EDITOR **HSIAO-YING LIN**
IEEE Member; hsiaoying.lin@gmail.com

Data-Centric Artificial Intelligence, Preprocessing, and the Quest for Transformative Artificial Intelligence Systems Development

Abdul Majeed and **Seong Oun Hwang**, Gachon University

Some data-centric artificial intelligence (AI) practices often resemble the preprocessing used in traditional AI systems; therefore, there is a possible misconception about the two approaches. This article differentiates these two aspects to guide the AI community to unlock the hidden potential of the data-centric AI paradigm.

Digital Object Identifier 10.1109/MC.2023.3240450
Date of current version: 3 May 2023

The data-centric artificial intelligence (DC-AI) concept was first coined by Prof. Andrew Ng in a livestream session hosted on 24 March 2021.[1] DC-AI gives a higher preference to data rather than iteratively improving only the complex code in AI models.[2] DC-AI can contribute greatly to advancing the technical efficacy of AI in solving many real-world complex problems that are unlikely to be solved with other approaches, such as model-centric AI (MC-AI).[3] The MC-AI approach mainly improves the architectural aspects, specifically the code, of AI models. A detailed introduction of both MC-AI and DC-AI is presented in Figure 1.

The main motivations behind these code-based developments (that is, MC-AI) are to overcome multiple issues in AI, such as larger parameters, stability problems, computing overhead, deficiencies in salient feature extraction, and model sizes. As a result, a variety of AI models have been developed, and more developments are underway in this context (that is, improving code in AI models) to further advance AI.[4] Here, we summarize six main research dimensions (most MC-AI) in which the AI community is investing a lot of effort and money in the ongoing era:

- improving the network architecture of models
- pruning/quantizing the network architecture by removing redundant weights/parameters
- developing new AI models (or upgrading existing versions)
- developing new frameworks for AI model training (that is, centralized → decentralized)
- devising new strategies for hyperparameter tuning to improve model performance
- expanding horizons of AI applications to multiple disciplines/problems.

As mentioned, the AI community pays relatively more attention to improving the architectural aspects in AI models rather than data. Currently, most researchers/engineers compare model accuracy (like competitions in Kaggle and AI conferences, such as the Neural Information Processing Systems) with a fixed dataset. However, data are an integral component for AI quality, and DC-AI is also mandatory to rectify as well as increase the adoption of AI technology. In addition, it can address longstanding problems of conventional MC-AI (for example, trustworthiness, reliability, and AI use for social good).[5] This is the right time to take practical steps to adopt and realize DC-AI, making a significant impact on our society.[6]

Unfortunately, the AI community and its researchers are hesitant to adopt DC-AI on a large scale, and they regard DC-AI as merely a substitute for the preprocessing in conventional AI models. However, that is not the case because DC-AI is a much more promising and handy technology that is not simply for preprocessing data. In this article, we take the first step toward communicating the promises of DC-AI, and we highlight the fundamental distinction between DC-AI and preprocessing to guide the AI community in the right direction. Our painstaking analysis can contribute to sparking further developments in DC-AI, leading to the development of transformative AI systems for the well-being of the general public around the world.

