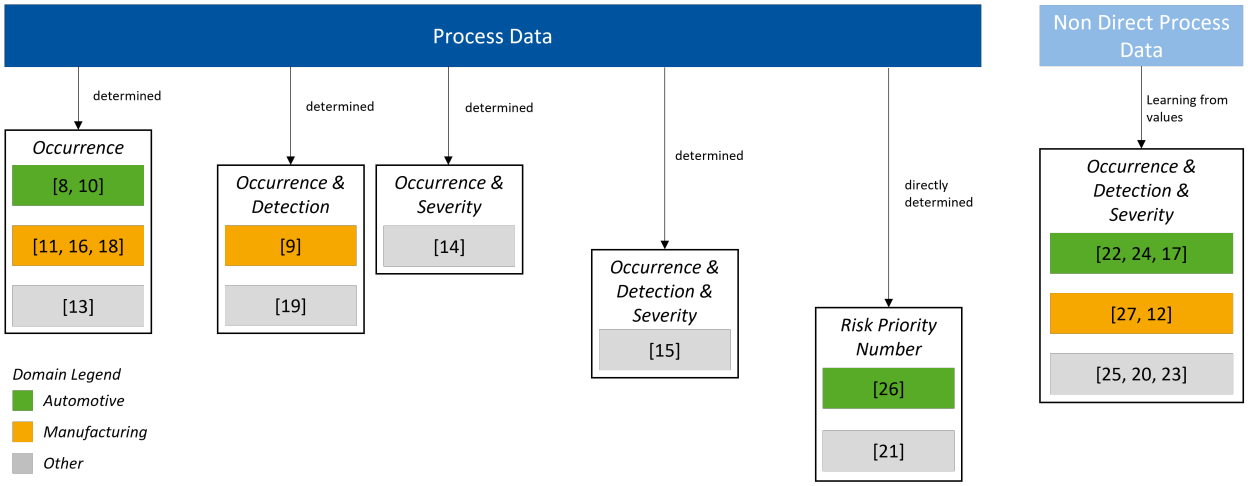


Figure 2: Overview of the risk factors, the underlying data and domains of the publications

at the intersection of AI and FMEA. Following a rigorous screening process, 20 publications were ultimately selected for data extraction. The PRISMA diagram in Figure 1 illustrates the flow of information through the review process, detailing the number of records identified, screened, and excluded, as well as the reasons for exclusion [7]. The main exclusion criteria included the lack of application of FMEA in conjunction with AI for determining *S*, *O*, *D*, or *RPN*.

The SLR identified a total of 622 records from four electronic databases: Scopus, IEEE Xplore, Web of Science, and PubMed. The search strategy was designed to capture a broad range of studies at the intersection of AI and FMEA. Following a rigorous screening process, 20 publications were ultimately selected for data extraction. The PRISMA diagram in Figure 1 illustrates the flow of information through the review process, detailing the number of records identified, screened, and excluded, as well as the reasons for exclusion [7]. The main exclusion criteria included the lack of application of FMEA in conjunction with AI for determining *S*, *O*, *D*, or *RPN*. Specifically, studies that utilized FMEA and data-driven methods but did not focus on the quantification of these risk factors were excluded. The extracted data, summarized in Table 1, provides detailed insights into each study. A notable trend in the publication years of the selected studies indicates a significant increase in research output in recent years. While the review considered publications from 2015 onward, the majority of the accepted studies were published after 2019, with five publications emerging in the year preceding this SLR. Figure 2 presents an overview on what risk factors were determined. Additionally, the figure gives information in which field the publications carried out their investigation. The selected studies come from a broad spectrum of industries, with the manufacturing and automotive industry being the predominant application field of data driven FMEAs. Some publications directly determined the *RPN* instead of each of the risk factors on their own. In Table 1, the risk factor shown is therefore the *RPN* and not the *SOD* [21, 26, 25]. The analysis revealed that the *O* factor was the most frequently determined risk factor across the studies. In instances where only one factor was assessed, it was typically the occurrence factor or the *RPN* value. Notably, studies such as Pathak et al. [14] determined both *S* and *O* factors, while Zhang et al. and Bouzembrak et al. [9, 19] focused on *O* and *D* factors . In cases where *RPN* was calculated, the other risk factors were often determined through expert judgment [11, 13]. Figure 2 depicts the different risk factors determined as well the industries in which the research was carried out. A diverse array of methodologies was employed across the studies, reflecting the complexity of the data-driven approaches utilized. These methodologies ranged from simple linear mappings [8] and rule-based decisions [10] to more sophisticated

