Department: Head
Editor: Name, xxxx@email

Technical Analysis of Data-Centric and Model-Centric AI

Abdul Majeed
Department of Computer Engineering, Gachon University, Seongnam 13120, South Korea

Seong Oun Hwang
Department of Computer Engineering, Gachon University, Seongnam 13120, South Korea

***Abstract*—The AI field is going through a dramatic revolution in terms of new horizons for research and real-world applications, but some research trajectories in AI are becoming detrimental over time. Recently, there has been a growing call in the AI community to combat a dominant research trend named model-centric AI (MC-AI) which only fiddles with complex AI codes/algorithms. The MC-AI may not yield desirable results when applied to real-life problems like predictive maintenance owing to limited or poor-quality data. In contrast, a relatively new paradigm named data-centric (DC-AI) is becoming more popular in the AI community. It entails data debugging and quality enhancement rather than solely improving code and/or algorithms and is expected to create a huge impact in the AI field. In this paper, we discuss and compare MC-AI and DC-AI in terms of basic concepts, working mechanisms, and technical differences. Then, we highlight the potential benefits of the DC-AI approach to foster further research on this recent paradigm. This pioneering work on (DC+MC)AI can pave the way to understanding the fundamentals and significance of these two paradigms from a broader perspective.**

ARITIFICIAL INTELLIGENCE (AI) is a transformative technology with a wide range of practical applications in diverse sectors such as healthcare, defense, cybersecurity, and robotics. During the COVID-19 pandemic, AI has been widely used to forecast daily case tallies, gauge the efficacy of interventions, trace hidden routes of transmission, predict the course of the epidemic, and analyze trends, to name just a few applications [1]. Recently, researchers/practitioners have been expanding the horizon of AI applications from simple problems to global issues such as climate change. Addressing such global issues by utilizing AI will have a very big impact on people around the globe [2]. Apart from these applications, the amalgamation of AI with technologies like blockchain, edge computing, and other industry 4.0 technologies is rapidly increasing and has opened up various innovative use cases. Based on the discussion here, it is fair to say that AI is on its way to becoming a very helpful tool for humans to execute many tasks.

Recently, AI has emerged as a strong competitor of humans as it can perform many tasks

in less time and with minimal cost than humans. More efforts are underway in developing artificial general intelligence systems (e.g., the systems that can closely mimic the way human performs tasks.). As a result, a large # of new AI models, architectures, and low-code/no-code tools have been developed. The abilities of machine learning models are expanding from the classification/prediction tasks to predictive maintenance and other complex tasks [3]. The deep learning models combined with the IoT and other technologies are helping to combat the shortage of experts and resources in healthcare sectors [4]. Also, the advancement in federated learning and contrastive learning are improving the privacy and usability of data. The developments in generative AI are assisting in curating more data to compensate for the deficiency of data and to improve the results of AI models. The latest generative AI tools such as ChatGPT have many innovative use cases (e.g., code writing, scientific paper writing, answering questions, virtual assistants, etc.) [5]. The forthcoming wave of AI will bring more powerful and innovative tools for diverse sectors.

Before the inception of DC-AI, most efforts were put into an MC-AI approach that puts special focus on improving the architectural aspects of AI models (e.g., modifying the network architecture, switching to a new model, reducing model size, and hyperparameter tuning). Using this approach when an AI model fails to yield the required performance, developers only improve architectural aspects. This might not apply to some scenarios when data are limited (or are of poor quality) and when further data acquisition is difficult owing to a limited budget. Another main drawback of this approach is $\times 2$ the data, meaning if an AI model fails to yield the required accuracy, the developers get more data irrespective of the fact that only a few images/features might be faulty. This can waste time and effort and increase computing overhead.

Thanks to the discovery of Prof. Andrew Ng, the deficiency of large datasets can easily be overcome by rigorously using the DC-AI approach [6]. In this approach, when an AI model gives poor performance, the developers need to inspect the data as well, rather than solely improving the code. DC-AI can overcome the potential drawbacks of the MC-AI approach and can reduce overhead by collecting the required images/features, rather than simply doubling the data. Furthermore, DC-AI can increase the accuracy of convolutional neural network (CNN) models by using even less, but good-quality data [7]. It can be widely applicable to scenarios where the commodity of data does not exist, or when getting more data is difficult. Thus far, very little is known about these two paradigms, and a concrete overview of their workflow and key differences remains unexplored. The main contributions of this work are summarized below.

