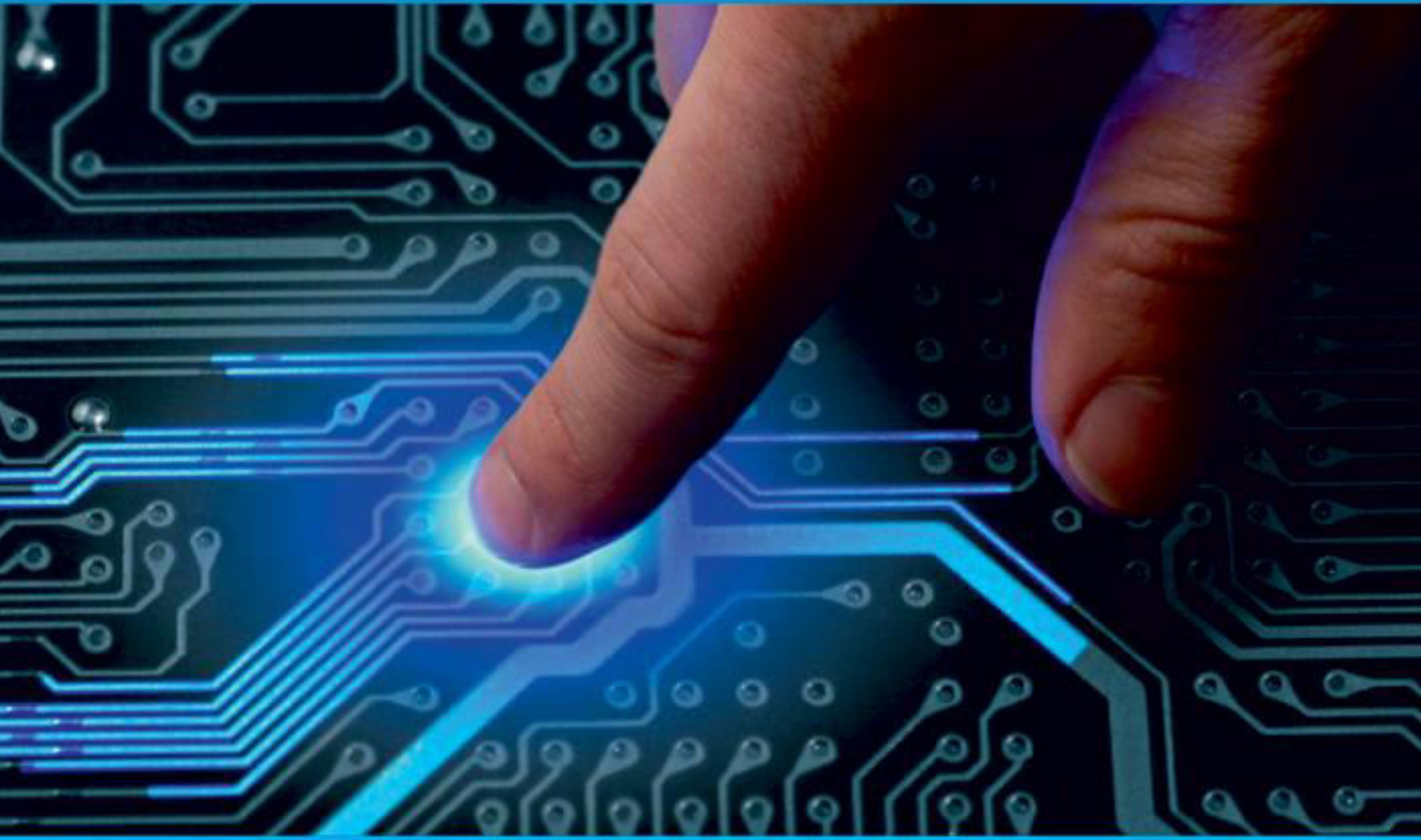
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 7, July 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