Table 1: Included Sources

Source	Risk Factor	Method	Data Format	Data Source	Industry
Geramian et al. (2016) [8]	O	Linear Mapping	Number of Failures	Assembly Of Car Doors	Automotive
Zhang et al. (2016) [9]	OD	Statistical Process Model, Information Entropy	Numerical Simulation and Degradation Data	Blast Furnace	Manufacturing
Belu et al. (2019) [10]	O	Rule-Based Decision	Number of Tests and Energy Consumption	Milling And Pressing of Pistons	Automotive
Pradhan et al. (2020) [11]	O	Probability of Occurence	Number of High-Risk Situations	Simulated Traffic Flow in A Factory	Manufacturing
Sader et al. (2020) [12]	SO, Impact (I) instead of D	Google AutoML (no further specification given)	Claims of Engineers	Agricultural Equipment and Machines	Manufacturing
Filz et al. (2021) [13]	O	Naive Bayes, Generalized Linear Model, Logistic Regression, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees	Weather Data, Airplane Maintenance Data, Flight Profiles and Usage Data	Airplane Maintenance	Aviation
Pathak et al. (2022) [14]	SO	Linear Mapping	Performance Indicators	Mixed-Mode Chromatography	Pharma
Ervural and Ayaz (2023) [15]	SOD	M-CRITIC	Process Data of System Failure Modes	Pasta Production	Food production
Jiang et al. (2023) [16]	O	Logistic Regression	Fault Occurrence	CNC Process Data	Manufacturing
Liang et al. (2023) [17]	SOD	TF-IDF, Word2Vec, k-Means	Social Media Comments	Car Failures	Automotive
Yu et al. (2023) [18]	O	Fuzzy Evaluation	Failure Occurrence	Assembly Process of Machining Center	Manufacturing
Bouzembrak et al. (2024) [19]	OD	Bayesian Network	Data Of Detected Frauds	Spices Supply Chains	Food supply chain
Collier et al. (2024) [20]	SOD	LLM	Recall Data	US Consumer Product Safety Commission Database	Consumer products
Du et al. (2024) [21]	Weights for RPN	Entropy Weighting and Fuzzy-TOPSIS	Reliability, Economy and Maintainability Indicators	Water Supply Components	Water supply
El Hassani et al. (2024) [22]	SOD	LLM	Review Data from Kaggle	Cars	Automotive
Song et al. (2024) [23]	SOD	Average Sentiment Polarity, Frequency, Information Entropy CRITIC Method	Reviews	Hotels in New York City	Hotel industry
El Hassani et al. (2025) [24]	SOD	LLM	Selected Social Media Comments	Car Parts	Automotive
Naranjo et al. (2025) [25]	RPN	Random Forest	Survey Results	Administrative Staff	Administration
Pandya et al. (2025) [26]	RPN	Monocular Depth Estimation	Simulation Data	Automotive Driving	Automotive
Xu (2025) [27]	SOD	LLM	Patent	Electric Scooter	Manufacturing

techniques such as Bayesian networks [19], logistic regression [16], and large language models (LLMs) [20, 22, 24, 27]. The choice of methodology was often influenced by the underlying data sources, which primarily consisted of real-world data derived from the processes under investigation. For instance, studies like Geramian et al. [8] focused on car failures, while Yu et al. [18] examined assembly processes in CNC machines. Only a few studies, such as Pandya et al. [26] have proposed methods that condense process date into a risk factor which can then be used for the calculation of the *RPN* [14, 19, 15]. The evaluation of applied methods varied significantly among studies. Some studies employed train-test splits to compare estimations [13, 16], while others confirmed the plausibility of results through expert validation [24]. Another applied evaluation technique consisted of comparing the ranking of the failure methods to other estimation methods [11, 15]. However, it is noteworthy that some studies did not evaluate the impact of risk factor estimation on the overall FMEA process [14, 10, 24].

