TechMB: Exploring the Potential of Vision Language Models for Interpreting Technical Drawings

Leonhard Kunz[1,*], Mario Klostermeier[1], Kokulan Thanabalan[2], Tatjana Legler[1], Martin Ruskowski[1]

[1] University of Kaiserslautern-Landau (RPTU), Gottlieb-Daimler-Straße 42, 67663 Kaiserslautern, Germany
[2] Trier University, Campus II, Behringstraße 21, 54296 Trier, Germany

** Korrespondierender Autor:*
Leonhard Kunz
Gottlieb-Daimler-Straße 42,
67663 Kaiserslautern,
Germany
☎ *+49 631/205757009*
✉ *leonhard.kunz@rptu.de*

Abstract

Vision Language Models (VLMs) have gained widespread adoption among end users. Their versatility has also sparked interest in applying them to more domain-specific challenges. This paper investigates the principal suitability of small-scale VLMs in the task of evaluating the manufacturability of parts based on a technical drawing by providing the Technical drawings for Manufacturability Benchmark (TechMB). A selection of small-scale VLMs is then tested using this benchmark. The results indicate that the models show potential for text extraction and interpretation of domain-specific terminology. However, they struggle with the reasoning about the manufacturing of the depicted parts and partly even with the delivery of concise and precise answers necessary for the targeted task.

Keywords

Benchmark, Visual Question Answering, CAD, Design for Manufacturability

 | DOI: 10.35199/dfx2025.19

1. Introduction

Design for Manufacturability (DfM) ensures that product designs can be efficiently and reliably produced. However, applying DfM best practices remains largely manual and experience-driven. Less experienced designers often fall into time-consuming feedback loops involving production and quality assurance [1]. Here, automated validation tools can provide immediate feedback on design violations, improving efficiency [2].

These tools typically analyze product geometries based on predefined rule sets. Commercial CAD-integrated software such as DFMPro [3] and the Boothroyd-Dewhurst DFMA suite [4] identify violations like small radii or deep pockets directly within the geometry. Although robust and deterministic, rule-based approaches are limited by their static nature and often require manual adaptation to different manufacturing environments [5].

More recently, AI-based approaches have gained attention. Machine learning models are data-driven methods that can capture subtle and context-specific patterns in manufacturing. However, they require large efforts in dataset preparation and may produce incorrect results due to their probabilistic nature. As a result, such models are still primarily found in research or early-stage applications and are not yet widely used in industry [6,7].

This paper investigates the capabilities of small Vision Language Models (VLMs), referred to as VLMs, in assessing product designs regarding DfM best practices. For this, it introduces the *Technical drawings for Manufacturability Benchmark (TechMB)* for visual understanding and reasoning on technical drawings and tests it on a selection of small-scale VLMs.

The structure of the paper is as follows: Section 2 reviews related work. Section 3 describes the proposed benchmark. Section 4 presents and discusses the results using several open-weight VLMs. Section 5 concludes the paper.

2. Related Literature

DfM validations ensure that a product design is compatible with cost-effective and efficient manufacturing processes. Traditionally, rule-based systems are used for this task [5]. However, recently data-driven alternatives gained traction due to progress in machine learning [6]. The work presented here builds the groundwork for the project XDP-Opt, which attempts to utilize a VLM to analyze technical drawings to detect problems with learned DfM rules [8]. Therefore, in the following, we will firstly assess the advantages and drawbacks of rule-based and data-driven DfM methods. Secondly, we will analyze the functionality and benchmarking of VLMs.

2.1. Comparing rule-based and data-driven approaches to DfM validations

The formulation of DfM rulesets offers several clear advantages. Their deterministic nature yields reliable validation results, as each check follows a precisely defined logic derived from expert knowledge, industry standards, or fundamental manufacturing constraints. Because rules can be encoded directly from specialists' experience, their decision paths remain transparent, easily audited, and relatively simple to trace [9]. This immediate feedback loop helps designers iterate quickly and ensures that process-specific considerations are taken into account from the outset [11].