- We explore two schools of thought (DC-AI and MC-AI) concerning the development of AI technology, and we identify opportunities to provide concrete technical details and insights about them. Specifically, we present a technical analysis of DC-AI and MC-AI paradigms, and we highlight the key differences between them.
- We pinpoint and describe six dimensions to systematically highlight the MC-AI approach of AI developments that remained unexplored in the current literature.
- We analyze different techniques that can be vital in realizing DC-AI, and we group them into three levels to systematically demonstrate what DC-AI entails.
- We demonstrate the potential benefits of DC-AI when solving many key issues in the current AI technology. To the best of our knowledge, this is the first work centering on DC-AI and MC-AI, and it can provide a good foundation for future research in this line of work.

This work's four differences from the published article [6] are: (i) It identifies and discusses the categories of MC-AI developments from the perspective of six dimensions, (ii) It identifies and groups techniques that can be vital to enhancing data quality and realizing DC-AI, (iii) It provides the workflow of MC-AI and DC-AI when solving a real-world problem using AI, and (iv) it pinpoints the potential benefits of DC-AI from a broader perspective than previously anticipated.

Model-centric AI and Data-centric AI

Fig. 1 illustrates the workflows of both DC-AI and MC-AI paradigms in real-world scenarios. We define both paradigms as follows.

- In MC-AI, developers usually pay more at-

This article has been accepted for publication in IT Professional. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MITP.2023.3322410

