A Review of Data-Centric Artificial Intelligence (DCAI) and its Impact on manufacturing Industry: Challenges, Limitations, and Future Directions

Michael Nieberl
BMW AG
Landshut, Germany
michael.nb.nieberl@bmw.de

Alexander Zeiser
BMW AG
Landshut, Germany
alexander.az.zeiser@bmw.de

Holger Timinger
Institute for Data and Process Science
University of Applied Sciences Landshut
Landshut, Germany
holger.timinger@haw-landshut.de

***Abstract*—With the advancement of big data, the scope and potential of Artificial Intelligence (AI) have acquired major momentum. Data-Centric Artificial Intelligence (DCAI) is one of the most emergent fields of study in the current era of digitalization. Many examples have proven the effectiveness of Machine- and Deep Learning methods. In industrial production, however, limitations are still present that hinder application on-line and in a series that goes beyond isolated use cases. One crucial issue is data precondition, i.e., data quality, consistency, and labeling. As DCAI addresses these issues, developments in this field have caught the attention of various experts. In summary, DCAI continues to be an exciting and promising field of study that enhances AI applicability. Several research works have been conducted in DCAI, but unfortunately, no comprehensive reviews have been conducted to summarize and highlight the results. This gap in knowledge inspired our work, that aims to answer well-structured research questions. The focus of this paper is to clarify the terminology used in DCAI while also distinguishing it from other AI-related problems. This helps to analyze the current standards and problems associated with DCAI. This paper summarizes current use cases of DCAI and their impact on industries. Through this detailed description, readers can understand the potential and benefits of using DCAI in different business sectors. The analysis of the latest methods employed by DCAI to achieve enhanced AI performance and outcomes provides valuable insights for professionals and organizations that strive to incorporate AI into their business.**

Keywords—artificial intelligence, machine learning, data-centric AI, Industry 4.0, manufacturing

I. INTRODUCTION

In recent years, Artificial Intelligence (AI) has witnessed remarkable advancements and has revolutionized industries and everyday life. The success of the rapid improvement of AI has numerous reasons, i.e., improved hardware [1] or growing datasets with high-quality data [2]. AI applications can be found in almost every domain, like cybersecurity [3], teaching [4], industry [5] or in medicine where it was used to fight the COVID-19 pandemic [6]. If there is one AI based application that stands out these days, it is ChatGPT. This Natural Language Processing tool uses advanced AI technology to comprehend and generate text in response to specific commands and can answer questions precise and humanlike [7]. That is the reason ChatGPT is strongly present in the media. The breakthrough about this technology is not only about the sophisticated algorithm like everyone is expecting, but the company's solution also centers on high-quality data, which leads to improved performance in machine learning. This approach was also recognized by Andrew Ng, a professor at Stanford University, which led him to launch a competition. This competition emphasizes the importance of data for ML performance [8]. The aim was to get the best performance out of a given and fixed model, only by improving the quality of data. This procedure can be described as data-centric AI, which focuses on the data quantity and quality. Through this, the potential of DCAI was shown. Since then, the focus has shifted from model-centric AI (MCAI), which focused on enhancing the architecture of machine learning (ML) models by giving a fixed set of data to improve their performance, to the mentioned DCAI. Although AI applications are used in various domains, the DCAI approach has not widely reviewed [9]. This is because the topic is still relatively new and has not been extensively researched. However, despite the widespread utilization of MCAI in the industry, there is a noticeable absence of research exploring the potential impact of DCAI on practices. This knowledge gap presents an opportunity to enhance the industrial utilization. The aim of the current work is to summarize the state-of-the art. The paper is divided into the following sections. It starts with an introduction to the research method. Afterwards, the definitions and differences between DCAI and MCAI are explained, followed by a presentation of related work and an overview of research approaches in DCAI with a focus on industry implementation. After that, a description of potential limitations in implementing and proposed solutions is presented. The paper ends with a summary, outlook, and discussion.

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00018

II. RESEARCH METHOLOGY:

The research process is based on Webster and Watson's "Analyzing the past to prepare for the future" [10]. As they recommend in their paper, our research process does not concentrate on only one geographical zone. The steps, which were conducted for literature acquisition, will be presented in a clear, detailed, and easy-to-understand format for the reader. The literature acquisition process in this review comprises five key steps, which are listed below. Every step will be explained.

- Step 1–Definition of review scope.
- Step 2–Define databases.
- Step 3–Define search terms.
- Step 4–Apply specific selection criteria.
- Step 5–Conducting the process.

