MIJRD Multidisciplinary International Journal of Research and Development

Multidisciplinary International Journal of Research and Development

Volume: 04 / Issue: 04 / 2025 - Open Access - Website: www.mijrd.com - ISSN: 2583-0406

Data-Centric AI: A Systematic Review of Methods, Challenges, and Future Directions

Suhasnadh Reddy Veluru[1], Sai Teja Erukude[2], and Viswa Chaitanya Marella[3]

[1,2,3]Kansas State University, Manhattan, KS 66502, USA

Email: [1]suhasnadhreddyveluru@gmail.com, [2]erukude.saiteja@gmail.com and [3]viswachaitanya20@gmail.com

Abstract— Model architectures for machine learning have become increasingly strong as models have developed over the last decade. However, we have started to plateau anytime we try only to improve performance through model architecture. This trend has led to another area of focus: the data itself, a fundamental yet largely forgotten aspect of an AI system. Data-centric AI (DCAI) describes systematically improving datasets to improve machine learning performance. In this paper, we thoroughly examine the landscape of DCAI by synthesizing the perspectives of recent developments in literature published between 2022 and 2025, focusing on data quality, data cleaning, data labeling, data augmentation, and data monitoring. We discuss the methods and tools of DCAI, the successful utilization of DCAI in healthcare, computer vision, and privacy-preserving synthetic data, and the significant difficulties DCAI faces, including cost, bias, and evaluation. Lastly, we discuss exciting future directions, including automating the data pipeline and moving to a more holistic approach to dataset model co-design. The transition from model-centered to data-centered development is foundational to developing better, more reliable, fairer, and more beneficial AI systems everywhere.

Keywords— Data-Centric AI, Data Quality, Data Augmentation, Bias Mitigation, Automation in AI Pipelines, Synthetic Data Generation.

I. INTRODUCTION

Artificial Intelligence (AI) has developed significantly in the last ten years, and this evolution has been underpinned by advancements in model family architectures such as convolutional neural networks, recurrent networks, transformers, and diffusion models (Zha et al., 2023; Ng, The Data-Centric AI Approach). However, as models grow more complex, gains from model-focused approaches seem to diminish. This has generated a significant field of research, and many papers now show that after a certain point, model improvements contribute less to the overall performance than the quality of the data (Zha et al., 2023; Zhou et al., 2024). This realization has fostered a nascent but rapidly growing movement of Data-Centric AI (DCAI), where the focus has shifted to improving datasets over models. DCAI asserts that limitations in today's systems are frequently not based on the insufficiency of models with respect to hyperparameter tuning; instead, they come from the quality of data. Increase the quality of the data, and it needs to be large, cleansed, fully represented, and encompass real-world environments (Zha et al., 2023; Zhou et al., 2024; Lu et al., 2023).

DCAI emphasizes treating data dynamically rather than the fixed data found in typical traditional datasets. Dynamic data captures include building the data incrementally while iterative engineering and refining the data. Dynamic data includes exposing inconsistencies in labeling, removing noise in the data, correcting bias in data,

Volume: 04 / Issue: 04 / 2025 - Open Access - Website: www.mijrd.com - ISSN: 2583-0406

and improving real-world representative data (Zha et al., 2023; Mazumder et al., 2022). The potential benefits of DCAI are significant in zero-sum, high-stakes professions such as those found in healthcare, finance, and autonomy, where improvements in flawed data can lead to a harmful practice (Gulamali et al., 2023). As AI systems proliferate into many aspects of society, the ethical obligation to ensure fairness, robustness, and transparency requires dataset-level ethical responsibilities.