However, to capture all relevant factors (e.g., materials, geometries, tool capabilities), rules inevitably become complex and highly parametric. Even then, they remain strictly limited to the cases anticipated during rule definition [9, 10]. Any novel feature or emerging manufacturing technique requires experts to update the rulebase accordingly. Maintaining and extending a comprehensive ruleset therefore demands deep specialist knowledge of both the underlying manufacturability constraints and the rule-definition formalism itself [12]. From an organizational standpoint, that also means instituting clear governance over who may modify

the rules to avoid conflicting or redundant entries. As a result, rule-based systems can become cumbersome to keep maintained, particularly in fast-evolving production environments [13].

Historically, most commercially available DfM software is currently built around a rule-based core. They are well integrated with common CAD systems and process the 3D model of the product to simplify validation checks [3, 4].

By contrast, data-driven DfM validation begins from a very different premise. Here, statistical models learn implicit design-to-manufacturability relationships from well-documented artifacts, such as root-cause-analysis reports, version histories with annotated change rationales, and process-parameter logs [14]. However, the high nonlinearity of machine learning models often renders their outputs opaque, reducing their interpretability [15]. Moreover, to establish robust correlations between geometric features and manufacturability outcomes, large volumes of high-quality, representative data must be collected and curated. This data-acquisition effort typically represents the largest investment in any purely data-driven approach [16].

However, particularly machine learning models can adapt automatically to complex, non-linear relationships in the design space, often detecting subtle interactions that simple rules cannot capture [17]. As new designs and production data accumulate, these systems can continuously integrate new correlations through novel data without requiring manual intervention from experts [18]. Table 1 offers a quick summary and a broad comparison of both approaches.

Table 1: Comparative summary of rule-based and data-driven DfM validations

Aspect	Rule-based approach	Data-driven approach
Knowlede Source	Expert-defined rules and standards	Historical data
Flexibility	Limited to predefined cases	High adaptability to new patterns
Maintenance	Manual updates by specialists	Continuously from novel data
Transparency	Well comprehendible through explicit formulation	Potentially difficult to comprehend through high model non-linearity
Data Requirements	Minimal	Substantial high-quality data

In practice, first commercial DfM software suites now incorporate machine learning models to extend the capabilities of rule-based evaluations. They are, e.g., enhancing the weighting of errors based on criticality, learning to detect critical past design patterns, or rule-prefiltering [19, 20]. Nonetheless, many applications are currently targeted at printed circuit board design and are closely integrated with proprietary tool chains.

Agnostic DfM validation tools that leverage the artifacts of the design process (e.g., CAD files, technical drawings, etc.) that are utilizing the strengths of data-driven approaches are sparse and not yet in commercial use [17]. To leverage these artifacts, they must contain all definitions relevant for manufacturing. Technical drawings are still the de facto standard in the communication between design and production in most manufacturing environments. Even though their highly standardized form carries all relevant information, they are seldomly considered as the basis for DfM validation, as their human-readable format makes them inherently difficult to be processed algorithmically [21]. However, advancements in computer vision and machine learning are beginning to address this challenge by segmenting and interpreting components within engineering drawings [21].

2.2. Vision Language Benchmarks for the manufacturing domain

Multimodal foundation models represent a class of large machine learning models that are pre-trained on general data with different modalities (e.g., text, images, audio). Most of these models are based on transformer architectures and have specialized encoders for specific modalities. VLMs specifically are built to process and align visual inputs and natural language. They can be used for a wide array of tasks, such as image captioning, visual question answering, visual reasoning, or cross-modal retrieval [22].

The majority of VLMs are trained on a range of data sources, mainly obtained through webscraping. These cover a broad range of all topics and image/text types available on the internet [23]. The availability of benchmarks consisting of domain-specific datatypes and tasks is essential to assessing the usefulness of different models for specific applications. They help to determine the most suitable model for further fine-tuning on domain-specific data. As such, the benchmark dataset needs to represent the underlying task and datatypes well enough to allow such a conclusion [24].

Several benchmark datasets for VLMs have been proposed in the past few years covering the engineering domain [25, 26, 27]. Some more general benchmarks try to explore the physical reasoning capabilities of VLMs [28] or datatype-specific reasoning on tables or diagrams [29]. However, the manufacturing domain is currently still underexplored in this field. Specifically, there are currently no public benchmarks for the evaluation of technical drawings. And data on visual reasoning tasks regarding DfM assessments based on technical drawings are currently a research gap.