Step1: At the beginning, it is mandatory to describe the scope of the literature review. Therefore, different research questions (RQ) were formulated:

1. How does DCAI differ from the model-centric approach and what is it exactly?
2. What kind of impact can DCAI achieve in the industry?
3. What are the challenges and the limitations of the DCAI?
4. How can the limitations be improved?

Step2: To ensure the most effective and comprehensive outcome of the extensive research process, a curated selection of various electronic databases and search engines was selected. The databases which were used are: IEEE Explore, ACM digital library, Scopus, and Springer Link. The decision to incorporate this mix of databases was based on insightful recommendations from experts in the field, specifically Brereton et al. [11].

Step3: Using the wright keywords for the search process is a mandatory step. Therefore, various keywords were identified and combined with different logical operators. The keywords are derived from the RQs. To extract the most relevant and informative sources, these keywords were synergized with a selection of different logical operators, resulting in an optimized and efficient search process. The keywords and the associated operators are shown in Table 1.

Table 1: Used Keywords

("Data-centric" **OR** "DCAI")
AND
("AI" **OR** "Artificial intelligence" **OR** "machine learning" **OR** "ML")
AND
("Production" **OR** "industry" **OR** "industrial processes")

Step4: Considering the defined research questions and the aim of this work to ensure high quality, a set of inclusion and exclusion criteria for the articles were defined. As this paper refers to Andrew NG's Campaign, the first exclusion criteria are that papers which are prior to 2021 are not considered. Regarding Webster & Watson [10], there is no focus on a specific journal or a specific location. The other exclusion and inclusion criteria are listed in Table 2.

Table 2: Inclusion and exclusion criteria for quality insurance

Criteria Type	Description	Criteria
Period	Articles are selected based on the time.	**Exclusion**: prior 2021
Language	Articles are excluded based on their language.	**Exclusion**: Articles that are not written in English
Relevance to research questions	The content of the paper must contribute to answer the RQs.	**Exclusion:** Not relevant to at least 1 RQ´s
Type of literature/source	Articles that fall into the category of gray literature.	**Exclusion**: Reports, working papers, speeches, poster sessions, dissertations, books, **Inclusion**: journal articles and conference proceedings
Accessibility	Not accessible in specific databases.	**Exclusion**: Not accessible

Step5: On 17 October 2023, the research process was conducted. The research began with a comprehensive web search through various databases using the search strings mentioned earlier, locating 548 papers on the subject. However, these papers underwent a screening process to eliminate duplicates, leaving 495 papers for further assessment. In the subsequent stages, the exclusion criteria were inserted into place to sort out irrelevant papers on the subject, leading to the removal of 422 articles. With a smaller pool of papers available, a meticulous process of title and abstract screening was conducted to identify the most pertinent and valuable papers to the research. Following that, a forward search was conducted using Google Scholar, alongside a backward search by screening the citations in the relevant literature. To ensure the overall quality of the review, it is important that the results got from both the forward- and backwards search are consistent with the predefined quality criteria. This resulted in a final selection of 32 papers deemed to be the most relevant and credible sources to provide a comprehensive and accurate analysis of the subject.

III. Related Work

Because of the novelty of DCAI, a comprehensive review of the basics is lacking. In this section, the efforts of other authors should be highlighted. As there are few review papers to which we can relate, we additionally introduce papers that do not exactly meet the inclusion criteria and do not pertain to the industry.

The paper titled "Systematic review on data-centric approaches in AI and ML" was written by Singh, P. (2023) [9] summarizes data-centric approaches used in AI and ML. The author reviewed 165 research articles and found six categories that improve this approach: feature selection and extraction, dimensionality reduction, missing value imputation, imbalance data handling, data augmentation, and data preprocessing. The article also discusses the challenges, limitations, and future potential of the DCAI approach.

Sukdeo, N. I., & Mothilall, D. (2023) [12] review the impact of AI in the Printing and Packaging industry. The study analyzes 33 articles on AI's effects on production, customer satisfaction, and employee training. The adoption of AI technologies is explored, and this includes discussing the implications such as increased productivity, reduced costs, and re-skilling of the workforce. Practical insights for managers, policymakers, and practitioners in the industry are provided by the paper.

Adeoye et al. (2023) [13] reviews the use of DCAI for head and neck cancer (HNC) treatment. They analyze 19 articles on data characteristics, feature engineering, validation techniques, and statistical results. The review underscores the challenges of comparing studies because of variations in features and outcome metrics. Various studies lacked data quality reporting, leading to potential bias and limited generalizability. The paper underscores the need for robust data management practices and their impact on precision medicine in HNC treatment.