4 Discussion

The findings of this SLR highlight significant trends and insights regarding the integration of data-driven methods within FMEAs. The increasing prevalence of AI applications in FMEA, particularly in recent years, underscores a shift in how risk factors are assessed and managed across various industries.

Different methods have been used for data-driven risk assessment. Methods are tailored to the specific use case and the underlying data. However, one of the most notable observations from the review is the rise of LLMs and their ability to streamline FMEA processes. These models have enabled the development of simpler pipelines for risk assessment, often imitating expert judgment. While this advancement presents opportunities for efficiency, it also raises questions about the reliability and validity of AI-generated assessments. The reliance on LLMs to replicate expert decision-making may not fully capture the nuanced understanding that human experts bring to the FMEA process. Therefore, it is crucial to consider the implications of using AI as a substitute for human expertise, rather than as a complementary tool that enhances decision-making.

The analysis revealed that the *O* factor is the most frequently determined risk factor across the studies reviewed. This trend suggests that the identification of potential failures is often prioritized over the assessment of *S* and *D* factors. One reason for this may lie in the relative simplicity of determining occurrence: simple statistical or machine learning (ML) methods, such as linear regression or linear mappings, can be readily applied to historical failure data to estimate *O*. Moreover, the definition of the *O* factor is inherently more interpretable and often directly derivable from available process data, making it particularly suitable for data-driven analysis. The ease of quantifying occurrence may contribute to this focus, as it allows for more straightforward data collection and analysis. However, this emphasis on occurrence could lead to an incomplete understanding of risk, as it may overlook critical insights that could be gained from a more balanced consideration of all risk factors. The sole focus on occurrence could lead to an incomplete understanding of risk, as it may overlook critical insights that could be gained from a more balanced consideration of all risk factors.

The diversity of data sources utilized in the studies indicates a wealth of possibilities for enhancing FMEA through data-driven approaches. The incorporation of real-world data, as opposed to simulated data, reflects a preference for practical applicability in real-world conditions. Additionally, in industries where FMEA is applied to series production, such as the automotive industry, there is the possibility for standardization and scaling of process data. This can even more so amplifies the credibility of model predictions for FMEA. Since FMEA is applied high-risk environments the severity of potential failures underlines the need for reliable and trustworthy

outputs from data-driven methods. Consequently, the quality of input data becomes a critical factor in determining the credibility of model predictions[2]. When data quality is insufficient, the validity of risk estimations may be compromised, thereby reducing the confidence that process owners can place in these models as accurate representations of real-world conditions. One potential approach to mitigate this limitation is the integration of simulated data, which can help fill gaps or validate findings derived from empirical datasets [28, 29]. However, the limited adoption of simulation-based methods in the reviewed studies raises concerns regarding the generalization and robustness of the proposed approaches across diverse operational contexts. Future research should explore the potential benefits of integrating simulated data to complement real-world findings, thereby enriching the overall analysis.

Furthermore, the inclusion of expert knowledge in the FMEA process is not merely a substitution for human input but rather an enhancement of the analytical framework. The studies reviewed demonstrate that expert judgment remains a valuable component in determining risk factors, particularly in cases where data may be sparse or ambiguous. This synergy between AI methodologies and expert insights can lead to more robust risk assessments and ultimately improve decision-making processes. In terms of methodological approaches, the review highlights a variety of techniques employed across the studies. The choice of methodology often reflects the specific context and data sources of each study. However, it is noteworthy that some studies did not evaluate the impact of their risk factor estimations on the overall FMEA process. This gap in evaluation underscores the need for a more comprehensive understanding of how different methodologies influence the effectiveness of FMEA outcomes.