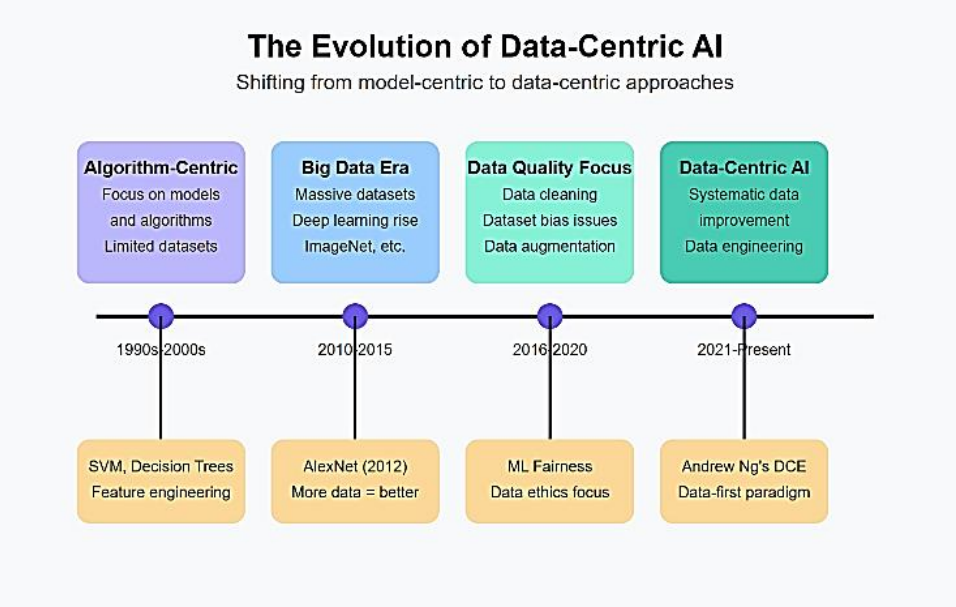

Figure 1: Timeline showing the paradigm shift from algorithm-centric approaches (1990s–2000s) to data-centric AI (2021–present), highlighting key milestones such as the rise of deep learning, focus on data quality, and the emergence of data engineering as a critical component.

II. RELATED WORK

Machine learning scholars were aware of the importance of data quality before the formal conception of Data-Centric AI (DCAI). The literature demonstrates that label noise, sampling bias, and lack of coverage can harm model generalization. However, the machine-learning community at large has prioritized new algorithms over data scrutiny (Zha et al., 2023; Zhou et al., 2024).

Zha et al. (Zha et al., 2023) developed the field by defining the data lifecycle with three main pieces: preparing training data, preparing inference data, and maintaining the datasets. Zha et al. (Zha et al., 2023) advocated for the iterative process of developing data through interventions including label clean-up and developing test datasets while emphasizing the iterative, uncertain nature of 'real-world' AI ecosystems.

Using this concept as a foundation, Zhou et al. (Zhou et al., 2024) suggested a framework for data quality that separated intrinsic (accuracy, completeness), contextual (relevance, timeliness), and accessibility (usability) data quality. Zhou et al. (Zhou et al., 2024) also documented the emerging tools associated with data profiling and cleaning, highlighting that applications and domains characterize the problem of data.

Various benchmarking initiatives, such as the DataPerf competition (Mazumder et al., 2022), have progressed in establishing measures of quality while prioritizing data quality as opposed to model quality. The tasks involved capturing mislabeled data, building a core dataset, and augmenting the dataset to improve robustness, with data interventions as the primary development process.

Applied developments indicate that DCAI is gaining relevance. Gulamani et al. (Gulamali et al., 2023) showed that by explicitly addressing hidden bias in healthcare datasets, they could improve fairness without having to change the model's architecture. Similarly, Sai Teja et al. (Erukude et al., 2024; Erukude, 2024) showed that CNN classifiers were sensitive to background perturbations, noting that simply forcing the classifiers to consider robustness in their dataset design provided another dimension to the problem. These examples reinforce an industry shift toward systematic data engineering as AI systems mature.

III. METHODOLOGY AND BACKGROUND ON DATA-CENTRIC AI

A. Data Collection

The success of an AI system depends on how good its data is. In Data-Centric AI (DCAI), data collection is about creating a large set of diverse, representative, and de-biased datasets. In the past, organizations would collect data opportunistically; in the DCAI paradigm, data collection means sampling specific groups to avoid demographics, environments, edge cases, etc., biases that limit model generalization.