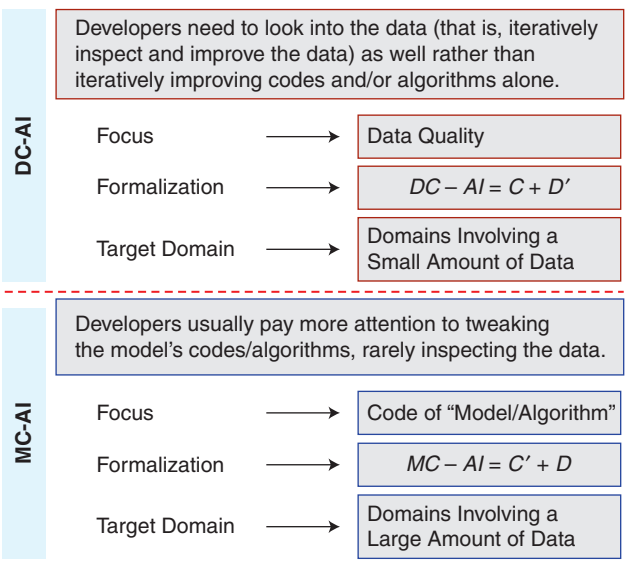

FIGURE 1. Introduction of DC-AI and MC-AI. The notations *D* and *C* refer to *data* and *code*, respectively. The prime sign (′) in formalization indicates priority.

BREAKTHROUGHS IN AND SCOPE OF DC-AI

This section presents the major breakthroughs in and scope of the DC-AI paradigm.

Accuracy enhancement in defect detection scenario

In some scenarios, there can be some predefined performance targets (for example, ≥90% accuracy) that need to be accomplished using AI. To this end, MC-AI alone may not help meet those targets because it often pays less attention to the data, which constitute an important component of AI quality. To that end, we present a scenario in which the accuracy target for defect detection in steel was set to 90%. Despite significantly improving the code, the accuracy was not improved much by using the MC-AI alone. In contrast, DC-AI was handy in augmenting

accuracy by 16.9%. This scenario affirms the need to employ DC-AI to accomplish certain performance targets while solving real-life problems.

Accuracy enhancement in anomaly detection scenarios

Accurately detecting data points that are anomalous in time series data is a very challenging task, and existing MC-AI-based techniques yield poor performance in real-time applications involving time series data. In this regard, a recent DC-AI-based approach obtained 100% accuracy in detecting data points that are anomalous.[7]

Training of deep neural networks with fewer data

Recently, there has been an increasing trend to use AI models on resource-constrained devices. To this end, reduced model size and fewer but good-quality data are of prime importance. In line with these trends, a recent DC-AI-based approach was developed to train a deep neural network (NN) model with many fewer data, leading to significantly lowering the model size by 1.54 × while improving accuracy by up to 5%.[8]

Domain-specific data for general-purpose language models

Recently, there has been an increasing focus on automating language models for social services, such as a mobile doctors. In these services, good-quality data can help answer precisely and accurately. Recently, a Stanford Center for Research on Foundation Models team developed a general-purpose language model with domain-specific data.[11] The proposed model is low-cost in terms of the computing budget while exploiting training recipes.

In addition, DC-AI can bring more benefits[12] in terms of AI model development time (10×), the time required from developing to deploying the AI model (65%), and accuracy enhancement (40%). Finally, it can help address other issues (that is, bias, explainability, etc.) in MC-AI.

The potential scope of DC-AI is not limited to the data-driven AI domain only. It can optimize the performance of conventional machine learning (ML) (both supervised and unsupervised) and deep learning models (such as NNs). In unsupervised learning, DC-AI can help with reducing the number of iterations in clustering, deciding optimal values of hyperparameters based on data size, forming balanced clusters, determining optimal convergence criteria, etc. Moreover, it is important to note that DC-AI may not be very beneficial in large-scale AI models that consume big data; however, it can contribute to reducing the computing budget and fixing social problems (for example, fairness, transparency, etc.) in them. Also, it can be a viable approach when those models yield poor performance owning to invalid and mislabeled instances.

> DC-AI can contribute greatly to advancing the technical efficacy of AI in solving many real-world complex problems that are unlikely to be solved with other approaches, such as model-centric AI.

DC-AI VERSUS CONVENTIONAL PREPROCESSING

Workflows for DC-AI and preprocessing, when applied to solving some real-world problems using AI, are illustrated in Figure 2. We can see that DC-AI evaluates data throughout the lifecycle of an AI project, whereas preprocessing is applied just once. In the collection phase, data quality is given higher preference than quantity, and data are evaluated based on relevance to the problem. Furthermore, the appropriate data providers that can provide high-quality data at the lowest possible cost are determined in this phase.