3. Benchmark

In this paper, we address the lack of available vision language benchmarks on technical drawings. The proposed TechMB benchmark is derived from the chain of thought behind manual DfM validations. Thus, we will first analyze the chain of thought behind DfM and collect and label the dataset accordingly. The TechMB benchmark is made publicly available on Huggingface[1].

3.1. Chain of thought DfM validation

Score-based manufacturing evaluation systems are a good starting point to determine the relevant indicators for manufacturability. Typically, these involve an assessment of the overall part definitions, like material declarations and tolerance guidelines. Then an identification of the geometric features and the allocation of suitable manufacturing processes. The manufacturability is then assessed based on feature parameters and process or tool constraints [30]. Derived from these considerations, we propose the following chain of thought:

1. What is the material declaration?
2. What general tolerance schemes are used?
3. What are the overall dimensions of the part?
4. What geometric features need to be manufactured?
5. What is the primary manufacturing process?
6. Does the choice of material, tolerances, and the overall geometric features lead to manufacturability problems for the choice of manufacturing process?

[1] https://huggingface.co/datasets/WSKL/techmb

The questions in the TechMB dataset are based in content on the questions formulated above.

3.2. Data Collection

The data for TechMB consists of manually created and annotated technical drawings. These technical drawings are created manually from the Fusion 360 Gallery segmentation dataset, which contains a wide variety of 3D models submitted by users of the CAD package Autodesk Fusion 360 [31]. The dataset was created to train and test models for segmentation of geometries in the associated modeling operations, like extrusions, revolutions, fillets, chamfers, ect. As such, it contains a high variety of objects of different topological and operational complexity. The TechMB dataset uses a random subset of 180 models of the Fusion 360 segmentation dataset (Version s2.0.1). The technical drawings are created from these manually by 6 different authors, each a (bachelor or master) student of mechanical engineering with entry-level experience in CAD modeling and the creation of technical drawings. The drawings are created using 8 different drawing templates with a variety in the arrangement and content of the title blocks. Nonetheless, the parts in [31] come without a functional context of what the parts are designed for. Thus, the definition of functional tolerances (default as well as specific) or materials is chosen freely. Figure 1 illustrates the process of dataset creation, of which step 3 presents most of the manual efforts.

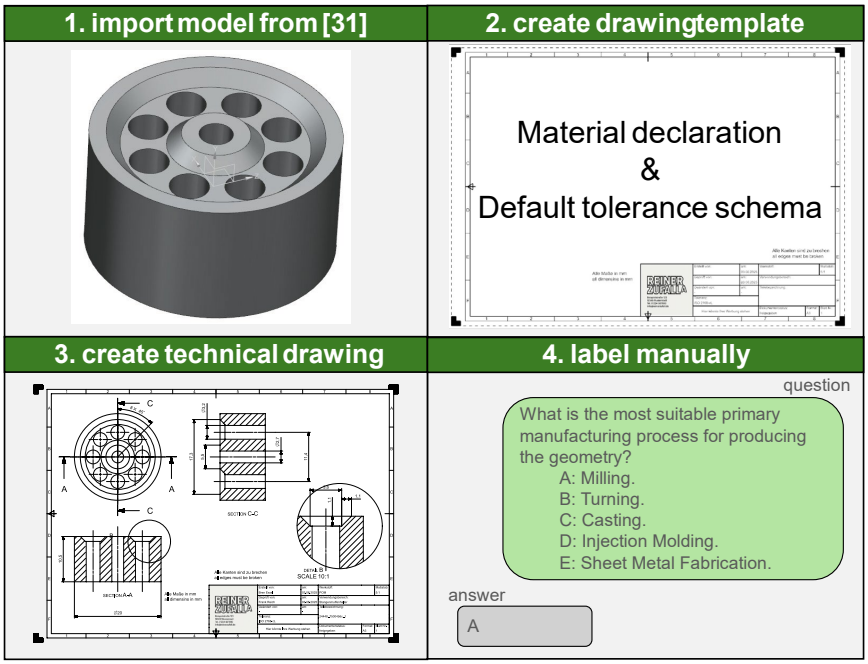

Figure 1: Dataset creation process, including all manual efforts.