Zah et al. (2023) [14] provide a comprehensive overview of data-centric approaches to AI. They explain data-centric AI and its emphasis on acquiring, preparing, and managing data. The paper covers data-centric AI concepts, including data sources, labeling, feature engineering, validation/evaluation, and management. The authors discuss the challenges and future directions of this field and summarize the benefits of data-centric AI. This paper is a valuable resource for researchers and practitioners in the field.

Based on the summarized related work, DCAI is becoming important in industry. However, there is still a need to further investigate and explore the potential of DCAI in various industrial applications. Therefore, this paper is necessary to provide a comprehensive understanding of the benefits and limitations of DCAI and examine its status in industry and propose future directions for research and development. Such insights would be valuable for industry practitioners, policymakers, and researchers interested in leveraging the power of DCAI to improve various aspects of industrial operation.

IV. Definition of Terms

Summarizing the related work, it stands out that no clear definition of terms is present in the study of DCAI. The following section defines the term DCAI based on our analysis. The difference between DCAI and MCAI is also explained in this chapter.

A. MCAI

The model-centric approach focuses on finding the best model with a fixed set of data, and improving model accuracy is achieved through hyperparameter tuning [19]. However, data can contain features that do not improve precision, resulting in overfitting, and erroneous data that is hard to detect in larger datasets and cannot be fixed by hyperparameter tuning [22]. To compensate for these weaknesses, the most common approach is to collect more data. Model-centric AI mainly focuses on optimizing model architecture and hyperparameters, with data created almost only once and kept the same throughout the AI system's development lifecycle [17]. However, this approach has been under considerable strain because of its vulnerability to adversarial samples, a narrow scope of business applicability, and low generalization capacity.

In model-centric AI, the success of the AI model is perceived to come from the sophistication of its design and model, not from the data used to train it. The traditional model-centric approach leaves little opportunity for revising and improving the quality of data systematically and progressively. Instead, data preparation is mostly performed at the onset through preprocessing steps, creating a static approach to data quality. The burden of dealing with data issues, such as data noise, is primarily left with models, reflecting the prevalent norm of data indifference in the AI community, which mainly sees data as just fuel for model training. This approach can cause data cascades with negative and unpredictable consequences in downstream AI deployments. The MCAI-approach is shown in Figure 1A.

B. DCAI

The DCAI approach is grounded in the understanding that the dynamism of data is an ever-evolving piece of infrastructure in an AI system. Therefore, DCAI emphasizes that the quality of data used for training an AI model must evolve along with the updated data to guarantee that the model stays relevant and accurate for its intended use [23]. The DCAI approach is an additional step to the traditional model-centric approach, which focus on developing the best model regardless of the quality of data [22]. DCAI comprises two steps. First, the data should be improved as well as possible to enhance the performance of a system and second, the model should be adjusted by hyperparameter tuning [24]. This approach can be seen in Figure 1B. By systematically refining the used dataset, DCAI practitioners avoid common sources of error that can arise from poor data quality or data imbalance. In summary, the DCAI

approach provides a framework for practitioners to develop and deploy AI models, which ensures that the data used in model training is continuously refined and benchmarked to maintain its quality [25]. DCAI refines the traditional model-centric established process by weighing features for the AI model at hand and improving the quality of data to reduce computational time, while also increasing the precision of the model with each iteration. Approaches and methods of DCAI are not newly developed for this purpose. It is rather about the attention that is given to high-quality data as a prerequisite for well-functioning AI models.

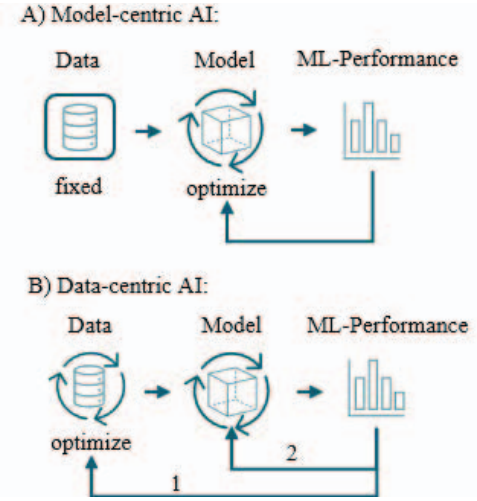

Figure 1: Comparison of the MCAI system (A) with the DCAI system (B)

V. Impact of DCAI on the Industry

The next chapter discusses the impact of DCAI on the industry. As mentioned before, the approach is not widespread now, but different companies already implemented a DCAI-system. The results of experiments and the impact on the environment are explained and summarized in this section.