In conclusion, the integration of AI and ML into FMEA presents both opportunities and challenges. While advancements in technology have the potential to enhance risk assessment processes, it is essential to maintain a critical perspective on the role of human expertise and the implications of relying on AI-generated insights. Future research should continue to explore the interplay between AI methodologies and expert judgment, as well as the impact of various data sources and methodologies on the overall effectiveness of FMEA. By addressing these considerations, the field can move towards a more nuanced and effective application of AI in production environments. While this SLR offers valuable insights into the integration of data-driven methods within the FMEA framework, several limitations must be acknowledged. First, the scope of this review was explicitly limited to applications within the context of FMEA. Although FMEA is a widely used methodology for risk assessment, many data-driven techniques identified in this review are equally applicable to related domains such as risk-based testing, reliability engineering, or predictive maintenance. These adjacent areas often share similar goals, namely the identification, quantification, and mitigation of risks yet operate within different methodological frameworks. As a result, valuable contributions from outside the strict FMEA literature may have been overlooked. Second, the review only considered studies published since 2015. This decision was made to capture recent developments in AI and ML, but it may have led to the exclusion of earlier works that could offer foundational insights, particularly those exploring the integration of statistical or rule-based approaches with FMEA. Given the long-standing history of FMEA since the 1960s, relevant pre-2015 contributions might still hold significant methodological value. Third, the findings underscore that the application of data-driven approaches to FMEA remains relatively limited. Despite the growing interest in leveraging data-driven methods for risk assessment, their integration into the well-established FMEA methodology is not yet widespread. This limited adoption restricts the empirical basis on which to evaluate the overall effectiveness and impact of such approaches on real-world FMEA outcomes. Finally, while the review captures a variety of data-driven methods and data sources, it must be noted that the quality and completeness of information reported in the original studies varied. In particular, many studies did not systematically evaluate the influence of their proposed methods on the broader FMEA process, such as decision-making or mitigation planning. As such, conclusions regarding the practical

utility and implementation readiness of data-driven FMEA methods should be drawn with caution.

5 Conclusion and Outlook

In this SLR, the goal was to identify data-driven methods to complement and improve the FMEA process. Relevant literature was selected and screened using the PRISMA method and 20 relevant publications were identified eventually. The goal was to identify the methods used in the papers and to which of the relevant FMEA factors they account. As the most prominent factor *O* was identified which suggests that the identification of potential failure modes is often prioritized, independent from the used method or the application domain. From the recently published studies, it can be concluded that the utilization of data-driven and AI methods will continue to increase, not only in manufacturing but also in enhancing FMEAs. Advances in data-driven risk modeling and AI, particularly the emergence of LLMs offer significant potential for integrating expert knowledge with diverse data sources. This integration can address traditional limitations of the FMEA process, such as subjectivity and time intensity, by enabling more objective and efficient risk assessments. Furthermore, the combination of expert knowledge and data-driven methods provides an opportunity to transition from static to dynamic FMEAs, where risk evaluations are updated in near real-time based on evolving operational data. However, achieving this transformation requires overcoming several challenges. The reliability and validity of data-driven approaches in high-stakes applications must be ensured, particularly in industries such as aerospace and healthcare, where the consequences of failure are critical. Future research should focus on enhancing the robustness of these approaches, for example, through the integration of simulation-based data to complement empirical datasets, thereby increasing the generalization and applicability of models across diverse operational contexts. Another promising avenue for future exploration is the standardization of data collection and processing practices to ensure compatibility and consistency across industries. The adoption of standardized frameworks could facilitate the scalability of data-driven FMEA approaches and support their broader implementation in industries beyond automotive and manufacturing, such as energy, transportation, and pharmaceuticals. Additionally, as the volume of process data grows, the development of methods to handle large-scale, heterogeneous datasets will become increasingly critical. Incorporating advancements such as multi-agent systems and collaborative AI frameworks may further enhance the ability to model complex interactions and provide more nuanced risk assessments. Multi-agent systems might be able to imitate an expert group in the FMEA process. Here agents trained or with access to data from different departments and technical backgrounds might identify and prioritize risk in a more profound way. Agents acting on different system level could aggregate risk and provide risk mitigation strategies for the complete system. Finally, while data-driven methods offer significant potential for improving FMEA, human expertise will remain indispensable, particularly in cases where data is sparse or ambiguous. Future efforts should focus on designing hybrid frameworks that optimize the collaboration between AI systems and domain experts. Such systems would ensure that AI serves not as a replacement but as a complement to human judgment, thereby enhancing the overall reliability and interpretability of risk assessments. By addressing these challenges and opportunities, the integration of data-driven methods into the FMEA framework can evolve into a transformative approach, enabling more effective, dynamic, and scalable risk assessment processes across a wide range of industries.