3.3. Data Labeling

The labels were assigned manually without being cross-checked by a second person. Some of them represent objective ground truths (e.g., direct text recognition); some labels are not unambiguous (particularly for visual reasoning tasks). Two strategies are employed to circumvent this problem as far as possible. The first strategy is the reformulation of a question as multiple choice. Thus, constraining the possible answer space. This helps the human annotator to choose the most likely answer and at the same time simplifies the automatic evaluation of the generated answer of a VLM. The second strategy separates reasoning and recognition tasks by transferring the problem to an example case and using information from the image. Thus, visual understanding can be tested, and the answer is specifically funneled

to an unambiguous case. An example is the recognition of general tolerances from the technical drawing and their application to a generic length. Figure 2 shows examples of the second tactic. The application of these tactics also helps to avoid debated metrics for the answer correctness, like BLEU or ROUGE scores. Nonetheless, not all questions could be successfully applied to all drawings if the ambiguity was still too high to allow the determination of a subjective ground truth.

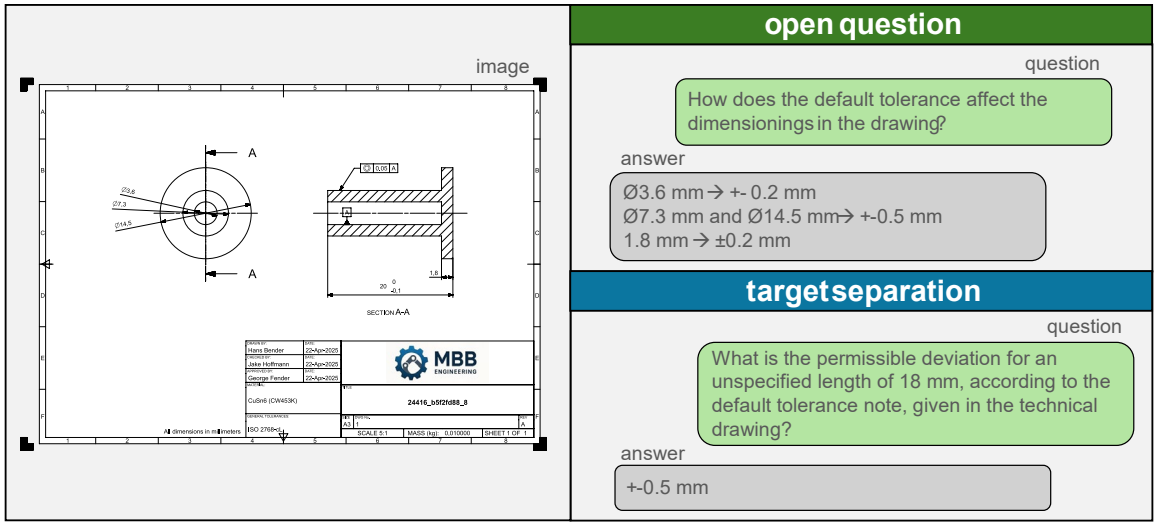

Figure 2: One question reformulation strategy applied to an example. It shows the separation of detection and reasoning onto a different target (the unspecified length of 18 mm). It simplifies the expected answer structure, decreases ambiguity, and makes it not necessary to extract both, dimensions as well as the default tolerance specification.

We furthermore distinguish the tasks for the VLM in the TechMB dataset into Optical Character Recognition (OCR), Visual Understanding (VU), and Visual Reasoning (VR). OCR means the simple recognition and extraction of specific text from the drawing. VU is the comprehension and understanding of visual elements in the drawing, like geometric features or the interpretation of tolerance specifications in domain-specific terminology. VR then declares the interpretation of the drawing as a whole of the depicted part and the reasoning of, e.g., a suitable manufacturing process. Table 2 shows the division of the quest dataset regarding the task as well as the answer evaluation type. The latter includes exact matches of ground truth strings and given answer (e.g., the material declaration: “POM”), and multiple-choice questions. Altogether, the TechMB dataset contains 947 question-answer pairs.