9940 572 462 6381 907 438 ijircce@gmail.com www.ijircce.com

| DOI: 10.15680/IJIRCCE.2023.1107004 |

Data-centric AI: Prioritizing Data Quality Over Model Complexity

Devendra Singh Parmar, Pankaj Gupta, Chetana Chouhan, Hemlatha Kaur Saran

Independent Researcher, USA

Independent Researcher, USA

Independent Researcher, India

Independent Researcher, India

ABSTRACT: In the past, Artificial Intelligence (AI) has been mostly driven by improvements in model architectures and computer power. New proof, on the other hand, shows that data, not models, is now a limit how well many AI systems work. This paper looks at how AI is moving away from a model-centric paradigm and towards a data-centric paradigm instead of focussing on making models that are more and more complicated. This study looks at industry reports, new literature, and benchmark case studies to show how important clean, consistent, and representative data is for better model performance, fairness, scalability, and moral integrity. Some of the most important points are how to make algorithms less biassed, how to make computers run faster, and new tools in industry that support data-centric methods. In its last part, the study stresses again how important data is as an equal partner in AI research and development. It also calls for more collaboration between different fields and new ways of doing data engineering.

KEYWORDS: Data-centric AI, Data quality, Machine learning, Model performance, Bias mitigation, Data efficiency, AI ethics, Benchmark datasets, Snorkel, Cleanlab.

I. INTRODUCTION

Deep neural networks and machine learning have advanced AI significantly in recent years. From computer vision's CNNs to NLP's transformer topologies, AI research and development has focused on constructing more complicated models. Massive datasets are used to train models, often ignoring data consistency, quality, and representativeness [1].

Any AI system relies on data and models. Data lets models learn, even though models outline the method. The classic model-centric paradigm suggests optimising hyperparameters, increasing layers, or refining training methods to boost performance measures. In benchmark and research challenges, this method has won several times [2]. However, it often assumes data is static and learning-friendly. Instead of examining and enhancing the dataset, the prevailing wisdom is to build bigger and more complicated models when model performance hits a wall or doesn't transfer well to real-world circumstances [3]. Recent studies and business initiatives have shown the limitations of a model-centric orientation. Even with advanced models, label noise, bias, incompleteness, and imbalance can negatively impact system performance, fairness, and interpretability. The data-centric AI paradigm was created to address this issue by emphasising data quality over model complexity [4]. Repeatedly adjusting the dataset while keeping the model essentially constant enhances learning. This improvement is especially important in sectors were obtaining high-quality, annotated data takes time or money. Many industries will benefit or suffer from data-centric AI [5]. Better-labeled and representative datasets can improve healthcare diagnostics and reduce algorithmic bias. Better sensor data annotation improves autonomous driving object detection and decision-making [6]. Strong fraud detection and risk assessment models may be constructed with consistent and accurate data in finance. Data quality assurance is becoming more critical for OpenAI, Google, and Tesla. This paradigm is becoming more relevant due to human-in-the-loop labelling, automated data validation, and dataset versioning [7].