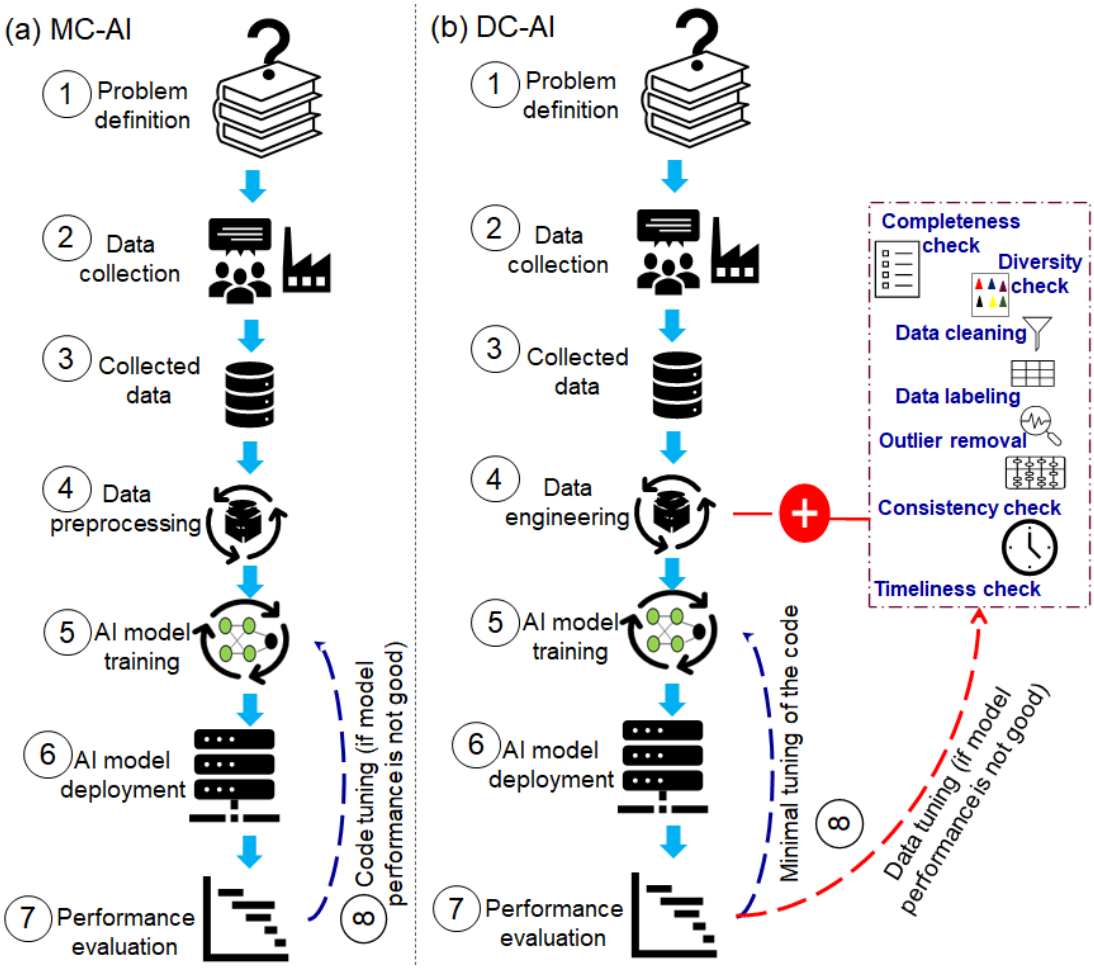

Figure 1: Workflows of (a) MC-AI and (b) DC-AI when adopted to solve real-life problems.

tention to optimizing the model's codes while rarely inspecting the data. MC-AI can be formally expressed in Eq. 1:

$$MC - AI = C^{'} + D \tag{1}$$

- In DC-AI, developers need to look into the data along with iteratively improving algorithms and/or codes. Specifically, developers should iteratively investigate and enhance the data along with tweaking the AI model. DC-AI can be formally expressed in Eq. 2:

$$DC - AI = C + D^{'} \tag{2}$$

In Eqs. 1 and 2, C and D refer to code and data. The $'$ sign over C and D indicates the priorities in the respective paradigm. In real settings, C is the code of any AI model, and D is data enclosed in any modality (e.g., table, images, text, etc.).

In MC-AI, developers improve the codes/algorithms only when the AI model yields poor results (Step # 8 in Fig. 1(a)). In contrast, both data and the AI model's codes/algorithms are jointly inspected in DC-AI when the AI model yields poor results (Step # 8 in Fig. 1(b)). Also, data are significantly improved in Step # 4 before being fed into AI models.

Six noteworthy dimensions of research/developments in model-centric AI

MC-AI has significantly contributed to advancing the technical potency of AI when solving many real-life problems. Some of the major problems are natural language processing, emotion detection, human activity recognition, and pandemic mitigation [8], [9]. Researchers have explored MC-AI from multiple perspectives, but most of those are related to improving the architectural

aspects (e.g., the code of AI models). To provide a clear overview of developments concerning MC-AI, we classify major research/developments into six broad dimensions, as discussed below.

1) *Ever-expanding horizons of AI applications*. In the beginning, AI was mostly confined to the computer science field and was used/investigated by computer scientists. However, with time, AI has expanded to many other disciplines. Currently, in most sectors, AI has been rigorously used to accomplish multiple objectives, and AI has taken over some jobs from humans, such as gauging the amount of liquid in water/wine bottles. Specifically, AI applications in the healthcare sector are booming. The recent pandemic sparked the use of AI in the healthcare sector [1]. Since data used to train AI models can vary from application to application, developers need to pay ample attention to the data for each particular application.
2) *Advancements in network architectures*. In the early days of AI development, the mapping of input (X) to output (Y) was governed by a few intermediate layers. In the perceptron model, computer scientists are interested in determining whether a mathematical function can map a vector of numbers to some specific target class or not. However, these methods are simple and require much greater human involvement. The landscape of AI models changed, and computer scientists became interested in limiting human involvement, so perceptron evolved into neural networks that require less feature engineering (or human involvement). Consequently, a simple perceptron model that can work well with simple data was enhanced to a complex neural network to solve problems in which the input can be images/videos. In other words, a simple binary mathematical function was replaced with layers and channels, and multiple interactions performed between layers determine the output. These enhancements in the network architecture improved the technical status of AI technology. Today, a mammoth of network architectures exist that can be used to solve any real-world problem.
3) *Optimization (pruning) in the architecture*. Now, there are plenty of AI models, each having a different workflow in terms of network size and architecture, data processing mechanism, number and types of parameters, convergence speed, etc. However, in some cases, AI models need to operate on resource-constrained and tiny devices, like microcontroller units (MCUs), and deploying complex AI models (e.g., a CNN) on such devices can be impossible. To this end, researchers/developers have explored various ways to reduce the size of AI models so they can operate in resource-constrained environments/devices [10]. Pruning and quantization techniques have been devised to remove redundant parameters/weights, skip some layers, and/or combine multiple layers to reduce the AI model's size [11]. Through such code-based techniques, memory and CPU time in AI models improved significantly.
4) *New model development*. The AI community has shown ever-increasing interest in the development of new AI models from slight modifications to prior models. The main motivations behind these developments are huge parameters, computing overhead, deficiencies in salient feature extraction, and model size. For example, the CNN was upgraded to a new variant named ResNet-18, which is relatively deeper and has more layers. Consequently, this new model has greater abilities in terms of extracting salient features from training data, and can yield better performance. However, ResNet-18 is prone to the vanishing gradient problem, and has greater complexity due to its many multiplication operations. Similarly, the inception AI model is needed when important global features are distributed in more parts of the images, and a fixed kernel size may not yield desirable results. The inception model goes wider, rather than as deep as conventional CNN models. Thus far, many AI models have been proposed, and more developments are expected in the near future. In Table 1, we