Table 2: Questions with their associated types and the count of each question and type in the TechMB Dataset.

Task ID	Type	Question	Count	
t031	OCR	What is the material specification of the depicted part?	180	360 (38%)
t033		What is the default tolerance schema for unspecified dimensions?	180	
t112	VU	What is the permissible deviation for an unspecified length of 18 mm, according to the default tolerance note, given in the technical drawing? A: 0.01 mm.\nB: 0.02 mm.\nC: 0.05 mm.\nD: 0.1 mm.\nE: 0.2 mm.\nF: 0.5 mm.\nG: 1.0 mm\nH: no specification.	178	438 (46%)
t214		What is the maximum extent of the depicted part along each dimension (width, height, and depth)? \A: [27.3;1.7;1.1].\nB:[21.8;2.9;1.5].\nC:[21.4;1.8;1.2].\nD:[20.0;2.2;1.2].	155	
t215		Which of the following geometric features are present in the part? A: Through-hole.\nB: Blind hole.\nC: Pocket.\nD: Chamfer.\nE: Fillet.	105	

t311	VR	What is the most suitable primary manufacturing process for producing the geometry? A: Milling.\nB: Turning.\nC: Casting.\nD: Injection Molding.\nE: Sheet Metal Fabrication.	134	149 (16%)
t323		Does the choice of material, tolerances, and the overall geometric features lead to manufacturability problems in the choice of manufacturing process? A: Sharp internal corner that cannot be milled.\nB: Overly tight tolerance on a non-critical surface.\nC: Unreachable hole for drilling.\nD: Bending radius to small.\nE: Hole depth exceeds tool length or depth-to-diameter ratio is too high.\nF: Thin wall that may deform during machining or casting.\nG: Threaded hole too close to edge, risking tool breakage or deformation.\nH: Chamfer or fillet are missing on critical mating edges.\nI: Material incompatible with specified feature.\nJ: No manufacturability issue.	15	

The TechMB dataset is provided in tabular form to simplify iterative processing. In conclusion, the final dataset contains the following fields in every row:

- **task_id:** ID of the specific question as declared in Table 2.
- **eval_type:** Classifier for the expected answer type (answer matching or multiple/single choice).
- **drw_id:** ID of the corresponding drawing, which is the same as the part ID in the Autodesk Fusion 360 dataset [31].
- **image:** Bit64-encoded image of the exported technical drawing.
- **drw_complexity:** Numeric complexity of the drawing.
- **question:** The question text.
- **answer:** The expected answer corresponding to the answer type.
- **label_confidence:** The confidence of the assorted labels in manual labeling (low, medium, high).

4. Results and Discussion

The following section describes a benchmark test conducted with 15 VLMs from different model families implemented in the VLMEval package [32]. The tested VLMs are a selection of open-weight models below 8 billion parameters, excluding larger models and service models available only via their APIs. This selection is drawn from a class of “desktop” VLMs that can be deployed locally, enabling direct integration into CAD applications. The selection of models in this benchmark prioritizes diversity in both architectural design and parameter count. The objective is to establish a broad comparative foundation that reflects a wide range of model families. Multiple versions of specific model series are included. Within each series, different model sizes are selected to enable a detailed analysis of scaling effects. Model evaluation is conducted in a zero-shot setting. Each model receives a technical drawing accompanied by a single, isolated question, without any additional example data. Questions are presented independently, preventing the models from carrying information across tasks or forming a coherent representation of the drawing. This setup allows for a focused assessment of each model’s ability to answer individual questions accurately. To standardize the interaction and facilitate automated evaluation, each question is embedded within a uniform system prompt:

*# **Context:***
You are an expert in analyzing technical drawings and manufacturing processes. Your task is to analyse technical drawings of components and answer different questions regarding the content and the validity of the depicted component.
*# **Task***
Analyse the technical drawing provided and answer the following question for the drawing and the depicted part:

{question}
*# **Response Format****
{answer format}
If you are not absolutely certain about the correct answer, return instead ???.
Limit your answer strictly to the given format and do not give any additional explanations.