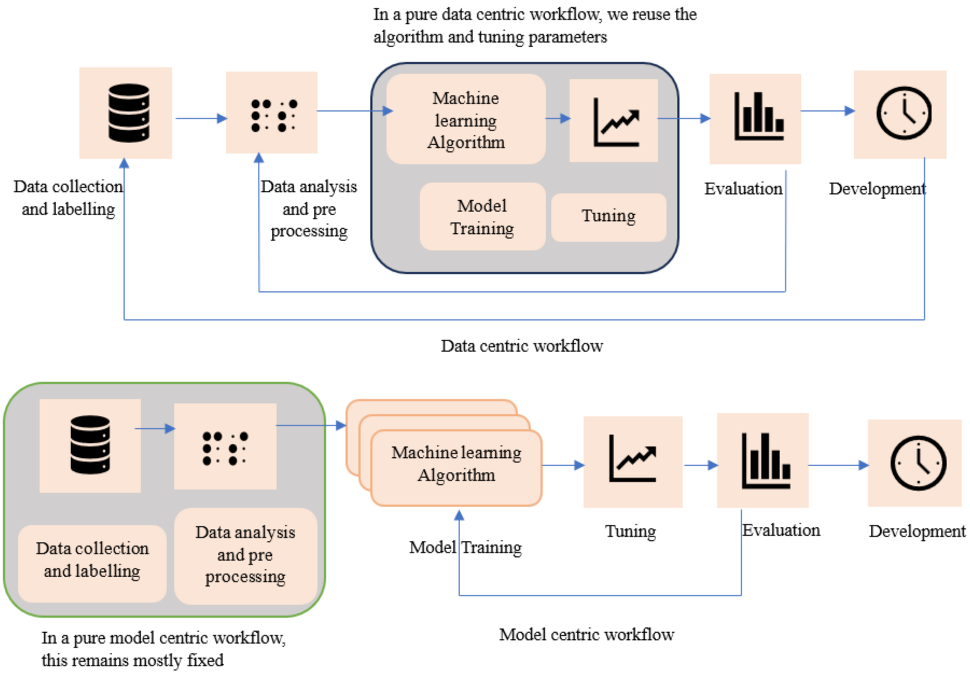

In a pure data centric workflow, we reuse the algorithm and tuning parameters
Data collection and labelling
Data analysis and pre processing
Machine learning Algorithm
Model Training
Tuning
Evaluation
Development
Data centric workflow
Data collection and labelling
Data analysis and pre processing
Machine learning Algorithm
Model Training
Tuning
Evaluation
Development
In a pure model centric workflow, this remains mostly fixed
Model centric workflow

Figure 1 Data-Centric AIPrioritizing Data Quality for Enhanced Performance [source: self-created]

Objective of the Paper

This paper aims to explore the concept of data-centric AI by analyzing the existing literature, comparing it with the model-centric approach, and highlighting the importance of high-quality data in achieving reliable AI outcomes. Through a secondary research methodology, this study synthesizes insights from academic and industry sources to underscore how improvements in data quality can drive performance gains, reduce bias, and promote ethical AI development. The structure of the paper includes a comprehensive literature review, discussion of thematic findings, and implications for future research and industry adoption.

II. LITERATURE SURVEY

Foundational Work in Model-Centric and Data-Centric AI

AI has always advanced by creating new models and algorithms, not by improving data. Core deep learning methods like [8] demonstrated neural network efficacy on ImageNet. Modern systems like ResNet, GPT, BERT, and EfficientNet use deeper layers, better activation functions, and larger parameter spaces to outperform simpler ones. Despite the large amount of training data, researchers often assumed it was enough for learning tasks and focused on model designs and training methods. Modern data-centric AI is a response to model complexity plateauing in many practical applications. Better data curation trumps model building and this strategy admits that data inconsistencies, label errors, and imbalanced representations can hinder performance, regardless of model sophistication The requirement for trustworthy and unbiased AI systems has driven data-centric viewpoints, in contrast to model-centric research that drove the deep learning boom.

Advocacy for Data-Centric AI: The Role of Andrew Ng

Dr. Andrew Ng is a data-centric leader and he influenced this change with his talks and writings. In his 2021 AI Conference keynote speech, Ng noted that most organisations have powerful models but lack high-quality, reliably tagged, and representative data [9]. He claims that data labelling, noisy annotation fixes, and enriching under-represented classes can improve performance more than weeks of hyperparameter optimisation or model switching. Ng's Data-Centric AI Competition proved that teams might improve model performance given selected data and a fixed model.

Dimensions of Data Quality: Consistency, Completeness, and Accuracy

AI systems use correctness, completeness, and consistency measures to evaluate data quality. Data consistency ensures common tagging and formats, eliminating model training interpretation. Missing features or labelling might cause underfitting and confusion, so make sure all relevant information is provided. Accuracy, or label or annotation correctness, is crucial to supervised learning. [10] found that filthy or uncurated data causes more machine learning "technical debt" than model errors. [11] used Confident Learning to identify and correct mislabeled data and showed that label noise alone can cause significant performance loss.