Active learning approaches are used to strategically select data samples based on uncertainty coverage or uniqueness (Zha et al., 2023). Adversarial sampling can also enhance a dataset with hard-to-classify examples that would be lost to standard sampling. The main interests here are not about datasets with numbers but datasets with representative coverage of the true complexity of the task environment.

Poor sampling could incur significant risk in high-risk domains, such as healthcare, finance, and autonomous systems. Hence, organizations now write formal data reports comprising dataset characteristics and collection procedures. The agenda of accessing rigor is complicated by various unavoidable limitations, and organizations document the purposes and ways datasets were intended to be used (Zhou et al., 2024).

In DCAI, data collection becomes a systematic design process rather than opportunistic data aggregation to ensure representativeness, fairness, and robustness.

B. Data Labeling

Once data has been collected, consistent, accurate labeling is paramount. Label noise is perhaps the most detrimental source of error in model performance, particularly with large volumes of supervised instances (Zha et al., 2023). If we mislabel only a small fraction of the examples, we can greatly change our model’s decision boundaries and generalization.

Manual labeling is standard; however, it is expensive and often unreliable because of subjective interpretations, annotator fatigue, or unclear labeling guidelines. To remedy this, a variety of quality practices are being used in the latest DCAI (data-centric artificial intelligence), e.g., clear labeling direction, majority voting from redundant labeling, and expert evaluation for harder cases.

Volume: 04 / Issue: 04 / 2025 - Open Access - Website: www.mijrd.com - ISSN: 2583-0406

Programmatic labeling approaches, such as weak supervision (Zha et al., 2023), are on the rise, given that they encode heuristic labeling functions and knowledge bases to label and leverage at scale. Sometimes, noise correction approaches (e.g., Snorkel) through statistical methods are used to estimate and suppress known or unknown noise.

Thus, modern data-centric annotation distinguishes between human-structured interventions and programmatic labeling that can happen at scale to facilitate more efficiency and reliability.

C. Data Cleaning

The datasets we use are likely to contain noise from real-world usage, such as mislabeled examples, corrupted files, duplicates, outliers, and more. In data cleaning, we need to systematically address nuisance variables so that the technologies developed do not rely upon artificial training instances that may affect learning.

There has been a recent emphasis on error detection rather than error correction (Zhou et al., 2024). Tools such as CleanLab employ cross-validation techniques and identify where the labels are uncertain to help find probable labeling errors. Statistical foils, detecting clustering inconsistencies, and similarity-based deduplication further improve the quality of the dataset.

As discussed in DCAI, one view of data cleaning is as a continuous or cyclical process rather than a one-off task (Zha et al., 2023). Models will be improved over time by using the latest models. However, models will also uncover new modes of failure, which require reviewing and cleaning an out-of-date dataset. Effective cleaning does more than simply correct the obvious categories; it also takes out the subtle mistakes that have a confounding presence and ensures the evaluation dataset is relevant to the underlying context of deployment.

With insufficient cleaning, models risk learning shortcuts, which may lead to catastrophic failure under real-world variations (Gulamali et al., 2023).

D. Data Augmentation

Data augmentation is necessary when datasets are small, unbalanced, or diverse. In the context of DCAI, augmentation can impose variations on raw examples to enrich the data without necessarily changing models or dataset structure (Lu et al., 2023).

In typical computer vision, data augmentation comes from rotation, scaling, flipping, color jittering, or cropping to synthesize additional data. In recent work, such as AutoAugment, RandAugment, or adversarial augmentation, augmentation strategies are learned, and the best set of transformations is selected to maximize validation performance (Lu et al., 2023).

In NLP, data augmentation is challenging due to the need to maintain semantic meaning. Data augmentation strategies that have been shown to work include back-translation, synonym replacement, paraphrasing, transformer-based generation, and adversarially perturbing text (Zha et al., 2023).

In the case of synthetic data generation, we are augmenting in the sense that we can now generate entirely new examples based on generative models (GANs, VAEs, and diffusion models). Generated data can be used in support

Volume: 04 / Issue: 04 / 2025 - Open Access - Website: www.mijrd.com - ISSN: 2583-0406

of privacy-preserving Learning, especially when access to real-world data is strongly restricted when learning in private spaces (e.g., Health care) (Lu et al., 2023).