This system prompt provides additional context and explicitly defines the expected response format, whether as free-text or multiple-choice. This structure improves consistency across responses and ensures compatibility with automated evaluation methods. Automatic evaluation is performed using predefined reference answers and is subsequently verified through manual review. Responses are classified in binary terms as either correct or incorrect, with no partial credit assigned. Final performance scores are calculated as the mean accuracy across all questions for each model.

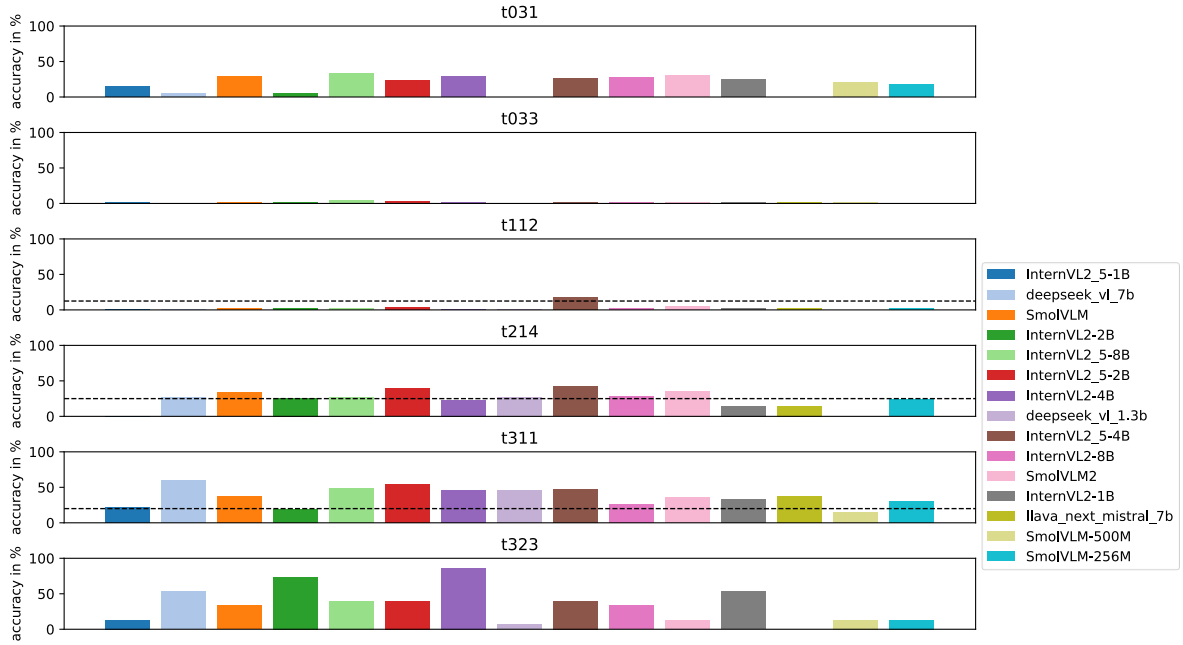

Figure 3: Accuracy per question for all evaluated visual language models (VLMs). The dashed line in tasks t112, t214, and t311 indicates the performance of a random guesser.

The results are summarized in Figure 3, which presents the accuracy achieved per question for each model. For the single-choice questions t112, t214, and t311, a dashed line indicates the expected performance of a random guesser, serving as a baseline for comparison. Notably, several models perform below this random baseline for questions t112 and t214, suggesting a failure to produce meaningful or valid outputs. No model demonstrates consistently strong performance across the full set of questions. From the perspective of practical deployment, the current quality of responses remains inadequate. Interestingly, for question t323, which involves visual reasoning and is among the more complex tasks, several models achieve relatively high accuracy. This suggests that some models are capable of handling isolated instances of high cognitive demand. Nonetheless, the overall findings indicate that small, locally deployable models, that have not been specifically fine-tuned, are presently limited in their ability to provide reliable user assistance in the investigated tasks. It should also be acknowledged that alternative prompt formulations may yield substantially improved performance.

