Interdisciplinary Systems for Global Management
Vol. 1 (2025), No. 2, pp. 76-85
DOI: https://doi.org/10.55578/isgm.2509.006

ARTICLE

Data Saturation Reliability Framework: A Framework for Optimising AI Input Feeds

Michael Mncedisi Willie [1,*], Siyabonga Jikwana [2,3], Lesiba Arnold Malotana [2], Zwanaka James Mudara [4]

[1] *Council for Medical Schemes, Pretoria, South Africa*

[2] *Gauteng Department of Health, Johannesburg, South Africa*

[3] *University of Pretoria, Pretoria, South Africa*

[4] *Vaal University of Technology, Vereeniging, South Africa*

[*]Corresponding author. Email: m.willie@medicalschemes.co.za

Received: 15 August 2025, Accepted: 29 August 2025, Published: 02 September 2025

Abstract

Artificial Intelligence (AI) systems increasingly rely on large and diverse data streams to support accurate, adaptive, and context-aware decision-making. However, beyond a certain point, adding new data can lead to diminishing or even negative returns due to redundancy, noise, and bias, a phenomenon known as data saturation. This paper introduces the Data Saturation Reliability (DSR) framework, a conceptual framework that optimises AI input feeds by balancing data volume, quality, and reliability. Drawing on principles from information theory, machine learning, and data governance, the DSR framework formalises saturation thresholds, signal-to-noise ratio assessment, temporal relevance, and dynamic feedback mechanisms as key factors for sustainable AI performance. By linking marginal information gain to input reliability, the DSR framework provides strategies to mitigate risks of over-saturation, bias propagation, and operational inefficiencies, while improving predictive accuracy and adaptive learning. The framework prioritises quality over quantity, encouraging intelligent curation of inputs rather than indiscriminate data collection. Applications include high-stakes fields such as healthcare diagnostics, financial forecasting, autonomous systems, and large-scale natural language processing, where real-time decision accuracy and reliability are vital. The paper highlights opportunities for empirical validation, cross-domain adaptation, and integration of DSR principles into AI lifecycle management and governance. Ultimately, the framework promotes shifting from "more data equals better performance" towards an optimal data balance that ensures operational effectiveness and ethical responsibility in AI deployment.

Keywords: Artificial Intelligence; Data Saturation; Reliability; Input Feeds; Signal-to-Noise Ratio; Feedback Mechanisms; AI Governance; Data Quality; Machine Learning; Optimisation

1. INTRODUCTION

As modern AI systems ingest ever-larger, faster, and more heterogeneous input streams (e.g., sensors, logs, crowdsourced reports), the field continues to operate under the long-standing assumption that "more data is better." However, empirical and operational evidence increasingly show that additional data can yield diminishing or even negative returns beyond a certain point unless their reliability is actively managed. We describe this inflexion as data saturation and argue that its interplay with input reliability remains under-theorised, particularly in multi-source, real-time pipelines. Historically, early research highlighted the sheer power of scale [1], while more recent scaling-law studies demonstrate that performance gains follow power-law dynamics that hinge on balancing parameters, compute, and data volume [2,3]. These developments call for a shift from indiscriminate data accumulation toward principled curation and reliability assessment of input streams.

Evidence increasingly indicates that quality rather than quantity governs the effectiveness of additional data at scale. Compute-optimal analyses reveal that misaligned dataset sizes relative to compute budgets waste resources and depress performance [2,3]. Redundant or noisy data accelerate

https://doi.org/10.55578/isgm.2509.006

saturation, with studies showing that mislabelled or contaminated datasets compromise generalisation and predictive accuracy across modalities [4,5]. While data-quality models address these issues through denoising, relabelling, or error detection [6,7], they often treat reliability as an isolated corrective step rather than as a dynamic factor shaping saturation itself.

Operational challenges exacerbate this problem: technical debt [8], cascading data errors [9], and biases amplified by naïve source aggregation make saturation-driven failures more likely in practice. While calibration techniques such as weighted aggregation and bias adjustments [10,11] improve reliability, they remain piecemeal solutions rather than part of an integrated theoretical framework.