Data-Centric Green Artificial Intelligence A Survey [15]: Data-centric green AI reduces environmental impact. AI model and training set size can be disadvantages. Chat GPT-3 emitted 552 tons CO2 during training, nearly 10 times a car's lifetime emissions. Energy consumption and data storage contribute to the worsening of the issue. Conserving energy involves saving less data. DCAI is key to reducing consumption. This paper presents 5Rs techniques (remove, restrict, reorder, replace, retrofit) to accelerate ML training and minimize storage. Their ML framework includes active learning, knowledge sharing, dataset distillation, data augmentation, and curriculum learning. Model accuracy can be affected, so balancing is crucial.

Data-Centric Model Development to Improve the CNN Classification of Defect Density SEM Images [16]: This study uses CNN to classify defects. The data-centered approach deals with imbalanced and insufficient datasets. 6 experiments were conducted, involving three tailored data sets with added values and classes. The first two models were used to derive a corresponding model. 6 tests showed that data significantly influences accuracy; 6 adjustments improved accuracy from 59.4% to 92.7%.

Anomaly Detection in Time Series Data using Data-Centric AI [17]: This analysis shows how data-centric AI techniques enhance anomaly detection and improve model performance. The authors address the challenge of mislabeling in supervised learning, which impacts system performance. Three experiments are conducted to evaluate their approach: using a model-based AI algorithm, a model-centric algorithm, and high label focus. Through optimized labeling, the authors reduced the loss from 0.579 to 0.0233 and improved accuracy by 23%. The paper underscores the significance of data quality and accurate labeling for optimal model performance.

Synthetic Training Data Generation for Convolutional Neural Networks in Vision Applications [18]: In this investigation, the authors describe the problem of object recognition using CNN. The problem with this task is the different angles from which images can be taken. It is explained based on object recognition on a bicycle and solved by data synthesis. For this purpose, synthetic data of a bicycle is generated at a certain angle and the algorithm is fed in for training. This extended training data set solves the problem of object recognition.

From Concept to Implementation: The Data-Centric Development Process for AI in Industry [19]: In this paper, the authors address the problem that "small- and medium-sized enterprises" have when creating machine learning applications. The focus is on the quality and quantity of the data. To solve this problem, a proposed solution was developed that should make it possible to implement DCAI in companies. The collaboration of various experts and the utilization of tools for data processing are crucial in this.

Next-generation Challenges of Responsible Data Integration [20]: This article provides a tutorial on data integration and accountability in machine learning pipelines. It emphasizes the need to address data quality and bias concerns in responsible data science. The authors discuss existing efforts, research opportunities, and challenges in responsible data integration. The tutorial focuses on evaluating tasks for data integration, measures for data stewardship, and techniques to achieve data stewardship.

A Data-Centric Approach to Design and Analysis of a Surface-Inspection System Based on Deep Learning in the Plastic Injection Molding Industry [21]: This paper suggests a deep learning inspection system for tampon applicators, using their properties for data collection and feature extraction. Testing found false positives resolved by two data preprocessing techniques. While overall performance declined, experiments showed improved recognition. The authors warn against relying solely on metrics, as data annotation inconsistencies can lead to decision ambiguities. Malfunctions

were resolved using various data-centered techniques in the field. These techniques were validated with experiments on normal and outlier image datasets.

VI. Potential Solutions of DCAI for Common Issues of AI Systems

This chapter addresses the general problems that can arise when implementing an AI system. Many of these causes have their problems in the dataset, which is why DCAI can provide a solution. While this list is not exhaustive, it presents the most common issues that may occur during implementing an AI system. We introduce several solutions to mitigate these problems.

A. Current Limitations:

Data collection. Data collection involves gathering data. While constructing new datasets from scratch is a common approach, it can be time-consuming. Another problem can be missing data [23]. Reasons for this are that the process is still relatively new or that the data has simply not yet been recorded. It can also happen that a feature that is still relatively new has not been measured often and therefore data is still missing for this specific task.

Limited Data: The problem about limited data is the difficulty in training AI models using limited datasets, which is often the case in manufacturing industries. The lack of extensive and diverse datasets in these industries can cause models that cannot capture complex relationships and patterns underlying the data [26]. Although machines can record numerous data through sensors, unfortunately, the relevance of the data is not always given and therefore cannot be used to train an AI.

Data Quality: Data quality describes the lack of comprehensive consideration of data quality requirements in the machine learning development pipeline. Different stakeholders may have distinct data quality preferences and requirements [27]. Granular data quality specifications established in data management are often insufficient.