In the data engineering phase, labeling performed by multiple labelers is assessed from a consistency point of view. The annotation process and sizes of the bounding box are also evaluated to filter out ambiguous labeling. Bounding boxes can also be used to align the values of attributes in synthetic as well as real data enclosed in a tabular form. They can also be used in medical image analysis, human activity recognition, etc. Also, class distributions [such as people's income (for example, ≥US$50,000 and <US$50,000) or disease distributions in medical data] can be determined via visualization, and all features are evaluated to generate representative data. We refer interested readers to get more insights about the iterative evaluations performed in the lifecycle of DC-AI from an informative blog.[13]

The evaluation techniques used in the preceding two phases are different from the ones for other phases, such as model building and hyperparameter tuning. In modeling building, it is assessed whether all samples were equally used in AI model training or not.[9] In hyperparameter tuning, it is evaluated whether the optimal set of parameters can yield consistent performance in unseen data or not.

A comparative analysis of DC-AI and preprocessing techniques is given in Figure 3. From the analysis, DC-AI encompasses more sophisticated techniques for data quality enhancement/enrichment than

preprocessing. Although preprocessing entails useful techniques that can improve data, many aspects concerning AI quality are often overlooked. For example, preprocessing does not stress the need to ensure greater diversity in the training data, which is vital for fair decision making in AI systems. Similarly, important checks, such as whether data are complete from all perspectives or not, whether data are up to date or not, and whether the data are relevant to the problem under investigation or not, are often overlooked in preprocessing. Furthermore, preprocessing is applied to data that were already collected, without paying attention to relevance. In addition, it offers no concrete solutions when data-collection budgets are small.

There are four main techniques in preprocessing, such as data cleaning, integration, transformation, and reduction. Data cleaning is an important preprocessing technique in which multiple operations (that is, outlier removal, imputing/ignoring missing values, eliminating noise, and addressing inconsistencies) are performed to clean data. Data integration is the process to combine data scattered across multiple sources/institutes, and it is often regarded as data warehousing. It is required to solve complex problems like detecting the existence of nodules by using computerized tomography scan images. Data transformation is employed to change the structure, value, or format of data. It can give proper meaning to the data, and it can be accomplished using normalization, generalization, feature selection, and aggregation operations. Data reduction deals with reducing the size of data without losing guarantees on analytical results. It is imperative while performing analytics on a massive amount of data.

In contrast, DC-AI puts more emphasis on data quality before the initiation of AI-powered solutions to any problem, and it can be a potential remedy for longstanding problems in AI.[10] It includes more techniques that are relatively more advanced than preprocessing and that can open up the black-box nature of AI models. There are three key building blocks of DC-AI: data-first strategy (DFS), intelligent data architecture (IDA), and data compliance. DFS alone provides many sophisticated techniques compared to traditional preprocessing. For example, it ensures data quality, availability, and observability through 14 different operations. IDA ensures proper versioning of data and ensures effective utilization of data in the lifecycle of product development. Data compliance ensures privacy-aware processing of data, and it provides practices that are imperative for the responsible use of data.[14] It is worth noting that some of these techniques may not be needed all the time, and, therefore, optimal techniques can be selected based on domain knowledge.