Thus, data augmentation is viewed in DCAI as the process of increasing dataset size with more examples and adding to the quality and diversity of data in order to ultimately create more robust models.

E. Data Maintenance

Data maintenance is an important but often underestimated part of DCAI. Because real-world environments are dynamic-characterized by distribution shifts, evolving populations, sensor drift, and changing adversaries, a static dataset cannot ensure durable model invariance.

Data maintenance requires continued monitoring for distributional changes; concept drifts, and ongoing label relevance (Zha et al., 2023; Zhou et al., 2024). Production systems typically analyze telemetry data generated during model deployment to surface new patterns that may be underrepresented or non-existent in the training data and thus feed this information back into retraining/development data gathering.

While dataset versioning has existed for some time, traditionally, it has lacked the robustness of versioning for models. However, contemporary systems such as Data Version Control (DVC) can systematically keep track of dataset changes and versioning so that reproducibility, audibility, and management of ML data pipelines are possible.

In DCAI, we must acknowledge that maintaining data quality is an ongoing activity, not a discrete task. In many respects, ensuring that the models continue adapting to environmental changes means the dataset is continually created and re-used for that purpose.

F. Automation in Data-Centric AI

Automation has now become an enabler of scale for DCAI methods. More manual early-stage methods of data cleaning, labeling, and augmentation meant more upfront work and data distribution manually. The huge rise in dataset size and the need to iterate quickly have led to the continuous development of data-automated systems.

For example, auto-labeling methods using foundation models or large language models (LLMs) enable near-instant bootstrapping for annotated datasets (Zha et al., 2023). Conversely, automated algorithms for label error detection can use unsupervised or semi-supervised methods to find error(s) or inconsistencies. Reinforcement learning agents can select features optimally, label features optimally, and execute the pipeline for generating synthetic features (Ying et al., 2025).

Of course, automation comes with new risks. Blindly relying on automation systems' outputs would mean propagating errors at scale without implementing further quality checks. Consequently, modern DCAI methods have begun promoting human-in-the-loop automation methods, wherein humans oversee, validate, and allow correcting misjudgments made by automated methods, keeping them trustworthy and valuable (Zha et al., 2023).

While it may feel melodramatic, the future of DCAI would imply that automation will be further embedded, yet allowing for continuous refinement with humans onboard brings the judgment of decision points.

Volume: 04 / Issue: 04 / 2025 - Open Access - Website: www.mijrd.com - ISSN: 2583-0406

VI. FUTURE DIRECTIONS

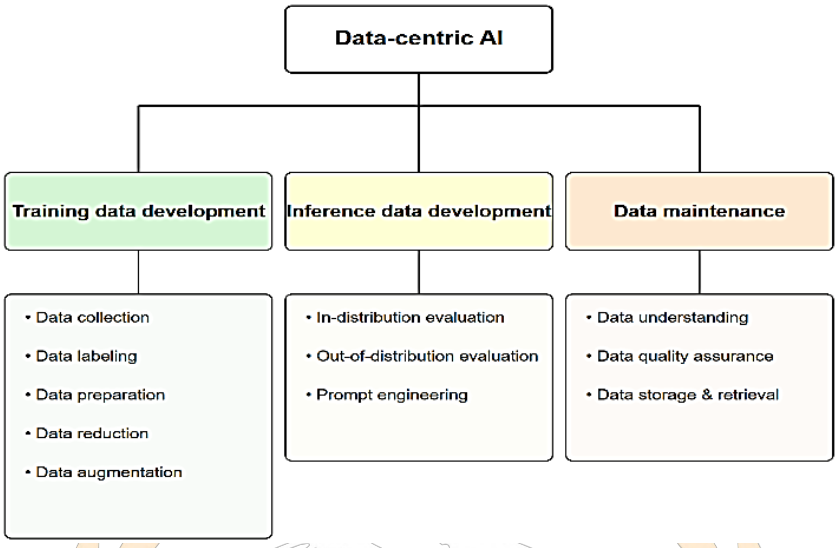

Figure 2: Overview of the main pillars of Data-Centric AI: Training Data Development, Inference Data Development, and Data Maintenance, with a focus on systematic data collection, preparation, evaluation, and quality assurance to improve machine learning outcomes.