Data Bias: Data bias has become a significant concern in AI systems. This bias often comes from imbalanced distributions of sensitive variables in the data [25]. From a DCAI perspective, several challenges arise:

1. How can bias in training data be mitigated?
2. How can evaluation data be constructed to expose unfairness?
3. How can data unbiasedness be maintained in a dynamic environment?

Failure to address biases introduced during training can cause biased behaviors of the AI system, contributing to concerns about fairness and a loss of trust [28]. However, the information about training data is often not effectively communicated to stakeholders, limiting their ability to understand and address these biases.

Labeling Quality from Data: The focus of investigation revolves around the accuracy of labeling, which reflects the extent to which the assigned labels align with the raw training data [23]. An example of illustration is a machine learning system used for traffic surveillance. If the system accurately detects all the cars and correctly classifies them as cars, and all the bikes as bikes, we can conclude that the classification quality is high. However, if there is an inconsistency in classification, where some bikes are labeled as cars and vice versa, the classification quality would be low and might cause dangerous situations on the road. Some observe in previous studies that AI accuracy positively influences user perceptions and trust in AI. Thus, based on this premise, it is predicted that showing accurate labeling practices and high-quality labeling will enhance the perceived credibility of the training data used in AI systems [29].

Noise features/Data poisoning: Noisy features can arise because of poisoning attacks during the training phase of machine learning models or because of low labeling quality. Poisoning attacks the training procedure by injecting maliciously designed data that targets features and/or labels [30]. While defense strategies such as adversarial training, knowledge distillation, feature squeezing, and separate classification networks have been explored to develop more robust models, the challenge of defending against poisoning attacks remains significant [30].

Missing labels: Considering this problem, the labels were not swapped, influenced, or incorrect, but simply not added or forgotten. This error means that the ML model cannot be trained accurately [30]. This affects primary supervised models, which depend on labels.

Class overlaps: The problem of class overlap refers to the issue where there is overlap among different classes in a machine learning classifier. This can lead to performance degradation in the classifier, as it may misclassify points or have less confidence in its predictions in the overlapping regions [31].

B. Solutions

Data centric explanations: The problem at hand is the lack of transparency in AI systems, making it difficult for end-users to understand the output. A potential solution is to provide data-centric explanations to end-users. These explanations involve communicating the training datasets used in machine learning systems, enabling end-users to acquire a deeper understanding. A study suggests that data-centric explanations can help develop trust in machine learning systems, particularly when the training data appears balanced. Conversely, revealing problems with the training data can negatively impact trust [28]. Data-centric explanations have a lesser impact on end-users' perception of system fairness. While some concerns were raised about the complexity of these explanations for end-users and the potential for confirmation bias, future research should explore the effectiveness of data-centric explanations for different audiences, such as journalists and decision-makers involved in acquiring AI systems.

Using machine learning operation platforms: One solution to reduce maintenance cost of AI applications is to use machine learning operation, MLOps-platforms, rather than spending resources on developing software from scratch. MLOps platforms provide necessary scaffolding software, which streamlines the production of an AI system and reduces the gap between proof of concept and production from years to weeks [22]. MLOps systems exist for both data-centric and model-

centric AI and can take on work such as data-labeling and data-cleaning, respectively. For MCAI, available MLOps tools include model store systems, model continuous integration tools, training platforms, and deployment platforms [26].

Using Data-Experts: Another step towards streamlining the use of AI systems is to have business-domain experts that guide or even perform data engineering, instead of relying solely on AI experts. This is because AI experts possess competence mainly in representing a domain in a format that enables algorithms to learn patterns, whereas domain experts possess comprehensive knowledge about specific business use cases and can hence provide domain-relevant representation of the world [30]. Domain experts can enhance the evaluation process by creating use cases that test the model in more domain-sensitive scenarios. By involving domain experts in the process of data engineering and evaluation, the use of AI becomes more accessible to a wider range of industries, resulting in a more widely used and beneficial AI system [22].

Augmentation: Data augmentation is a technique that can address a wide range of problems in various domains. It involves creating new training samples by altering existing ones, which can improve model performance and help overcome issues such as overfitting, class imbalance, and limited data availability.

Synthetic Data: In contrast to augmented data, synthetic data is newly generated data. As an approach to generate synthetic data, domain randomization can be adapted, either by simulation or by abstraction of the real domain [32]. Another approach that can be used for data augmentation are generative adversarial networks (GAN) [33]. Considering the problem of low data volume, lacking data diversity or trustful labels, the synthetic data production opens auxiliary possibilities for model development, especially if combined with transfer learning.