This paper introduces the Data Saturation Reliability (DSR) framework, which extends beyond existing saturation and data-quality models by linking marginal information gain directly to source reliability and pipeline health. Unlike traditional saturation frameworks, which define a plateau in performance, DSR formalises saturation as the point where the expected utility of new inputs measured via uncertainty reduction, error elasticity, or entropy gain falls below a reliability-adjusted threshold. In doing so, DSR reframes the question from "how much data is enough" to "how much reliable information is being added per unit cost." We propose design levers, including data-centric quality improvement [12], active learning for high-value samples [13], and truth-discovery source calibration that can delay or mitigate saturation. Conceptually, DSR offers a unified perspective and testable predictions for multi-source, real-time AI systems that must reconcile speed, scale, and trust.

2. LITERATURE REVIEW

Data saturation is commonly defined as the point at which adding new data no longer yields proportional informational value or performance gains in AI models. In qualitative research, Saunders et al. [14] emphasise the varied interpretations of saturation and stress the need for consistent operationalisation tied to research design and analytic framework. Applied to AI, saturation highlights that after a threshold, additional input data can not only cease to improve model performance but may also degrade it through redundancy, noise, or conflict. Thus, AI systems benefit most not from raw volume, but from high-quality, contextually diverse, and reliable data. In multi-source settings, calibration strategies such as weighting or probabilistic truth discovery are essential to preserve reliability, reduce bias, and optimise outcomes, paralleling hybrid human–AI systems where algorithmic efficiency still requires human judgement [15].

While Braun and Clarke [16] question whether saturation is always a sufficient justification in qualitative research, and Guest et al. [17] treat it as a marker of adequacy, these interpretations must be adapted to AI's unique conditions of continuous learning, high-dimensional features, and shifting environments. Effective performance depends less on dataset size and more on data quality, governance, and diversity [18]. Algorithms such as InfoGrowth [19] demonstrate how targeted data cleaning and selection can improve performance in dynamic streams, illustrating the growing recognition that quality and timeliness drive utility more than scale alone.

Reliability further complicates this picture. Studies show that data volume, provenance, validity, and contextual relevance all shape outcomes [20]. This holds across sectors from consumer analytics to education [21] and is especially critical in healthcare, where heterogeneous datasets and stakeholder complexity demand careful management [22,23]. Stream management techniques [24] and filtering approaches [25] help mitigate redundancy and noise. However, cascading reliability issues persist, particularly in high-stakes contexts like healthcare and finance [26].

Against this backdrop, the DSR framework makes a distinct contribution by integrating the concepts of saturation and reliability into a unified framework. DSR enables systematic monitoring of input reliability, alignment with dynamic environments, and integration into governance frameworks. This approach advances model robustness and efficiency and strengthens transparency, accountability, and ethical integrity in AI-driven decision-making.

3. THEORETICAL FRAMEWORK

3.1. Foundations of Data Saturation Reliability Framework

The DSR framework emerges at the nexus of information theory, signal-to-noise ratio principles, and the AI/machine learning lifecycle, positing that system performance is contingent on the optimal interplay between the volume and quality of input data streams [27]. While larger datasets can initially enhance model learning and generalisation, a saturation threshold exists beyond which additional data may offer diminishing, or even detrimental, returns on predictive accuracy. This principle is exemplified in the work of Kedziora and Marciniak [28], who integrated fuzzy logic with robotic process automation to improve data validation in financial systems, achieving a 67% reduction in false rejection rates while maintaining 97% accuracy, highlighting the role of intelligent data filtering in balancing data volume and reliability. Similarly, Budnikov, Bykova, and Yamshchikov [29] demonstrate that large language models (LLMs) achieve their generalisation capacity through underlying information-theoretic mechanisms, which facilitate robust performance across diverse natural language processing tasks. These findings are reinforced by Ajiboye, Arshah, and Qin [30,30], who show that predictive models trained on larger, representative datasets generally achieve superior accuracy and stability, underscoring the necessity of both dataset sufficiency and quality. These studies highlight a delicate balance in AI systems: while abundant data is essential for improving and refining models, too much data can hinder efficiency and reduce predictive accuracy, echoing the economic principle of diminishing returns as it applies to AI data ingestion and model performance.