Empirical Evidence: Better Data Beats Complex Models

Instead of complicating models, empirical studies have demonstrated that better data is more helpful. The ImageNet Re-labeling Project is notable for manually re-labeling part of the validation set. The re-labeled version outperformed underperforming models, suggesting label problems were obscuring the models' true capabilities. Another example shows that enhancing COCO (Common Objects in Context) annotations and segmentations can improve item recognition and segmentation performance even with older models [12]. These results suggest that upgrading the dataset can yield better results than improving the model. Programmatic data labelling and moderate supervision from Stanford University's Snorkel system add support. Snorkel-based pipelines outperform hand labelling when equipped with dependable labelling functions and quality monitoring. This suggests that structured and scalable data management may fully exploit even complex models.

Benchmarks and Case Studies Supporting the Data-Centric Approach

Model innovation has traditionally been driven by benchmark datasets, but new case studies show that benchmark quality is just as important. To test models on real-world distribution shifts and cross-group fairness, [13] created the Wilds benchmark. Despite their accuracy, models trained on traditional datasets like CIFAR-10 or MNIST cannot generalise to real-world situations because to their lack of variety or bias. Data versioning and auditability processes in Weights & Biases and Data Version Control emphasise the requirement to monitor data change throughout model development. A key component of data-centric AI workflows, these tools let teams evaluate data changes without model changes.

III. METHODOLOGY

To investigate the rising profile of data-centric AI, this study takes a secondary data-based method, drawing from a variety of sources such as technical reports, academic white papers, and peer-reviewed academic articles. The main goal was to collect reliable, high-quality data that represents both theoretical advancements and real-world applications in the field. Relevance to current AI trends was prioritised when sources were selected according to clear criteria, with a focus on peer-reviewed journals and conferences from IEEE, ACM, Springer, and arXiv, as well as reputable industry insights from OpenAI, Google AI, and DeepMind. Previous publications were also considered. Research comparing model-centric and data-centric approaches, as well as articles discussing data management techniques and data quality, were given extra consideration. In order to find and filter literature, we employed tools like Google Scholar, IEEE Xplore, ScienceDirect, and Semantic Scholar. Through the use of a qualitative thematic synthesis approach, we combed through the gathered literature in search of commonalities and differences in data-centric AI research, as well as new developments and difficulties in the field. This method focus on how a change in emphasis from model complexity to data integrity is shaping present-day research and industry practices, and it allowed for the extraction of crucial information about the practical influence of high-quality data on the performance of AI models.

IV. ANALYSIS AND DISCUSSION

The trend towards data-centric AI shows that AI experts agree that data quality often trumps model complexity. This thematic analysis examines four main themes from the secondary literature: (1) data quality's effect on model accuracy, (2) data's role in promoting equity and reducing bias, (3) high-quality data's efficiency and scalability benefits, and (4) industry and academia's growing interest in data-centric AI frameworks and tools.

Theme 1: Impact of Data Quality on Model Accuracy

One of the most surprising findings is that data quality affects AI model precision and dependability. Model-centric AI has traditionally focused on improving architectures (e.g., by deepening neural networks, increasing connectivity, or implementing better optimisation functions), but a large body of research shows that cleaning, standardising, and labelling data often yields better results than adjusting model parameters. In natural language processing (NLP), huge language models like GPT-3 and GPT-4 were developed using massive volumes of data and carefully vetted corpora [14]. OpenAI researchers found that removing harmful, low-information, or noisy input increased the model's ability to understand complicated queries and respond logically. Even advanced models trained on low-quality or hostile text input produce hallucinations, inappropriate content, and unreliable outcomes. Data curation is a priority in computer vision due to ImageNet and COCO. Cleaning ImageNet of mislabeled or ambiguous photos increases ResNet and VGG's top-1 and top-5 accuracy without changing the model architecture. A study showed that a ResNet-50 model trained on a curated version of ImageNet performed identically to a more complicated architecture trained on the noisy dataset, demonstrating how data quality can reduce model complexity. Medical imaging, speech recognition, and satellite image processing have seen similar trends. In these high-stakes settings, inadequate or inadequately annotated data might lead to dangerous misclassifications or missed abnormalities. Even with smaller footprints, models trained on clean, context-rich datasets are more robust and generalisable.