Tackle Data Noise: To tackle this problem, noisy data must be filtered out of the dataset. Unfortunately, there is a problem if a data point is close to an overlap region. To overcome this is issue, two improvements should be explained. First, by using a new algorithm to detect the overlap region and prune the potential candidates for noise, and second, by implementing effective neighborhood-based strategies to evaluate neighboring samples and suggested labels to improve the accuracy of noise detection [31].

High Quality Data Labeling: High quality in labels can be achieved with different techniques. The easiest but most time-consuming way is manual labeling, where humans add the label based on their own judgement. A more time efficient way is to automate the labeling effort. One way to achieve this is by using semi-supervised learning approaches. This technique aims to derive the labels from a smaller data set with labeled data. Therefore, a model can be trained to make predictions for the unlabeled data [34]. Another automated way for data labeling is active learning. It uses correctly classified data provided by the user in a feedback loop to reduce the amount of annotated data needed to train a model.

VII. Discussion:

What is DCAI and how does it differ from the model-centric approach? In some papers DCAI was defined as a single-step process which focus primarily on the data. This paper defines it as a part of a two-step process. Hyperparameter tuning, which is associated with MCAI, should also be in the focus when an AI-system is optimized. If these two steps are considered, optimum results should be achieved.

What is the impact that DCAI can achieve in the industry? DCAI has already been used successfully for individual applications. These cases are briefly presented here. Unfortunately, there is no documentation of failed cases based on this approach in the current literature. Such documentation can help the user avoid making the same mistakes or provide a basis for further research.

What are the challenges and the limitations of the DCAI? Besides the usual challenges that can affect AI-systems, DCAI faces additional ones. The computing capacity, which is needed for generating synthetic data, is one of these challenges that must be tackled. In addition, synthetic data is difficult to produce, depending on the application. This can also be the case if there is still very little data available. Another challenge that needs to be overcome is the development of an evaluation algorithm with which it is possible to assess whether the data basis is sufficient.

How can the limitations be improved?
To improve transparency, data-centric explanations can be provided to end-users, which involves communicating the training datasets used in machine learning systems to enable a deeper understanding of the output. Using MLOps can reduce the cost of AI applications while providing necessary scaffolding software to produce AI systems.
To address domain-specific issues, relying on domain experts instead of solely AI experts can guide or even perform data engineering, resulting in more accessible and widely used AI systems. Augmentation techniques such as data augmentation, synthetic data generation, and tackling data noise can enhance data quality, reduce overfitting, and class imbalance, and improve the accuracy of noise detection.
Finally, high-quality data labeling can be achieved through a variety of techniques, such as manual labeling, semi-supervised learning approaches, and active learning, which can reduce the amount of annotated data needed to train a model.

VIII. Summary and Outlook:

This paper explains the terms DCAI and MCAI and distinguishes between them. Once the terminology had been clarified, the problems that can occur when implementing an AI-system were explained. Based on this, solutions were also shown. Real deployment reports and other experiments related

Table 3: Findings and future directions

RQ	Findings	Future direction
RQ1	DCAI emphasizes data optimization as a distinct initial step, which, when combined with subsequent model-centric hyperparameter tuning, can yield optimal AI performance.	Future efforts may focus on creating unified frameworks that merge data-centric and model-centric methods for comprehensive AI optimization.
RQ2	DCAI's adoption could lead to enhanced AI reliability and efficiency in the industry by learning from both its successes and undocumented failures.	Future research should focus on cataloging DCAI failures to improve AI development strategies.
RQ3/RQ4	DCAI shares common issues with MCAI, but actively addresses these challenges.	Future research should aim at enhancing computational efficiency, improving synthetic data generation techniques, and creating robust data sufficiency evaluation algorithms for DCAI.

AI-system were explained. Based on this, solutions were also shown. Real deployment reports and other experiments related to the topic were also summarized. This should encourage readers to consider the data-centered AI/ML. These findings and their derived future directions are listed in Table 3. Overcoming the before discussed limitations requires the adoption of comprehensive data quality guidelines, effective communication systems, and stringent monitoring and validation processes. Addressing these challenges is vital to ensure the development of reliable and trustworthy AI systems in DCAI. Additional to this, the focus should be on developing systems to generate synthetic data.

IX. References

[1] T. Baji, "Evolution of the GPU Device widely used in AI and Massive Parallel Processing," in *2018 IEEE 2nd Electron Devices Technology and Manufacturing Conference (EDTM)*, Kobe, 2018, pp. 7–9, doi: 10.1109/EDTM.2018.8421507.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017, doi: 10.1145/3065386.