provide a comparative analysis of famous AI models along with relevant technical details. From the analysis, we can see that AI developers are rapidly advancing AI models to address issues stemming from AI use/governance.

Table 1: Comparative analysis of AI models (adopted from [12]).

Network (Yr)	Salient feature	Parameters	Top5 accuracy	FLOPs
AlexNet (2012)	Deeper	62M	84.7%	1.5B
VGGNet (2014)	FS kernels	138M	92.3%	19.6B
Inception (2014)	W-P kernels	6.4M	93.3%	2B
ResNet-152 (2015)	SC connections	60.3M	95.51%	11B

Abbreviations: FLOPs: floating point operations, FS: fixed-size, W-P: wider-parallel, SC: shortcut, M: million, B: billion

5) *Advancement in data modalities and AI model training*. Before 2016, data acquisition at some central place was necessary in order to train AI models. However, with the introduction of the federated learning (FL) concept, data acquisition at some central place is no longer required, and AI models can still be trained in a distributed manner [13]. With the inception of FL, AI developers shifted towards new data modalities and optimized ways of AI model training by using local data. Recently, many challenges and problems with FL deployment concerning client selection, model distribution, and privacy and security of the FL ecosystem have been observed by researchers [14]. Therefore, developers are exploring ways to optimize the performance of FL ecosystems by utilizing improved codes/algorithms.
6) *Hyperparameter tuning to yield better performance*. In most cases, developers try different combinations of hyperparameters such as batch size, α, sampling technique, filter size, cyclical momentum, etc., to yield desirable results with AI models. The selection of optimal combinations of hyperparameters for diverse applications is challenging, and needs careful implementation [15]. In conventional AI implementations, the trial-and-error method is adopted to determine the optimal values of hyperparameters, leading to extensive computing overhead. In case of poor performance, developers either change hyperparameter values or change the network structure, which can slow down the development of AI models.

In most of the dimensions discussed above, the main focus of developers revolves around the code; they rarely inspect the data. However, the data constitute a vital component in AI technology development and significantly contribute to the quality of AI systems. In Table 2, we summarize all six dimensions and highlight their main priorities. From the analysis, we found that MC-AI gives minimal preference to the data, and therefore, fiddling with code may not yield desirable solutions to many industrial problems.

Table 2: Six research dimensions centered on the MC-AI approach.

Dimension #	Main priority (or focus)	Data investigation
1	Ever-expanding horizons of AI applications	Yes
2	Improving network architectures of AI models	Rare
3	Pruning/optimizing the network architecture	Rare
4	Proposing new models of AI (or upgraded versions)	Rare
5	Devising new modalities and AI model training	Rare
6	Improving hyperparameters of AI models	Rare

Three-level data-centric AI paradigm

DC-AI is a very recent paradigm that explores ways to improve data quality in order to enhance the performance of AI models [16], [17]. Table 3 presents the core approaches of the DC-AI.