5. Conclusion

The presented work introduces the TechMB dataset as a benchmark for the manufacturability evaluations of parts based on their technical drawings. The drawings are manually created from 3D models from the Autodesk Fusion 360 Segmentation dataset. They are manually annotated based on several questions regarding the content of the drawing and individual manufacturability reasoning. TechMB contains in total 947 question-answer pairs targeting simple text feature extraction or more complex understanding of the drawing as well as reasoning on the depicted part. The study includes a benchmark test with 15 VLMs up to 8 billion parameters. The results indicate that the out-of-the-box performance of the tested open-weight models is not sufficient for this specific task.

Future tests with larger models might show improved performance but would defeat the point of a local application integrated into the CAD environment. Iterative prompt engineering could potentially mitigate particularly problems with the generation of concise and viable answers. Task specific fine-tuning could also help to achieve a higher answer quality but would potentially require high amounts of specific data that is currently not available publically. Nonetheless, the fine-tuning could also involve a broader range of contextual data, like extracts from textbooks, or written DfM guidelines. This kind of information could also be utilized using Retrieval Augmented Generation Pipelines. Lastly, frameworks like Federated Learning could help to open the door to a privacy-preserving usage of private data and therefore mitigate the limits of a fine-tuning strategy.

Acknowledgement

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projektnummer (543073350) - Acronym: XDP-Opt. The following people have enabled this work by creating technical drawings Niklas Krell, Pascal Wohnsiedler, Matthias Bohr, Eljas Völker, Nico Conrad. We are thankful for their support.

References

[1] Alam, Md Ferdous et al..: From Automation to Augmentation: Redefining Engineering Design and Manufacturing in the Age of NextGen-AI. In: *An MIT Exploration of Generative AI* (2024)

[2] HCL Software: *HCL DFMPro for leading Aerospace and Defense Manufacturer*. URL https://dfmpro.com/case-study/dfmpro-for-aerospace-industry/. - abgerufen am 2025-07-11

[3] HCL Software: *Powerful Design for Manufacturing Software | HCL DFMPro*. URL https://dfmpro.com/. - abgerufen am 2025-07-11

[4] Boothroyd Dewhurst: *DFMA® Software: Design for Manufacture and Assembly*. URL https://www.dfma.com/. - abgerufen am 2025-07-11

[5] Campi, Federico; Favi, Claudio; Germani, Michele; Mandolini, Marco: CAD-integrated design for manufacturing and assembly in mechanical design. In: *International Journal of Computer Integrated Manufacturing* Bd. 35 (2022), Nr. 3, S. 282–325

[6] Zhang, Ying; Yang, Sheng; Dong, Guoying; Zhao, Yaoyao Fiona: Predictive manufacturability assessment system for laser powder bed fusion based on a hybrid machine learning model. In: *Additive Manufacturing* Bd. 41 (2021), S. 101946

[7] Makatura, Liane et al.: Large Language Models for Design and Manufacturing. In: *An MIT Exploration of Generative AI* (2024)

[8] Submitted: Kunz, Leonhard et al.: XDP-Opt: Experience-Based Design Process Optimization for Industrial Manufacturing. In: Potsdam, 2025, [Manuscript submitted]

[9] Hedberg, Thomas D.; Hartman, Nathan W.; Rosche, Phil; Fischer, Kevin: Identified research directions for using manufacturing knowledge earlier in the product life cycle. In: *International Journal of Production Research* Bd. 55 (2017), Nr. 3, S. 819–827

[10] Cao, Jianpeng; Vakaj, Edlira; Soman, Ranjith K.; Hall, Daniel M.: Ontology-based manufacturability analysis automation for industrialized construction. In: *Automation in Construction* Bd. 139 (2022), S. 104277

[11] Chirumalla, Koteshwar et al. Exploring feedback loops in the industrialization process: A case study. In: *Procedia Manufacturing* Bd. 25 (2018), S. 169–176

[12] Kang, SungKu et al.: Extraction of Formal Manufacturing Rules from Unstructured English Text. In: *Computer-Aided Design* Bd. 134 (2021), S. 102990

[13] Lee, Hyeonji et al.: Rule-based design verification for mechanical parts with dynamic rule subset selection. In: *Scientific Reports* Bd. 15 (2025), Nr. 1, S. 17153