[3] A. Vehabovic *et al.,* "Data-Centric Machine Learning Approach for Early Ransomware Detection and Attribution," in *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*, Miami, FL, USA, 2023, pp. 1–6, doi: 10.1109/NOMS56928.2023.10154378.

[4] L. Chen, P. Chen, and Z. Lin, "Artificial Intelligence in Education: A Review," *IEEE Access*, vol. 8, pp. 75264–75278, 2020, doi: 10.1109/ACCESS.2020.2988510.

[5] G. Zeba, M. Dabić, M. Čičak, T. Daim, and H. Yalcin, "Technology mining: Artificial intelligence in manufacturing," *Technological Forecasting and Social Change*, vol. 171, p. 120971, 2021, doi: 10.1016/j.techfore.2021.120971.

[6] X. Yue, H. Li, and L. Meng, "AI-based Prevention Embedded System Against COVID-19 in Daily Life," *Procedia computer science*, early access. doi: 10.1016/j.procs.2022.04.021.

[7] M. Javaid, A. Haleem, and R. P. Singh, "A study on ChatGPT for Industry 4.0: Background, potentials, challenges, and eventualities," *Journal of Economy and Technology*, vol. 1, pp. 127–143, 2023, doi: 10.1016/j.ject.2023.08.001.

[8] E. Strickland, "Andrew Ng, AI Minimalist: The Machine-Learning Pioneer Says Small is the New Big," *IEEE Spectr.*, vol. 59, no. 4, pp. 22–50, 2022, doi: 10.1109/MSPEC.2022.9754503.

[9] P. Singh, "Systematic review of data-centric approaches in artificial intelligence and machine learning," *Data Science and Management*, vol. 6, no. 3, pp. 144–157, 2023, doi: 10.1016/j.dsm.2023.06.001.

[10] J. Webster and R. T. Watson, "Analyzing the Past to Prepare for the Future: Writing a Literature Review," *MIS Quarterly*, vol. 26, no. 2, pp. xiii–xxiii, 2002. [Online]. Available: http://www.jstor.org/stable/4132319

[11] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, "Lessons from applying the systematic literature review process within the software engineering domain," *Journal of Systems and Software*, vol. 80, no. 4, pp. 571–583, 2007, doi: 10.1016/j.jss.2006.07.009.

[12] N. I. Sukdeo and D. Mothilall, "The Impact of Artificial Intelligence on the Manufacturing Sector: A Systematic Literature Review of the Printing and Packaging Industry," in *2023 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, Durban, South Africa, 2023, pp. 1–5, doi: 10.1109/ICABCD59051.2023.10220486.

[13] J. Adeoye, L. Hui, and Y.-X. Su, "Data-centric artificial intelligence in oncology: a systematic review assessing data quality in machine learning models for head and neck cancer," *J Big Data*, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00703-w.

[14] D. Zha *et al.,* "Data-centric Artificial Intelligence: A Survey," Mar. 2023. [Online]. Available: http://arxiv.org/pdf/2303.10158v3

[15] S. Salehi and A. Schmeink, "Data-Centric Green Artificial Intelligence: A Survey," *IEEE Transactions on Artificial Intelligence*, pp. 1–18, 2023. doi: 10.1109/TAI.2023.3315272. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85171536422&doi=10.1109%2fTAI.2023.3315272&partnerID=40&md5=7a89e5885c113ba09ad72a4543cfd531

[16] C. Kofler, C. A. Dohr, J. Dohr, and A. Zernig, "Data-Centric Model Development to Improve the CNN Classification of Defect Density SEM Images," in *IECON 2022 – 48th Annual Conference of the IEEE Industrial Electronics Society*, 2022, pp. 1–6, doi: 10.1109/IECON49645.2022.9968911.

[17] C. Hegde, "Anomaly Detection in Time Series Data using Data-Centric AI," in *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, Bangalore, India, 2022, pp. 1–6, doi: 10.1109/CONECCT55679.2022.9865824.

[18] H. Vietz, T. Rauch, and M. Weyrich, Eds., *Synthetic Training Data Generation for Convolutional Neural Networks in Vision Applications*, 2022, doi: 10.1109/ETFA52439.2022.9921534.

[19] P. -P. Luley, J. M. Deriu, P. Yan, G. A. Schatte, and T. Stadelmann, "From Concept to Implementation: The Data-Centric Development Process for AI in Industry," in *2023 10th IEEE Swiss Conference on Data Science (SDS)*, 2023, pp. 73–76, doi: 10.1109/SDS57534.2023.00017.