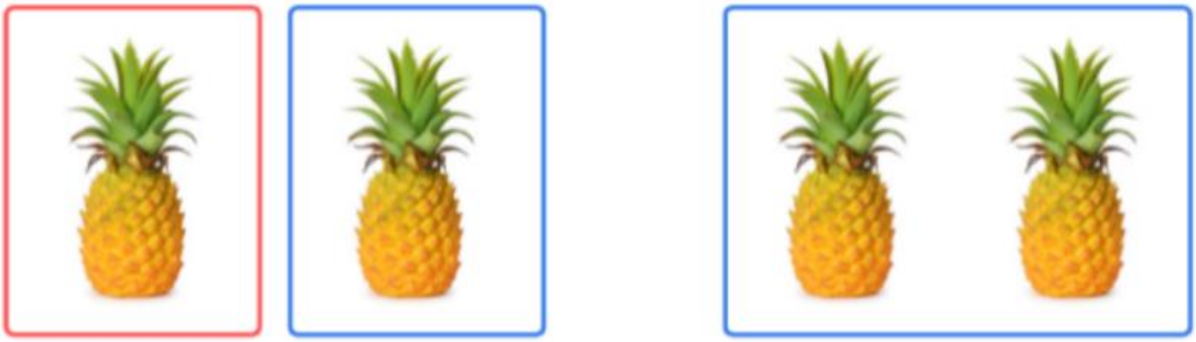


Figure 2 Different approaches to drawing bounding boxes[15]

Theme 2: Fairness, Bias, and Ethical AI through Better Data

Data is crucial to establishing if AI systems are fair and ethical, according to the second study review subject. AI bias has become a major issue due to real-world examples of systems demonstrating racial, gender, or cultural prejudices, sometimes due to biassed training data.

According to seminal studies by Joy Buolamwini and Timnit Gebru, commercial facial recognition algorithms overclassify persons of colour, especially women, relative to lighter-skinned people. Lack of diversity in training datasets caused the problem, not the model's architecture. These findings demonstrate the necessity for demographic, language, and contextual training data diversity [16]. Andrew Ng, a vocal supporter of data-centric AI, emphasises rebalancing datasets to improve minority user group results without new models. Examples include adding under-represented languages to speech recognition systems. Recent research has examined data-level bias mitigation measures such reweighting samples, de-biasing labels, and demographically balancing training data. Fairlearn and AI Fairness 360 audit datasets to help developers improve fairness measures. These tools recommend starting fairness actions during data collection and preparation before model training.

Responsible AI has made data sourcing ethics—using consent-based data, honouring cultural sensitivity, and deleting harmful content—crucial. Using clean, ethical data reduces performance bias and protects companies from legal and reputational issues.

Theme 3: Efficiency and Scalability with Good Data

Along with accuracy and ethics, high-quality data is essential to AI system efficiency and scalability. Data curation reduces training computational effort, speeds convergence, and improves model resource efficiency and generalisability. Due to the rising cost of training large-scale models, this is crucial. Research shows that data deduplication, error correction, and normalisation can dramatically shorten training epochs needed to reach target performance. With a deduplicated JFT dataset, Google Research found that training a vision transformer (ViT) to the same performance level took 30% fewer iterations, saving millions of GPU hours.

Industry examples reinforce these findings. Tesla's iterative strategy for refining autonomous driving models includes model tweaking, edge-case driving scenarios, accurate annotation, and feedback into the training pipeline [17]. Meta (Facebook) has invested in data platforms that prioritise high-quality, high-resolution pictures and curated interaction logs to improve content recommendation algorithms without expanding models. Spotify uses highly filtered user engagement data without bot activity or noise to better model retraining and user personalisation. From brute-force modelling to strategic data engineering, companies can scale AI across applications while reducing computational and environmental costs.

Even with limited infrastructure, smaller companies can compete by emphasising data collection and preparation above model experimentation. Data-centric AI makes AI development easier for everyone, which is great.