3.2. Key Constructs and Relationships

The DSR framework is structured around four interrelated constructs:

1) Data Volume – The quantity of data required to train, validate, and optimise machine learning and deep learning models, where sufficient data is essential for identifying patterns, improving predictive accuracy, and ensuring generalisation across diverse real-world contexts [18].

2) Data Quality – The degree to which data meets the requirements for its intended use, encompassing characteristics such as accuracy, completeness, consistency, reliability, and relevance within a computer science context [31,32].

3) Saturation Threshold – The inflexion points where further data acquisition fails to provide proportional improvements and may introduce noise [33].

4) Reliability Index: A composite measure that evaluates the trustworthiness of AI outputs under prevailing data stream conditions. It reflects a system's consistent performance while accounting for security, trust, resilience, and agility and is operationalised through sub-metrics and assessment tools such as risk and vulnerability analyses [34,35].

DSR framework proposes that Data Quality moderates the relationship between Data Volume and AI Performance, while the Saturation Threshold acts as a critical boundary condition. The Reliability Index is the operational measure for determining optimal data input strategies.

3.3. Dynamic Feedback Loops

Unlike static data quality frameworks, the DSR framework incorporates dynamic feedback loops that continuously evaluate input streams against performance metrics. These loops enable the AI system to adaptively filter, prioritise, or discard incoming data based on evolving operational requirements. In real-time environments such as autonomous driving or algorithmic trading, this feedback mechanism ensures that the AI remains responsive to situational changes while avoiding performance degradation from over-saturation. AI-driven intelligent feedback loops transform raw user interactions into actionable insights, enabling continuous, real-time product innovation, proactive decision-making, and rapid iteration cycles aligned with customer needs [36].

3.4. Theoretical Contribution

The DSR framework advances AI theory by challenging the conventional assumption that "more data automatically yields better performance," advocating an optimal data equilibrium approach instead. Embedding performance-driven saturation monitoring into AI lifecycle management extends traditional data governance models, bridging technical optimisation with ethical and operational imperatives. This aligns with findings that integrating AI with Master Data Management (MDM) improves data quality, consistency, and governance efficiency across industries, while addressing challenges such as redundancy, silos, and ethical concerns [37]. Simultaneously, effective AI accountability necessitates well-defined structures of answerability, including authority recognition, interrogation, and limitation of power, supported by an architecture of seven features that collectively facilitate compliance, reporting, oversight, and enforcement [38]. Beyond these organisational contexts, the DSR framework demonstrates cross-domain applicability, from healthcare diagnostics to large-scale natural language processing, providing a structured methodology to balance continuous data inflows with decision reliability and system trustworthiness.

4. RESEARCH METHODS

4.1. Research Design

This study adopts a conceptual research design grounded in a theory-building approach, drawing from established design science principles and theoretical synthesis [39]. The aim is to conceptualise and formalise the DSR framework by integrating insights from information theory, machine learning performance analysis, and data governance literature. Rather than empirically testing the theory at this stage, the research focuses on developing a coherent framework that can later be operationalised in domain-specific applications such as autonomous systems, healthcare diagnostics, and financial forecasting.

4.2. Data Sources

The study is based on a systematic literature review of peer-reviewed articles, technical reports, and industry white papers published between 2000 and 2025, of which 12 out of 43 references (~28%) were published in the last three years (2023–2025), 20 (~47%) in the last five years (2021–2025), and 26 (~60%) in the past ten years (2016–2025), indicating that a majority of the references are relatively recent, with a substantial portion from the last five years and a smaller, but notable, fraction from the very recent three-year period. Databases consulted include Scopus, IEEE Xplore, ACM Digital Library, and Google Scholar. Keywords such as data saturation, AI input streams, machine learning performance decay, data governance, and signal-to-noise ratio in AI were used in Boolean combinations to ensure comprehensive coverage. Sources were screened for relevance, methodological quality, and conceptual depth, focusing on studies addressing the relationship between data volume, data quality, and AI performance. **Table 1** summarises the references by citation type, showing the predominance of journal articles alongside conference papers, preprints, books, technical reports, and other sources.