IV. CASE STUDIES AND APPLICATIONS

The Data-Centric AI principles already demonstrate measurable impacts across data quality, augmentation, and management. In health care, (Gulamali et al., 2023) presented AEquity, comparing multiple model architectures to identify the training data set that would enhance fairness across the subpopulations without loss of predictive performance while mitigating and identifying potential biases in clinical data sets using AI assistance.

In the computer vision (CV) context, (Erukude et al., 2024; Erukude, 2024) described the development of latent model biases in their CNN classifiers when the model became overly reliant on irrelevant background features, creating multiple transformed datasets of the defined initially datasets to significantly reduced model uncertainty allowing them to center the analysis on features of interest.

As far as generating synthetic data takes things to a new level. For example, Lu et al. (Lu et al., 2023) describe how high-fidelity synthetic data sets can be generated utilizing GANs, VAEs, or diffusion models. This can mitigate bias for semantically reliant models based on human-labeled datasets.

It allows greater training when working within potentially privacy-sensitive environments like health care without needing the addition of more annotated datasets. For example, Luc et al. show examples of images that published synthetic images of radiology, which enabled pre-training diagnostic models without patient disclosures.

Through data performance (Mazumder et al., 2022) augmentations, the participants received ranking based on performing quality data tasks such as error correction and core data-set selections, which improve data quality and are designed to improve model quality.

Overall, these actions show that the systematic strengths of data curation and the complexity of models have comprehensible potential to provide robust, fair, and scalable AI systems.

V. CHALLENGES AND LIMITATIONS

Despite progress, many challenges exist that inhibit the scaled adoption of data-centric AI (DCAI). Annotation quality remains a persistent problem. Supervised learning depends on correctly labeled data, and manual annotation is expensive, slow, and often inconsistent, especially in domain-specific contexts like healthcare and legal (Zha et al., 2023; Zhou et al., 2024). Crowdsourcing significantly lowers costs but often lacks domain expertise. Programmatic, automatic, or programmatic labeling has potential but cannot supplant human-in-the-loop for many sensitive, nuanced tasks (Zha et al., 2023).

Bias and fairness aren't going away, either. Adding more data does not mean that you have dealt with systemic biases in the samples, to begin with (Gulamali et al., 2023; Erukude et al., 2024; Erukude, 2024). Properly remediating this requires evaluating subgroup imbalances and then applying techniques such as counterfactual data augmentation. We could benefit from better modeling and standardized frameworks to audit for bias across domains (Zhou et al., 2024; Mazumder et al., 2022).

The challenges with error detection and data deduplication continue. Naturally, tools like CleanLab aim to automate and significantly improve error detection; however, a little bit of labeling noise and errors, context-dependent errors, often still require human presence and determination (Zha et al., 2023; Zhou et al., 2024); and, if anything, will not this show that bad decisions on clean flagging could create systemic vulnerabilities if left unchecked.

The proliferation and fragmentation of data tooling also inhibit scalability. Many tools are used to solve problems dealing with an isolated step within a pipeline, profiling, cleaning, augmenting, but not integrated pipelines with end-to-end whole-system data improvements (Zhou et al., 2024). This fragmentation generates engineering overhead and decreases adoption momentum.

Finally, even once we can do all of this cleaning, we have no way of reliably estimating the impact on downstream data intervention; it is difficult to evaluate the effect of work on data quality. We can consider the standardized metrics of performance (e.g., accuracy and F1 score) by gluing together model optimizations, but to accurately consider data improvements requires defined evaluation settings with differing systems, or often an ablation experiment or longitudinal monitoring.

So, while we have seen some benchmarks emerge, like DataPerf (Mazumder et al., 2022), we still lack the ability to compare across modalities using standard or reliable methods.