It is important to note that some preprocessing techniques are used with DC-AI. However, the correct order and combinations as well as a relevancy analysis are rigorously performed before their actual use. The role of preprocessing is limited (for example, it ends once the model is trained/deployed), and it can only marginally improve some aspects of AI (such as accuracy). Furthermore, preprocessing is a one-time event (that is, before training an AI model). In contrast, DC-AI is an iterative process of data quality enhancement/assessment and is expected to bring a dramatic revolution in AI that is

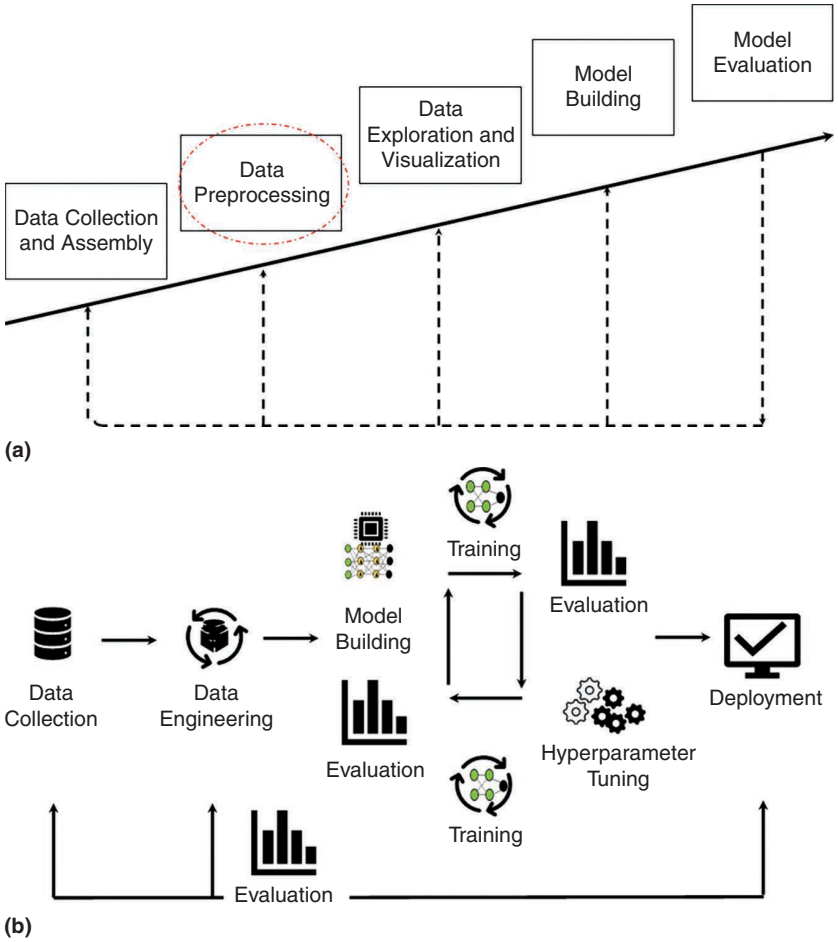

FIGURE 2. Workflows for preprocessing and DC-AI when applied to solving real-world problems. (a) Preprocessing in practice. (b) DC-AI in practice.

not possible with MC-AI and/or preprocessing alone, and it can likely be adopted to solve global issues (for example, climate change, supply chain disruptions, and event prediction). The amalgamation of DC-AI with preprocessing, MC-AI, and other relevant techniques is imperative to advancing AI technology.

It is worth noting that DC-AI has certain challenges that require the urgent attention of the AI community. For instance, DC-AI recommends higher diversity in training data—but, sometimes, it is hard to meet the required diversity criteria owing to data aggregation from identical domains or groups of people. Data completeness checks are mandatory in DC-AI, but this can be a hard task, particularly when the data engineers are not experts in the field of application. Handling a massive amount of data in which most instances are manually labeled poses serious challenges in the process of data auditing/governance. In the absence of automated tools, fixing data quality problems by identifying faulty parts from large-scale data is challenging. Also, applying DC-AI to numerous data types is tricky because data engineers may not be intimate with all data types/formats. Finally, one benefit of deep learning (over conventional ML) is a simplified data engineering process; therefore, most DC-AI criteria may be inherently fulfilled, and DC-AI may no longer be needed.