[20] F. Nargesian, A. Asudeh, and H. V. Jagadish, "Next-Generation Challenges of Responsible Data Integration," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 2023, pp. 1256–1259, doi: 10.1145/3539597.3572727.

[21] Donggyun Im, Sangkyu Lee, Homin Lee, Byungguan Yoon, Fayoung So, and Jongpil Jeong, "A Data-Centric Approach to Design and Analysis of a Surface-Inspection System Based on Deep Learning in the Plastic Injection Molding Industry," *Processes*, 1895 (22 pp.)-1895 (22 pp.), 2021, doi: 10.3390/pr9111895.

[22] O. H. Hamid, "From Model-Centric to Data-Centric AI: A Paradigm Shift or Rather a Complementary Approach?," in *2022 8th International Conference on Information Technology Trends (ITT)*, 2022, pp. 196–199, doi: 10.1109/ITT56123.2022.9863935.

[23] D. Zha, K.-H. Lai, F. Yang, N. Zou, H. Gao, and X. Hu, "Data-Centric AI: Techniques and Future Perspectives," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 5839–5840, doi: 10.1145/3580305.3599553.

[24] Y. Zhong, L. Wu, X. Liu, and J. Jiang, "Exploiting the Potential of Datasets: A Data-Centric Approach for Model Robustness," Mar. 2022. [Online]. Available: http://arxiv.org/pdf/2203.05323v1

[25] M. H. Jarrahi, A. Memariani, and S. Guha, "The Principles of Data-Centric AI," *Commun. ACM*, vol. 66, no. 8, pp. 84–92, 2023. doi: 10.1145/3571724. [Online]. Available: https://doi-org.ub-proxy.fernuni-hagen.de/10.1145/3571724

[26] O. H. Hamid, "Data-Centric and Model-Centric AI: Twin Drivers of Compact and Robust Industry 4.0 Solutions," *Applied Sciences (Switzerland)*, vol. 13, no. 5, 2023. doi: 10.3390/app13052753. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85149675012&doi=10.3390%2fapp13052753&partnerID=40&md5=a13dee6ed8cf8b6c5b5711ed5792cea5

[27] M. Priestley, F. O'donnell, and E. Simperl, "A Survey of Data Quality Requirements That Matter in ML Development Pipelines," *J. Data and Information Quality*, vol. 15, no. 2, pp. 1–39, 2023, doi: 10.1145/3592616.

[28] A. I. Anik and A. Bunt, "Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan, Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn, and S. Drucker, Eds., 2021, pp. 1–13, doi: 10.1145/3411764.3445736.

[29] C. Chen and S. S. Sundar, "Is this AI trained on Credible Data? The Effects of Labeling Quality and Performance Bias on User Trust," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg Germany, A. Schmidt et al., Eds., 2023, pp. 1–11, doi: 10.1145/3544548.3580805.

[30] S. E. Whang, Y. Roh, H. Song, and J.-G. Lee, "Data collection and quality challenges in deep learning: a data-centric AI perspective," *The VLDB Journal*, vol. 32, no. 4, pp. 791–813, 2023. doi: 10.1007/s00778-022-00775-9. [Online]. Available: https://link.springer.com/10.1007/s00778-022-00775-9

[31] H. Patel, S. Guttula, N. Gupta, S. Hans, R. S. Mittal, and L. N, "A Data Centric AI Framework for Automating Exploratory Data Analysis and Data Quality Tasks," *J. Data and Information Quality*, 2023. doi: 10.1145/3603709. [Online]. Available: https://doi-org.ub-proxy.fernuni-hagen.de/10.1145/3603709

[32] A. Zeiser, B. Ozcan, B. van Stein, and T. Back, "Evaluation of deep unsupervised anomaly detection methods with a data-centric approach for on-line inspection," *Computers in Industry*, p. 103852, 2023, doi: 10.1016/j.compind.2023.103852.

[33] H. Vietz, T. Rauch, and M. Weyrich, "Synthetic Training Data Generation for Convolutional Neural Networks in Vision Applications," in *2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA)*, 2022, pp. 1–6, doi: 10.1109/ETFA52439.2022.9921534.

[34] A. Vangala, A. K. Das, N. Kumar, and M. Alazab, "Smart Secure Sensing for IoT-Based Agriculture: Blockchain Perspective," *IEEE Sensors Journal*, vol. 21, no. 16, pp. 17591–17607, 2021, doi: 10.1109/JSEN.2020.3012294.