Declaration on Generative AI

During the preparation of this work, the authors used Chat-GPT-4 in order to: Grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] R. Gupta, S. Tiwari, P. Chaudhary, Large language models, in: Generative AI: Techniques, Models and Applications, volume 241 of *Lecture Notes on Data Engineering and Communications Technologies*, Springer, Cham, 2025. doi:10.1007/978-3-031-82062-5_5.

[2] A. Chakhrit, N. E. H. Benharkat, I. H. M. Guetarni, A. Guedri, M. Chennoufi, A literature review of risk assessment approaches in failure mode and effects analysis (2025). URL: https://www.emerald.com/insight/content/doi/10.1108/IJQRM-01-2025-0042/full/html. doi:10.1108/IJQRM-01-2025-0042.

[3] S. E. Schwarz, C. Thümmel, M. Kuebler, B. Doostkam, A. Albers, Fields of action in the continuous and early validation of product profiles in product engineering, in: Proceedings of the International Conference on Industrial Engineering and Operations Management, IEOM Society International, Augsburg (Greater Munich), Germany, 2024. doi:10.46254/EU07.20240226.

[4] AIAG & VDA, FMEA-Handbook. Design FMEA, Process FMEA, Supplemental FMEA for Monitoring & System Response, first edition ed., 2019.

[5] P. L. Belcaro, Fmea: Der operative-verfahrensseitige aspekt, in: 9-Object-Model FMEA, Springer Vieweg, Wiesbaden, 2025. doi:10.1007/978-3-658-45745-7_4.

[6] Din en 60812 analysis techniques for system reliability - procedure for failure mode and effects analysis (fmea) (iec 60812:2006); german version en 60812:2006 (2006).

[7] M. J. Page, et al., The prisma 2020 statement: an updated guideline for reporting systematic reviews, BMJ (Clinical research ed.) 372 (2021) n71. doi:10.1136/bmj.n71.

[8] A. Geramian, M.-R. Mehregan, N. Mokhtarzadeh, M. Hemmati, Fuzzy inference system application for failure analyzing in automobile industry, International Journal of Quality & Reliability Management 34 (2017) 1493–1507. doi:10.1108/IJQRM-03-2016-0026.

[9] J. Zhang, C. Hu, X. He, X.-s. Si, D. Zhou, Risk evaluation for deteriorating systems with accuracy analysis of parameter estimation, 2016, pp. 1–6. doi:10.1109/PHM.2016.7819939.

[10] N. Belu, L. Ionescu, N. Rachieru, Risk-cost model for fmea approach through genetic algorithms – a case study in automotive industry, IOP Conference Series: Materials Science and Engineering 564 (2019) 012102. doi:10.1088/1757-899X/564/1/012102.

[11] N. Pradhan, P. Balasubramanian, R. Sawhney, M. H. Khan, Automated risk assessment for material movement in manufacturing, Gestão & Produção 27 (2020) e5424. doi:10.1590/0104-530X5424-20.

[12] S. Sader, I. Husti, M. Daróczi, Enhancing failure mode and effects analysis using auto machine learning: A case study of the agricultural machinery industry, Processes 8 (2020) 224. doi:10.3390/pr8020224.

[13] M.-A. Filz, J. Langner, C. Herrmann, S. Thiede, Data-driven failure mode and effect analysis (fmea) to enhance maintenance planning, Computers in Industry 129 (2021) 103451. doi:10.1016/j.compind.2021.103451.