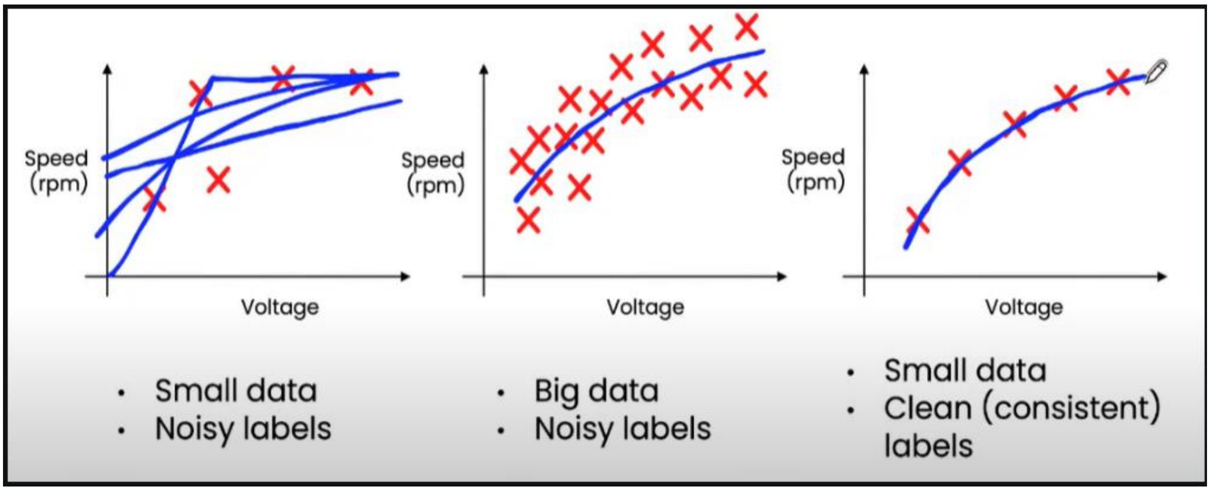


Figure 3Importance of consistency in small datasets [15]

Theme 4: Industry and Research Trends Toward Data-centricity

Finally, the data-centric AI ecosystem includes standards, community activities, and tools. As the field develops, open-source platforms and tools let practitioners evaluate and improve their datasets before modelling. Researchers at Stanford University developed Snorkel for programmatic tagging with no control. It takes less time and money to produce tagged datasets while maintaining precision. Cleanlab, an open-source program, discovers mislabeled data points and suggests modifications to improve datasets and downstream models.

Other solutions like Great Expectations and Label Studio provide data validation, label tracking, and quality assurance. These technologies include testing suites and version control, indicating a move towards treating datasets like models [18]. Data has become an emphasis in benchmarking. MLCommons' DataPerf project evaluates dataset quality and utility. Understanding how well the data allows generalisation and fairness is more important than evaluating model performance (accuracy, F1-score). Workshops and tracks at academic and commercial conferences like ICML and NeurIPS emphasise data quality, dataset documentation, and ethical data sourcing. Experts increasingly view data-centric AI as a fundamental part of machine learning rather than a speciality. This model is also used by startups and large cloud providers. Labelbox, Scale AI, and AWS SageMaker Ground Truth manage labelling, quality, and dataset versions. Traditional model-centric processes ignored these techniques.

V. LIMITATIONS

Though limited, this research provides a comprehensive overview of the new data-centric AI paradigm. Secondary sources such industry reports, white papers, and peer-reviewed studies inform the research. The results are limited by the breadth, depth, and accessibility of the current literature, which may not reflect all recent advances or unknown discoveries in the area. The research also addresses AI in general, although it may not have enough focus on data-centric techniques in healthcare, finance, and law. The research stresses supervised learning tasks like object detection and classification. Data quality is crucial in reinforcement learning and unsupervised learning, however these areas may not get enough attention in the present conversation. More research is needed to apply data-centric approaches to more machine learning paradigms.