FUTURE PROSPECTS OF THE DC-AI PARADIGM

Soon, DC-AI will foster AI use in many commercial sectors, and AI can vastly contribute to the social good. For example, it can thwart the doubling data rule of MC-AI, which can contribute to lowering computing overhead. It encourages higher diversity in training data, which can help prevent unfair decision making with AI. It can also contribute to understanding the outcomes of AI models by offering higher transparency in

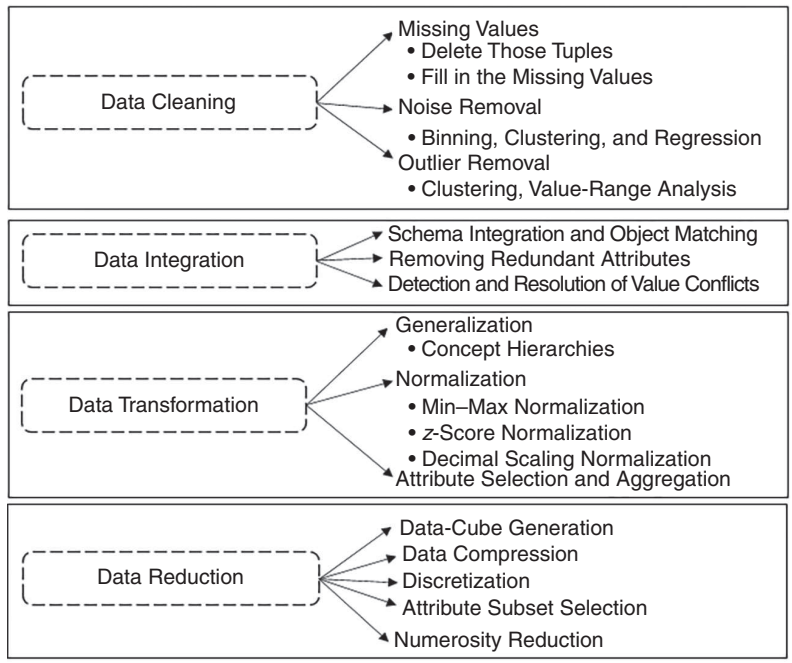

FIGURE 3. Overviews of the main techniques concerning (a) preprocessing and (b) DC-AI. SLA: service level agreement.

the training data as well as equal data utilization in model training. Here, we demonstrate the possible advantages that can likely be gained from DC-AI in the near future:

- It can significantly reduce computing overhead by collecting only necessary data (rather than doubling the data) when AI models yield deficient performance.
- It can foster AI use in industrial domains involving limited (or poor-quality) data.
- It can benefit from high-quality pretrained AI models rather than needlessly building models from scratch. Since DC-AI suggests taking advantage of the pretrained model as much as possible while improving data, a single line of code, such as `import BERT as bt`, can be sufficient to use a high-quality pretrained model with slight modifications in any new application.
- It can provide strong candidates to address performance issues (such as accuracy) by identifying vulnerabilities in data and correcting them before building AI models.
- It can augment the adoption of AI technology in multiple sectors, and make AI affordable for small businesses (for example, in the retail industry). There can be a lack of computing resources for small businesses owing to lower share/sales. However, building large-scale models requires computing infrastructure that can increase the operational cost for such businesses. Given the fact that DC-AI can yield desired results with the amalgamation of limited data and pretrained models, it can contribute to making AI affordable for small businesses.
- It can increase the lifespan of AI models by preventing data/concept drifts beforehand.
- It can increase the robustness and trustworthiness of AI by providing greater visibility in the training data and evaluating data *N* times.
- It can lessen any unintended consequences from AI technology on humans as well as make AI widely accessible and realizable.

> The amalgamation of DC-AI with preprocessing, MC-AI, and other relevant techniques is imperative to advancing AI technology.

Apart from these possible prospects cited, DC-AI can contribute to making AI more responsible and transformative.