[14] M. Pathak, P. Pokhriyal, I. Gandhi, S. Khambhampaty, Implementation of chemometrics, design of experiments, and neural network analysis ..., Biotechnol Prog 38 (2022) e3252. doi:10.1002/btpr.3252.

[15] B. Ervural, H. Ayaz, A fully data-driven fmea framework for risk assessment on manufac-

turing processes using a hybrid approach, Engineering Failure Analysis 152 (2023) 107525. doi:10.1016/j.engfailanal.2023.107525.

[16] S. Jiang, Z. Liu, J. Chen, A dynamic failure mode and effect analysis (fmea) method for cnc machine tool in service, Journal of Physics: Conference Series 2483 (2023) 012047. doi:10.1088/1742-6596/2483/1/012047.

[17] D. Liang, F. Li, X. Chen, Failure mode and effect analysis by exploiting text mining and multi-view group consensus for the defect detection of electric vehicles in social media data, Annals of Operations Research 340 (2023). doi:10.1007/s10479-023-05649-z.

[18] L. Yu, T. Zhang, H. Tian, Z. Yang, A. Liu, F. Zhang, Fmeca of cnc machine tool design stage based on cbwm and dea, Quality and Reliability Engineering International 40 (2023) 154–169. doi:10.1002/qre.3312.

[19] Y. Bouzembrak, N. Liu, W. Mu, A. Gavai, L. Manning, F. Butler, H. J. P. Marvin, Data driven food fraud vulnerability assessment using bayesian network: Spices supply chain, Food Control 164 (2024) 110616. doi:10.1016/j.foodcont.2024.110616.

[20] Z. A. Collier, R. J. Gruss, A. S. Abrahams, How good are large language models at product risk assessment?, Risk Analysis (2024) 1–24. doi:10.1111/risa.14351.

[21] X. Du, D. Wu, J. Gai, An improved fmea method combining component importance and fuzzy topsis, in: Proceedings of SRSE 2024, 2024, pp. 264–271. doi:10.1109/SRSE63568.2024.10772502.

[22] I. El Hassani, T. Masrour, N. Kourouma, D. Motte, J. Tavčar, Integrating large language models for improved failure mode and effects analysis (fmea): a framework and case study, Proceedings of the Design Society 4 (2024) 2019–2028. doi:10.1017/pds.2024.204.

[23] W. Song, W. Rong, Y. Tang, Quantifying risk of service failure in customer complaints: A textual analysis-based approach, Advanced Engineering Informatics 60 (2024). doi:10.1016/j.aei.2024.102377.

[24] I. El Hassani, T. Masrour, N. Kourouma, J. Tavčar, Ai-driven fmea: integration of large language models for faster and more accurate risk analysis, Design Science 11 (2025) e10. doi:10.1017/dsj.2025.7.

[25] J. E. Naranjo, J. Alban, M. Balseca, D. Villagómez, M. Falconi, M. Garcia, Enhancing institutional sustainability through process optimization: A hybrid approach using fmea and machine learning, Sustainability 17 (2025) 1357. doi:10.3390/su17041357.

[26] M. A. Pandya, P. C. Siddalingaswamy, S. Singh, Fmea-based safety analysis of monocular depth estimation for autonomous vehicles, in: 2025 International Conference on Artificial Intelligence and Data Engineering (AIDE), Nitte, India, 2025, pp. 907–912. doi:10.1109/AIDE64228.2025.10987392.

[27] M. Xu, Enhancing fmea with chatgpt: Structured outputs, qualitative evaluations, and ai-human hybrid fmea, 2025, pp. 1–6. doi:10.1109/RAMS48127.2025.10935221.

[28] P. Schlegel, D. Buschmann, M. Ellerich, R. H. Schmitt, Methodological assessment of data suitability for defect prediction, Quality Innovation Prosperity 24 (2020) 170–185. doi:10.12776/qip.v24i2.1443.

[29] S. Hao, et al., Synthetic data in ai: Challenges, applications, and ethical implications, https://arxiv.org/abs/2401.01629, 2024. ArXiv preprint arXiv:2401.01629.