VI. CONCLUSION

This research emphasises data-centric AI, shifting the focus from model complexity to data quality, consistency, and ethics. High-quality data enhances model accuracy, equity, bias reduction, efficiency, scalability, and following new research and market trends, as shown by theme analysis. Data is now considered as AI systems' main asset rather than the only driver of performance. Data versioning, automatic labelling, auditing, and validation tools must be strengthened for AI development to be scalable and reliable. Future research should examine the role of interdisciplinary collaboration, particularly between data engineers and machine learning specialists, in creating varied, technically sound, contextually relevant, and ethical datasets. By focussing on data, AI researchers can improve systems that mirror society's values.

REFERENCES

[1] N. Polyzotis and M. Zaharia, "What can data-centric AI learn from data and ML engineering?," *arXiv preprint arXiv:2112.06439*, 2021.

[2] M. Hajij, G. Zamzmi, K. N. Ramamurthy, and A. G. Saenz, "Data-centric AI requires rethinking data notion," *arXiv preprint arXiv:2110.02491*, 2021.

[3] L. J. Miranda, "Towards data-centric machine learning: a short review," *ljvmiranda921.github.io*, 2021.

[4] T. Parmar, "Data-centric Approach to Decision Making in Semiconductor Manufacturing: Best Practices and Future Directions," 2021.

[5] H. Kurban, P. Sharma, and M. Dalkilic, "Data expressiveness and its use in data-centric AI," in *Proc. NeurIPS Data-Centric AI Workshop*, Dec. 2021.

[6] H. Amrani, "Model-centric and data-centric AI for personalization in human activity recognition," Ph.D. dissertation, Univ. of Milano-Bicocca, 2021.

[7] D. Alvarez-Coello, D. Wilms, A. Bekan, and J. M. Gómez, "Towards a data-centric architecture in the automotive industry," *Procedia Computer Science*, vol. 181, pp. 658–663, 2021.

[8] V. Theodorou, I. Gerostathopoulos, I. Alshabani, A. Abelló, and D. Breitgand, "MEDAL: An AI-driven data fabric concept for elastic cloud-to-edge intelligence," in *Proc. Int. Conf. Adv. Inf. Netw. Appl.*, Cham: Springer, 2021, pp. 561–571.

[9] V. S. A. Poolla and B. K. Mandava, "Understanding the Challenges and Needs of Requirements Engineering for Data-centric Systems," 2021.

[10] X. Liu, Y. Zhou, and A. Rau, "Smart card data-centric replication of the multi-modal public transport system in Singapore," *J. Transp. Geogr.*, vol. 76, pp. 254–264, 2019.

[11] V. Virk, "Leading in a Data Centric Society," *J. Intell., Confl., Warfare*, vol. 3, no. 3, pp. 55–58, 2021.

[12] P. Chatarasi et al., "Marvel: A data-centric approach for mapping deep learning operators on spatial accelerators," *ACM Trans. Archit. Code Optim. (TACO)*, vol. 19, no. 1, pp. 1–26, 2021.

[13] G. Kadam, E. Smirni, and A. Jog, "Data-centric reliability management in GPUs," in *Proc. IEEE/IFIP Int. Conf. Dependable Syst. Netw. (DSN)*, 2021, pp. 271–283.

[14] J. Zigon, "Can machines be ethical? On the necessity of relational ethics and empathic attunement for data-centric technologies," *Social Research: An International Quarterly*, vol. 86, no. 4, pp. 1001–1022, 2019

[15] https://neptune.ai/blog/data-centric-vs-model-centric-machine-learning

[16] H. Subramonyam, C. Seifert, and M. E. Adar, "How can human-centered design shape data-centric AI?," in *Proc. NeurIPS Data-Centric AI Workshop*, Dec. 2021.

[17] T. Chuprina, D. Mendez, and K. Wnuk, "Towards artefact-based requirements engineering for data-centric systems," *arXiv preprint arXiv:2103.05233*, 2021.

[18] C. J. Bartel, "Data-centric approach to improve machine learning models for inorganic materials," *Patterns*, vol. 2, no. 11, 2021.