Table 3: Salient approaches of the DC-AI paradigm that can make AI more effective.

Level 3	Level 2	Level 1
DFS	Data quality	Consistency Accuracy Completeness Timeliness Metadata Effectiveness Relevance
	Data availability	Readiness Service level agreement
	Data Observability	Properties Patterns Metrics States Statistics
IDA	Visibility of all data	Higher knowledge about data Data versioning
	Moving data to the right place	Effective data utilization Efficiency
	Seamless access to data	Removing property lock-in
	Control on data	Monitoring data flows
DC	Correct data use	Define access levels Document unfair practices
	Ethically complaint data use	Explain the risk of data misuse Transparent models

Abbreviations: DFS= data-first strategy, IDA= intelligent data structure, DC= data compliance

Specifically, we classify various approaches that can be employed as part of DC-AI into three levels, described below.

1) *First level:* This includes 24 basic approaches that come under the DC-AI umbrella. Most approaches can be applied to the initial phase of AI system development. For example, it is vital to collect only relevant and necessary data concerning the problem, and there exists a data relevance approach at this level to ensure it. Similarly, sensitive data classification and risk of misuse can be assessed to guarantee better protection of sensitive data in the life cycle. Furthermore, analysis of data completeness is desirable at this level to prevent performance degradation issues in AI systems. Satisfying most approaches at the first level can prevent inadvertently propagating data-specific biases to the other levels, which can contribute to the development of effective AI systems to solve real-world problems.
2) *Second level:* This level includes nine different approaches that are relatively more sophisticated and advanced than in the prior level. These approaches enhance the quality of data and can be employed to determine if the data are complete from most aspects. These approaches empower AI developers to have strong control over the data, and therefore, all parts can be equally used in the training/development of AI systems. Most of the approaches at this level are multi-criteria, meaning multiple coefficients can be used to quantify the level of each approach. Eq. 3 is an example of data quality estimation using a multi-criteria method:

$$\begin{aligned} D_q = {} & w_1 \times Acc + w_2 \times Con + \\ & w_3 \times Com + w_4 \times Met + \\ & w_5 \times Tim + w_6 \times Rel + w_7 \times Eff \end{aligned} \tag{3}$$

where D_q refers to data quality, and Acc, Con, Com, Met, Tim, Rel, and Eff denote accuracy, consistency, completeness, metadata, timeliness, relevance, and effectiveness, respectively. These parameters can be quantified using mathematical formulas or numerical scores given by domain experts based on data judgment. The formula for computing Acc is expressed below.

$$Acc = \frac{C}{A} \tag{4}$$

where C denotes the # of samples that are recognized correctly, and A denotes the total # of samples in a dataset.
Similarly, the value of Con can be quantified using the below equation.

$$Con = \frac{C_{index}}{R_{index}} = \frac{(\frac{\lambda_{max} - n)}{n-1}}{R_{index}} \tag{5}$$

where C_{index} is the consistency index, R_{index} is the random consistency index, and n is the # of observation in the data. The value of R_{index} is determined using a lookup table by passing n as a parameter.
The Com value can be quantified via Eq.6.

$$Com = 1 - (\frac{\Phi}{n} \times 100) \tag{6}$$

where Φ denotes the # of missing values, and n denotes the total # of entries. For example, for a dataset having 500 records with 110 missing values, the Com is 78%.
Met is detailed information of a column/dataset. It is ideal to analyze the metadata before building ML models. For example, the distribution skew is a very common problem in ML, and it can be computed using the below formula.

$$Met = \frac{c_M}{c_m} \tag{7}$$

where c_M denotes the instances from the major class, and c_m denotes the instances from minor class. In real cases, the distribution/frequencies of the column can be computed and used in Met.
The Tim can be quantified by taking the difference between data curation time (T_c) and data use time (T_u).

$$Tim = T_c - T_u \tag{8}$$

The value of Tim can be compared with some threshold t to decide about data acceptance/unacceptance as expressed below.

$$Tim = \begin{cases} unacceptable, & \text{if } Tim \geq t \\ acceptable, & \text{otherwise} \end{cases} \quad (9)$$

The Rel can be quantified using Eq. 10.

$$Rel = \frac{S_f}{T_f} \quad (10)$$

where S_f denotes the salient features, and T_f refers to the total # of features. Rel can also be used to draw relevant samples out of the total samples.