To make DCAI less of a one-off isolated project or a failed art experiment in the data pipeline and to establish data as mature and truly transformative practice within AI development, we need to overcome the following barriers, library of costs for annotation in domain-specific contexts, bias remediation, error detection; data integration, the risks of being over-reliant on proper, yet automated data, and the poorly defined evaluations of data pain as a programmatic process.

Volume: 04 / Issue: 04 / 2025 - Open Access - Website: www.mijrd.com - ISSN: 2583-0406

VI. FUTURE DIRECTIONS

Data-centric AI (DCAI) will continue to mainstream, and research will seek to solve current constraints on scalability, automation, fairness, and measurement. One possibility is the advancement of fully automated data pipelines. Automation has been useful for previously isolated aspects of the pipeline, such as labeling and cleaning data. Future systems might adopt reinforcement learning or generative AI models to include human-in-the-loop automated data selection, labeling, and augmentation based on real-time human feedback (Zha et al., 2023). The growing presence of Auto-Data systems, parallel to the prevailing autoML, may help define the next wave of scalable AI.

The joint optimization of datasets and models, making datasets adaptable and model architectures, will change how and what we expect from future pipelines (Mazumder et al., 2022). Datasets have typically been conceptualized as fixed; however, a model that adopts a dataset and architecture together will perform better. For instance, the training of large language models such as GPT -4 on hand-curated, human-aligned datasets shows how model and dataset are evolving jointly.

Data will continue to be the primary element to measure fairness and the extent to which biases are denied. Public awareness of the problem has improved.

However, there are still significant gaps in standardized frameworks for reporting, auditing, correcting, and detecting bias at the dataset level across all datasets (Zhou et al., 2024; Gulamali et al., 2023).

In future pipelines, the integration of counterfactual data generation and synthetic balancing will likely emerge as everyday tools run during data generation to ensure equity.

The technical ubiquity of new data types will also expand. Unlike images or text, time series, and graph data must also be augmented and cleaned to reflect temporal constructs or relational data (Zha et al., 2023; Mazumder et al., 2022). Establishing modality-centric approaches for DCAI frameworks and how to organize and conduct data analysis is still a major research challenge.

There is still more work to be done when it comes to benchmarking. Current competitions, such as DataPerf (Mazumder et al., 2022), have developed and implemented many of the testing standards for evaluating data quality, but the discussion could be much broader than that.

Future benchmarks should also account for robustness, fairness, distributional shifts, and cost-effectiveness where rate limiting restricts data curation practices within preset annotation budgets.

In conclusion, DCAI has the potential to encompass multiple disciplines and spark transformational change. Once models' baseline capabilities are achieved, improvements in AI will be less about more complicated models and more about improving the quality of datasets.

Datasets, as a new type of tools, workflows, and frameworks, can potentially elevate the visualization of dataset engineering as a fundamental function within designing any intelligent system.

Volume: 04 / Issue: 04 / 2025 - Open Access - Website: www.mijrd.com - ISSN: 2583-0406

VI. CONCLUSION

Data-centric AI presents a unique paradigm shift for the AI discipline in how machine learning systems are constructed, developed, and deployed. The AI community has a historical obsession with algorithmic and model-centered innovation, often ignoring the systematic evolution of the quality and integrity of training data. Now, theoretical, and empirical studies are emerging that show the impact of data quality on model performance often has a larger influence than the complexity of architecture (Zha et al., 2023), (Mazumder et al., 2022).

As DCAI promotes the inclusion of data as a first-order engineering input rather than simply as an input, DCAI shifts the development of ML to a circular task of continuous dataset improvement. Improving the accuracy of labels, reducing bias, increasing diversity, revealing errors, and keeping data active have been shown to produce measurable improvements in fairness, robustness, and generalization without changing the model structure (Zha et al., 2023; Gulamali et al., 2023).

As witnessed in some high-stakes areas like healthcare, autonomous agents, and financial decisions, data-centricity has enhanced model metrics while mitigating ethical and operational risks (Gulamali et al., 2023; Erukude et al., 2024; Erukude, 2024).