Table 1. Distribution of References by Citation Type

Citation Type	N
Journal Articles	27
Conference Papers / Proceedings	7
Preprints (arXiv / Research Square / ResearchGate)	5
Book Chapters / Books	4
Technical Reports	1
Miscellaneous / Other (Datasets, Surveys, Websites, etc.)	3

4.3. Analytical Approach

The literature was subjected to thematic content analysis [40], enabling the identification of recurring patterns, theoretical gaps, and conceptual overlaps. The study followed three stages:

• Data Reduction – Filtering and coding relevant literature according to predefined categories: data volume effects, quality measures, saturation thresholds, and reliability assessment.

• Pattern Recognition – Identifying correlations between constructs and mapping them to known theoretical models such as the Law of Diminishing Returns and Shannon's Information Theory [27,41-43].

• Framework Synthesis – Integrating findings into a structured conceptual model that defines the key constructs, their interrelationships, and the feedback mechanisms underpinning the DSR framework.

4.4. Validity and Reliability Considerations

While this is a conceptual study without direct empirical testing, construct validity was strengthened through triangulation of multiple scholarly sources and theoretical perspectives. Reliability was ensured by maintaining a consistent coding protocol and conducting repeated reviews of literature categorisation to minimise bias. Future empirical studies can build on this framework to test its predictive validity in specific AI domains, thereby extending the external validity of the DSR framework.

5. ANALYSIS AND RESULTS

5.1. Data Saturation and Volume

AI model performance does not increase linearly with data volume; additional data can produce diminishing or negative returns beyond a certain saturation threshold due to redundancy, noise, or conflicting information. This emphasises the importance of balancing quantity with contextual relevance, representativeness, and computational constraints [1-3,12].

5.2. Data Quality and Reliability

High-quality, accurate, and diverse datasets are more critical than sheer volume. Data errors, mislabelled inputs, or heterogeneous sources can degrade predictive accuracy, reduce generalisation, and introduce biases. Corrective methods such as advanced denoising, feature selection, and probabilistic truth discovery are essential for maintaining reliability [31,37].

5.3. Saturation Thresholds and Performance Limits

The DSR framework formalises a performance inflexion point, beyond which additional inputs no longer improve and may even degrade model outcomes. Operationalising this threshold through reliability indices and adaptive monitoring enables systematic oversight of AI input streams [33,34].

5.4. Dynamic Feedback Loops and Adaptive Learning

Intelligent feedback mechanisms allow AI systems to continuously filter, prioritise, or discard incoming data. These loops are critical for real-time environments, supporting proactive decision-making, continuous learning, and alignment with evolving operational requirements, including user-centric product innovation [36].

5.5. Accountability, Governance, and Cross-Domain Applicability

AI reliability and accountability require structured governance mechanisms, including clearly defined accountability, authority recognition, and monitoring processes. Integrating AI with master data management enhances industry governance, data consistency, and operational ethics. DSR principles apply broadly, from healthcare diagnostics to large-scale NLP systems, providing a structured methodology to balance data inflows with system trustworthiness [37,38].

Building on these findings, the following section introduces the DSR conceptual framework, which integrates the insights on data volume, quality, saturation thresholds, and adaptive learning into a structured model for systematic AI performance management.

6. CONCEPTUAL FRAMEWORK: DSR FRAMEWORK

The DSR framework provides a structured lens to understand how multi-source input data streams are transformed into reliable system outcomes through iterative evaluation and intervention, as depicted in **Figure 1**. At the entry point, heterogeneous data streams are subjected to reliability factors such as provenance, validity, and relevance. These factors collectively inform the data saturation threshold, ensuring that only sufficiently reliable data advances decision-making. System outcomes measured in performance, trust, and bias mitigation are end goals and serve as feedback loops, continuously refining reliability criteria and informing subsequent data collection cycles.

DSR interventions such as data quality enhancement, active learning, and truth discovery are central to this framework. They strengthen the reliability threshold and promote adaptive improvement. This cyclical process illustrates the dynamic relationship between data inputs, reliability mechanisms, and outcomes, ensuring conceptual clarity and methodological rigour.