The Eff can be quantified using Eq. 11.

$$Eff = \frac{A}{D} \quad (11)$$

where A is the achieved accuracy/data, and D is desirable accuracy/data.

In Eq. 3, w_i, where $i = 1$ to 7, denotes weights of each parameter. The range of w_i coefficients is between 0 and 1 (e.g., $w_i > 0$), and $\sum_{i=1}^{7} w_i = 1$. The optimal values of coefficients can be specified by domain experts or can be adjusted based on the importance/problem. It is worth noting that some of the above parameters can be quantified using the built-in functions in some software (e.g., MS Excel) or assigning 1/0 based on expertise. Also, the data quality estimation can vary depending on the data modality [18]. By guaranteeing most approaches, data-specific bias can be significantly restrained.

3) *Third level:* This level includes three distinct approaches (a.k.a. building blocks) concerning DC-AI. These approaches are more cutting-edge and advanced than the bottom and intermediate levels. In these blocks, all of the previous 33 approaches are analyzed, and further opportunities are explored to improve data quality. In addition, the decision can be made to reassess the downstream approaches depending on the problem at hand. For example, in some cases, all data cannot be sensitive, and therefore, less attention can be paid to the data compliance block, compared to the other two blocks. Similarly, if we need to address only social problems in AI systems, then the data-first strategy requires closer attention, compared to the other two blocks. In some cases, pilot projects (or prototypes) can be developed to assess the efficacy of these components before building the actual product/system. It is important to note that some approaches may not be needed all the time, and therefore, further investigation is needed to choose suitable DC-AI approaches in real-world scenarios.

By using all approaches encompassed in these three levels, good-quality data can be curated.

Potential Benefits of DC-AI

DC-AI explores various ways to make AI technology more effective for human beings. In the future, it can bring many benefits to AI developers and consumers. We summarize the potential benefits of the DC-AI in Fig. 2.

- ⌘ Significant reduction in computing overhead
- ⌘ Reduction in time from AI model development to deployment by using pretrained models
- ⌘ Higher adoption of AI in the data-constrained domains (e.g., limited or low-quality data)
- ⌘ Possible remedy for performance issues by identifying and fixing data-related vulnerabilities
- ⌘ Potential solution to AI technology affordance & fragmentation issues by using good data
- ⌘ Extend lifespan of AI models and increase the success rate in real-world scenarios
- ⌘ Better AI adoption and higher robustness by offering higher visibility in training data
- ⌘ Lower societal risks (AI misuse) by using high quality data and improved environment health
- ⌘ Able to solve global issues (e.g., climate change) by ensuring data relevance and completeness
- ⌘ Contribute to opening the black-box nature of most AI models by explaining data use in them
- ⌘ Foster responsible and fair use of AI technology

Figure 2: Potential benefits of the DC-AI.

For example, DC-AI can reduce computing overhead by identifying faulty parts of the data and fixing (or augmenting) them rather than doubling the data as MC-AI does. It can foster AI's transition from academic labs to the market by exploiting the benefits of pretrained models to the extent possible, rather than rebuilding models from scratch. It can extend AI adoption to multiple domains involving limited data. DC-AI can also enhance accuracy by correctly labeling the data and involving multiple domain experts in

the labeling process. It can make AI technology more accessible and understandable to the non-expert. It can extend the AI models' lifespan by paying ample attention to data drift during development. DC-AI can increase robustness by ensuring data availability and control. It can solve the longstanding problems in AI (unfair decisions, explainability, trust, etc.) by applying data engineering practices. It can likely solve global challenges of climate change and supply chain disruptions through data-tailored actions. Lastly, it can assist in understanding the curious nature of AI models by explaining how data are used in them. It can contribute to the responsible use of AI, which is an urgent requirement amid the rapid rise in AI applications.