While DCAI holds promise for a better machine learning future than algorithmic-centric ideas, significant challenges must be considered. Issues such as the cost of accurate annotation, the difficulty of detecting nuanced biases and mistakes, the limitations of existing tooling being fragmented, and the danger of too much automation create significant hurdles for widespread adoption (Zhou et al., 2024; Mazumder et al., 2022).

Advances across research, engineering practices, and the regulations that any DCAI could bring will need to be coordinated. A more substantial investment in modality-specific methods of data engineering, fairness-aware data workflows, and unified pathways for benchmarking is required to see the DCAI paradigm mature.

AI advancement will become increasingly synonymous with data-centric approaches in the short- and long-term. With models approaching their theoretical limits in capacity, the next oversized steps in machine learning and AI will eventually arise not because there were many-layered networks but because they were made using better, cleaner, and smarter datasets.

During the next decade, the composite evolution of Auto-data systems, dataset-model co-design, bias-resilient data workflows, and multi-modal DCAI methods will determine the course of DCAI and machine learning in general (Mazumder et al., 2022)

This paper aims to provide a foundational synthesis of the key pillars, methods, applications, challenges, and avenues for future work in Data-Centric AI. Adopting a data-centric perspective to build AI systems to be fair, robust, scalable, and ethically aligned is simply the necessary alternative, not an optional choice.

The shift to data-centricity is less of a slight adjustment in technical choices and more of a radical rethinking of how our computational intelligence relates to the data-learning process, a rethinking that will come to shape the future of AI innovation.

Volume: 04 / Issue: 04 / 2025 - Open Access - Website: www.mijrd.com - ISSN: 2583-0406

REFERENCES

[1] D. Zha, Z. P. Bhat, K.-H. Lai, F. Yang, Z. Jiang, S. Zhong, and X. Hu, "Data-centric Artificial Intelligence: A Survey," *arXiv preprint arXiv:2303.10158*, 2023. [Online]. Available: https://arxiv.org/abs/2303.10158

[2] A. Ng, "The Data-Centric AI Approach With Andrew Ng," *Scale AI Exchange*, [Online]. Available: https://exchange.scale.com/public/videos/the-data-centric-ai-approach-with-andrew-ng

[3] Y. Zhou, F. Tu, K. Sha, J. Ding, and H. Chen, "A Survey on Data Quality Dimensions and Tools for Machine Learning," *arXiv preprint arXiv:2406.19614*, 2024. [Online]. Available: https://arxiv.org/abs/2406.19614

[4] Y. Lu, D. Qian, X. Cao, J. Hu, X. Ma, and C. Zhou, "Machine Learning for Synthetic Data Generation: A Review," *arXiv preprint arXiv:2302.04062*, 2023. [Online]. Available: https://arxiv.org/abs/2302.04062

[5] M. Mazumder, A. Ramaswamy, M. Mitzenmacher, A. Rostamizadeh, P. Mineiro, and T. Hazan, "DataPerf: Benchmarks for Data-Centric AI Development," in *Proc. ICML Workshop on Data-Centric AI*, 2022. [Online]. Available: https://dataperf.org

[6] F. F. Gulamali, D. Ghosh, P. Riley, and J. Zou, "An AI-Guided Data Centric Strategy to Detect and Mitigate Biases in Healthcare Datasets," *medRxiv preprint*, 2023. [Online]. Available: https://doi.org/10.1101/2023.11.06.23298164

[7] S. T. Erukude, A. Joshi, and L. Shamir, "Identifying Bias in Deep Neural Networks Using Image Transforms," *Computers*, vol. 13, no. 12, p. 341, 2024. [Online]. Available: https://doi.org/10.3390/computers13120341

[8] Erukude, S. T. (2024). *Identifying Bias in CNN Image Classification Using Image Scrambling and Transforms* (Doctoral dissertation, Kansas State University).

[9] W. Ying, C. Wei, N. Gong, X. Wang, H. Bai, A. V. Malarkkan, S. Dong, D. Wang, D. Zhang, and Y. Fu, "A Survey on Data-Centric AI: Tabular Learning from Reinforcement Learning and Generative AI Perspective," *arXiv preprint arXiv:2502.08828*, 2025. [Online]. Available: https://arxiv.org/abs/2502.08828