[14] Papageorgiou, Konstantinos et al.: A systematic review on machine learning methods for root cause analysis towards zero-defect manufacturing. In: *Frontiers in Manufacturing Technology* Bd. 2 (2022), S. 972712

[15] Watson, David S.: Conceptual challenges for interpretable machine learning. In: *Synthese* Bd. 200 (2022), Nr. 2, S. 65

[16] Escobar, Carlos A.; McGovern, Megan E.; Morales-Menendez, Ruben: Quality 4.0: a review of big data challenges in manufacturing. In: *Journal of Intelligent Manufacturing* Bd. 32 (2021), Nr. 8, S. 2319–2334

[17] Ghadai, Sambit; Balu, Aditya; Sarkar, Soumik; Krishnamurthy, Adarsh: Learning localized features in 3D CAD models for manufacturability analysis of drilled holes. In: *Computer Aided Geometric Design* Bd. 62 (2018), S. 263–275

[18] Antony, Jibinraj et al.: Adapting to Changes: A Novel Framework for Continual Machine Learning in Industrial Applications. In: *Journal of Grid Computing* Bd. 22 (2024), Nr. 4, S. 71

[19] Kim, Namjae et al.: A new era DFM solution for yield enhancement using machine learning (ML). In: Lafferty, N. V.; Grunes, H. (Hrsg.): *DTCO and Computational Patterning III*. San Jose, United States: SPIE, 2024 — ISBN 9781510672147 9781510672154, S. 35

[20] GLOBALFOUNDRIES: *GLOBALFOUNDRIES and Cadence Add Machine Learning Capabilities to DFM Signoff for GF's Most Advanced FinFET Solutions*. URL https://gf.com/gf-press-release/globalfoundries-and-cadence-add-machine-learning-capabilities-dfm-signoff-gfs-most/. - abgerufen am 2025-07-11. — GlobalFoundries

[21] Zhang, Wentai et al.: Component segmentation of engineering drawings using Graph Convolutional Networks. In: *Computers in Industry,* Bd. 147 (2023), S. 103885

[22] Ghosh, Akash et al.: Exploring the Frontier of Vision-Language Models: A Survey of Current Methodologies and Future Directions, arXiv (2024)

[23] Zhang, Jingyi; Huang, Jiaxing; Jin, Sheng; Lu, Shijian: Vision-Language Models for Vision Tasks: A Survey. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* Bd. 46 (2024), Nr. 8, S. 5625–5644

[24] Rädsch, Tim et al.: Bridging vision language model (VLM) evaluation gaps with a framework for scalable and cost-effective benchmark generation, arXiv (2025)

[25] Doris, Anna C. et al.: DesignQA: A Multimodal Benchmark for Evaluating Large Language Models' Understanding of Engineering Documentation. In: *Journal of Computing and Information Science in Engineering* Bd. 25 (2025), Nr. 2, S. 021009

[26] Li, Ming; Zhong, Jike; Chen, Tianle; Lai, Yuxiang; Psounis, Konstantinos: EEE-Bench: A Comprehensive Multimodal Electrical and Electronics Engineering Benchmark. In: 2025, S. 13337–13349

[27] Yi, Dongyi et al.: MME-Industry: A Cross-Industry Multimodal Evaluation Benchmark, arXiv (2025)

[28] Chow, Wei et al.: PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding, arXiv (2025)

[29] Masry, Ahmed et al.: ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning, arXiv (2022)

[30] Yeo, Changmo; Cheon, Sanguk; Mun, Duhwan: Manufacturability evaluation of parts using descriptor-based machining feature recognition. In: *International Journal of Computer Integrated Manufacturing* Bd. 34 (2021), Nr. 11, S. 1196–1222

[31] Lambourne, Joseph G.et al.: BRepNet: A Topological Message Passing System for Solid Models. In: 2021, S. 12773–12782

[32] Duan, Haodong at al.: VLMEvalKit: An Open-Source ToolKit for Evaluating Large Multi-Modality Models. In: *Proceedings of the 32nd ACM International Conference on Multimedia*. Melbourne VIC Australia: ACM, 2024 — ISBN 9798400706868, S. 11198–11201