Lastly, DC-AI can effectively contribute to efficiently extracting causal and temporal relations from a limited time series corpus. It can contribute to developing lightweight models that can create general causal graphs as a representation that can be used to forecast future time series data/values. DC-AI provides means to curate synthetic time series corpus which can be used to enhance the performance of AI models when combined with original time series corpus. DC-AI can also contribute to causal features extraction in a short time which improves the robustness and computing efficiency of applications involving limited time series corpus.

Discussion

Outcomes of the research: This work uncovers the details of the MC-AI and DC-AI for the readers, which can pave the way to understanding these two schools of thought used for AI performance enhancement. The benefits of DCAI explored in this work can contribute to making AI more beneficial, in particular, overcoming the societal risks of AI which are debatable issues around the globe. It pinpoints that DC-AI can likely open the black box nature of AI, which may assist in understanding the workflow of most AI models. Explainable and fair AI systems are an urgent need in some specific sectors such as healthcare. It provides a new perspective toward building more data engineering techniques as well as systematically applying DC-AI techniques that can extend the application horizon of AI. It provides workflows for DC-AI systems, which can allow the development of prototypes and proof of concepts. Lastly, this work aligns with recent trends toward making data AI-ready and improving data quality before feeding to AI models.

*Guidelines regarding the use cases for which DC-AI should be used:*DC-AI can be applied in any situation, however, it is more suitable to those scenarios involving fewer, no, and/or low-quality data. Furthermore, it is an ideal candidate when getting more data is either difficult or data collection budgets are very small. It is also handy when data is coming from diverse sources and in different modalities (i.e., tables, time series, etc.). DC-AI is also needed in scenarios where outcomes/decisions are solely made based on training data such as chatbot [19]. DC-AI is inevitable in sensor-powered systems such as IoT-based medical diagnosis systems and predictive maintenance. For example, in the predictive maintenance use case, it can contribute to data alignment, consistency, and fusion, which can be useful in identifying faulty machines proactively. Autonomous vehicles collect data from various sensors to make reliable decisions in real time. To this end, DC-AI can play a vital role in preventing data incidents. In a climate change scenario, it can ensure relevant data collection, and quality enrichment, which can help better understand the dynamics of climate change. In Tiny ML, it can prevent the possibility of data drift, and extensive re-training of AI models. It is also very important in equity and inclusion scenarios to prevent conflict by curating diverse data. Recently, it has been used to detect fraud in healthcare applications by simply preparing and understanding data [20]. Similarly, there are many use cases where DC-AI can be used such as language models, time-series forecasting, epidemic analysis, human activity recognition, etc. Lastly, data is the cornerstone for AI developments, and therefore, DC-AI can vastly contribute to enhancing its quality, leading to significant performance enhancement in AI.

Concluding Remarks

This article provided an in-depth analysis of MC-AI and DC-AI, which are two main research trends of AI technology development. MC-AI is a widely used approach that leverages AI to solve real-world problems, but it mostly focuses on improving only the code in AI models. Recently,

many ill effects of MC-AI, such as limited applicability in data-constrained domains, societal risks, and extensive computing overhead, have been observed that require urgent solutions to develop lightweight, safe, and reliable AI solutions. DC-AI is a fledgling paradigm and will provide the means to fully/partially resolve most of the drawbacks in MC-AI by rigorously improving one of the key elements (i.e., data) of the AI ecosystem, rather than improving only the code. Exploring ways to properly amalgamate these two approaches by identifying suitable application scenarios will enhance the AI transition from academic labs to the market, and will eventually solve many longstanding problems in conventional AI. Our work is an initial step toward representing the efficacy of DC-AI, which can pave the way to enabling AI for social good.