DC-AI is a modern paradigm with lots of opportunities to enhance the technical efficacy of AI in terms of development, deployment, adoption, and governance. Challenges remain, such as overcoming misconceptions between DC-AI and preprocessing, becoming a complementary approach to MC-AI, and inherently fulfilling DC-AI in some cases, but they will be conquered as DC-AI use increases with AI developments. Finally, DC-AI is still in its infancy, and joint efforts from multiple stakeholders are imperative for adopting it systematically in various domains so that AI can become more effective in future endeavors.

ACKNOWLEDGMENT

This work was supported by a National Research Foundation of Korea grant funded by the Korea government (2020R1A2B5B01002145).

REFERENCES

1. 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, Apr. 2022, doi: 10.1109/mspec.2022.9754503.
2. A. Wu, "A chat with Andrew on Mlops: From model-centric to data-centric AI," 2021. [Online]. Available: https://www.youtube.com/06-AZXmwHjo
3. O. H. Hamid, "From model-centric to data-centric AI: A paradigm shift or rather a complementary approach?" in *Proc. 8th Int. Conf. Inf. Technol. Trends (ITT)*, 2022, pp. 196–199, doi: 10.1109/ITT56123.2022.9863935.
4. A. Anwar, "Difference between AlexNet, VGGNet, ResNet, and inception," *Towards Data Sci.*, Jun. 2019. [Online]. Available: https://towardsdatascience.com/the-w3h-of-alexnet-vggnet-resnet-and-inception7baaaecccc96
5. W. Liang et al., "Advances, challenges and opportunities in creating data for trustworthy AI," *Nature Mach. Intell.*, vol. 4, no. 8, pp. 669–677, Aug. 2022, doi: 10.1038/s42256-022-00516-1.
6. L. J. Miranda. "Towards data-centric machine learning: A short review." GitHub. Accessed: Dec. 22, 2022. [Online]. Available: https://ljvmiranda921.github.io/
7. C. Hegde, "Anomaly detection in time series data using data-centric AI," in *Proc. IEEE Int. Conf. Electron., Comput. Commun. Technol. (CONECCT)*, 2022, pp. 1–6, doi: 10.1109/CONECCT55679.2022.9865824.
8. M. Motamedi, N. Sakharnykh, and T. Kaldewey, "A data-centric approach for training deep neural networks with less data," 2021, *arXiv:2110.03613*.

9. T. Gebru et al., "Datasheets for datasets," *Commun. ACM*, vol. 64, no. 12, pp. 86–92, Nov. 2021, doi: 10.1145/3458723.
10. E. Jeczmionek and P. A. Kowalski, "Input reduction of convolutional neural networks with global sensitivity analysis as a data-centric approach," *Neurocomputing*, vol. 506, pp. 196–205, Sep. 2022, doi: 10.1016/j.neucom.2022.07.027.
11. A. Venigalla, J. Frankle, and M. Carbin. "BioMedLM: A domain-specific large language model for biomedical text." MosaicML. Accessed: Dec. 23, 2022. [Online]. Available: https://www.mosaicml.com/blog/introducing-pubmed-gpt
12. *Data-Centric AI*. (2022). Landing AI. [Online]. Available: https://landing.ai/data-centric-ai-old/
13. D. Berscheid. "Data-centric Machine Learning: Making customized ML solutions production-ready." Dida. Accessed Dec. 26. 2022. [Online]. Available: https://dida.do/blog/data-centric-machine-learning
14. Responsible Data Science. (2016). Mission. [Online]. Available: https://redasci.org/

ABDUL MAJEED is an assistant professor in the Department of Computer Engineering at Gachon University, Seongnam 13120, Korea. Contact him at ab09@gachon.ac.kr.

SEONG OUN HWANG is a professor in the Department of Computer Engineering at Gachon University, Seongnam 13120, Korea. He is a Senior Member of IEEE. Contact him at sohwang@gachon.ac.kr.

Digital Object Identifier 10.1109/MC.2023.3263914