ACKNOWLEDGMENT

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

REFERENCES

1. M. AL-Hashimi and A. Hamdan, "The applications of artificial intelligence to control covid-19," in *Advances in Data Science and Intelligent Data Communication Technologies for COVID-19*. Springer, 2022, pp. 55–75.
2. W. Leal Filho, T. Wall, S. A. R. Mucova, G. J. Nagy, A.-L. Balogun, J. M. Luetz, A. W. Ng, M. Kovaleva, F. M. S. Azam, F. Alves *et al.*, "Deploying artificial intelligence for climate change adaptation," *Technological Forecasting and Social Change*, vol. 180, p. 121662, 2022.
3. O. Surucu, S. A. Gadsden, and J. Yawney, "Condition monitoring using machine learning: A review of theory, applications, and recent advances," *Expert Systems with Applications*, vol. 221, p. 119738, 2023.
4. Z. Lv, J. Guo, and H. Lv, "Deep learning-empowered clinical big data analytics in healthcare digital twins," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2023.
5. A. Bahrini, M. Khamoshifar, H. Abbasimehr, R. J. Riggs, M. Esmaeili, R. M. Majdabadkohne, and M. Pasehvar, "Chatgpt: Applications, opportunities, and threats," in *2023 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, 2023, pp. 274–279.
6. E. Strickland, "Andrew ng, ai minimalist: The machine-learning pioneer says small is the new big," *IEEE Spectrum*, vol. 59, no. 4, pp. 22–50, 2022.
7. 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, 2022.
8. S. Geravesh and V. Rupapara, "Artificial neural networks for human activity recognition using sensor based dataset," *Multimedia Tools and Applications*, vol. 82, no. 10, pp. 14 815–14 835, 2023.
9. A. H. Shamman, A. A. Hadi, A. R. Ramul, M. M. A. Zahra, and H. M. Gheni, "The artificial intelligence (ai) role for tackling against covid-19 pandemic," *Materials Today: Proceedings*, vol. 80, pp. 3663–3667, 2023.
10. Q. Huang, "Weight-quantized squeezenet for resource-constrained robot vacuums for indoor obstacle classification," *AI*, vol. 3, no. 1, pp. 180–193, 2022.
11. J. Torres-Tello and S.-B. Ko, "Optimizing a multispectral-images-based dl model, through feature selection, pruning and quantization," in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2022, pp. 1352–1356.
12. A. Anwar, "Difference between alexnet, vggnet, resnet and inception," *Medium-Towards Data Science*, 2019.
13. Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, "Federated learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 13, no. 3, pp. 1–207, 2019.
14. A. Majeed, X. Zhang, and S. O. Hwang, "Applications and challenges of federated learning paradigm in the big data era with special emphasis on covid-19," *Big Data and Cognitive Computing*, vol. 6, no. 4, p. 127, 2022.
15. L. Wu, G. Perin, and S. Picek, "I choose you: Automated hyperparameter tuning for deep learning-based side-channel analysis," *IEEE Transactions on Emerging Topics in Computing*, 2022.
16. L. Schmarje, M. Santarossa, S.-M. Schröder, C. Zelenka, R. Kiko, J. Stracke, N. Volkmann, and R. Koch, "A data-centric approach for improving ambiguous labels with combined semi-supervised classification and clustering," in *European Conference on Computer Vision*. Springer, 2022, pp. 363–380.
17. A. Majeed and S. O. Hwang, "Data-centric artificial intelligence, preprocessing, and the quest for transformative artificial intelligence systems development," *Computer*, vol. 56, no. 5, pp. 109–115, 2023.
18. I. Taleb, M. A. Serhani, C. Bouhaddioui, and R. Dssouli, "Big data quality framework: a holistic approach to

continuous quality management," *Journal of Big Data*, vol. 8, no. 1, pp. 1–41, 2021.

19. U. M. Fayyad, "From stochastic parrots to intelligent assistants—the secrets of data and human interventions," *IEEE Intelligent Systems*, vol. 38, no. 3, pp. 63–67, 2023.
20. J. M. Johnson and T. M. Khoshgoftaar, "Data-centric ai for healthcare fraud detection," *SN Computer Science*, vol. 4, no. 4, p. 389, 2023.

Abdul Majeed is an Assistant Professor with the Department of Computer Engineering, Gachon University, South Korea. His research interests include privacy-preserving data publishing, information privacy, federated learning, data-centric AI, and machine learning. Contact him at ab09@gachon.ac.kr.

Seong Oun Hwang is a Professor with the Department of Computer Engineering, Gachon University, South Korea. He is a senior member of IEEE. His research interests include cryptography, data-centric AI, cybersecurity, and artificial intelligence. He is the corresponding author of this article. Contact him at sohwang@gachon.ac.kr.