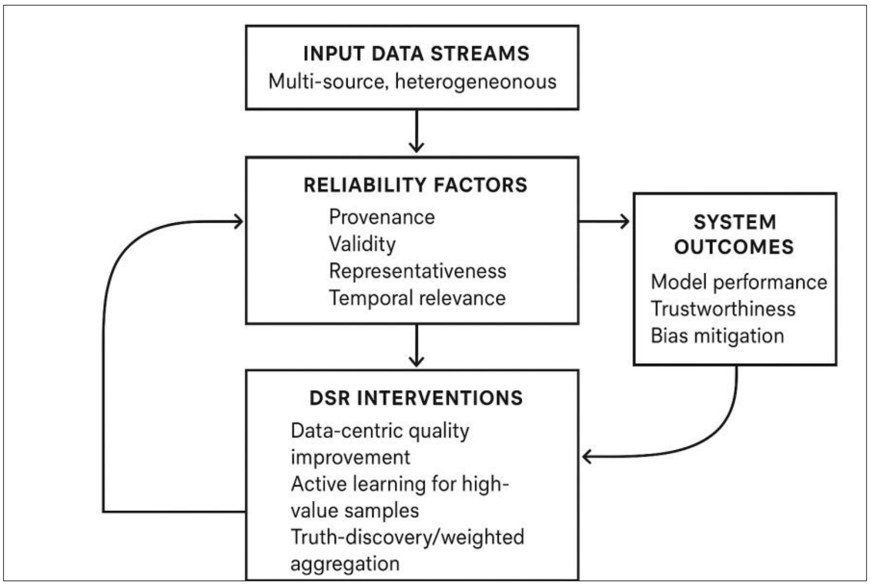

Figure 1. Conceptual framework of the DSR

7. DISCUSSION

This study aimed to develop and formalise the DSR framework, providing a conceptual framework for understanding the interplay between data volume, quality, and reliability in AI systems. Against the backdrop of increasingly heterogeneous, high-velocity input streams in modern AI applications, our findings challenge the prevailing assumption that "more data automatically yields better performance" [1]. Instead, the DSR perspective emphasises an optimal data equilibrium, maximising predictive gains while mitigating redundancy, noise, and bias. This approach aligns with broader trends in AI theory emphasising principled data curation, quality assurance, and reliability monitoring as critical complements to model architecture and compute resources [2,3].

This study emphasises that data quality is a key factor influencing AI performance. High-quality, contextually relevant, and diverse datasets emerged as more influential than raw data volume, reinforcing previous assertions regarding the necessity of representative inputs for generalisation and stability [31,37]. The study illustrates that unchecked data accumulation can exacerbate performance degradation in multi-source pipelines, not only through statistical maximising predictive gain, but also via operational risks such as technical debt, cascading errors, and bias amplification [9,11]. These findings extend the theoretical discourse on the law of diminishing returns to AI input streams, highlighting the need for a nuanced understanding of information utility relative to computational and operational constraints.

Identifying saturation thresholds offers a pragmatic tool for operationalising reliability within AI lifecycles. By formalising the inflexion point at which additional data ceases to contribute meaningfully to model performance, the DSR framework provides actionable insights for designing adaptive pipelines and feedback mechanisms. Dynamic feedback loops, as observed in intelligent user feedback systems [36], demonstrate the practical value of this approach, allowing AI systems to prioritise high-value data and discard or downweight inputs that fail to enhance predictive utility. Such mechanisms are particularly salient in real-time environments, where rapid decision-making must balance scale, speed, and trustworthiness [12,13].

From a governance perspective, our findings highlight the interconnectedness of accountability, reliability, and operational ethics. Effective AI systems require delineated structures of answerability and oversight, encompassing authority recognition, interrogation, and enforcement mechanisms [38]. Integration with master data management practices further strengthens these dimensions, improving data consistency, transparency, and compliance across diverse industrial contexts [37]. This dual focus on technical and organisational safeguards reinforces the notion that DSR is not purely a computational framework but a holistic lens encompassing the socio-technical ecosystem in which AI operates.

The DSR framework advances the theoretical and practical understanding of AI input management, offering a structured methodology for balancing data abundance with reliability, quality, and ethical accountability. It invites scholars and practitioners to reconsider prevailing assumptions regarding data volume, advocating for a shift from indiscriminate accumulation to principled, reliability-informed data strategies. These insights have immediate implications for AI system design, lifecycle governance, and cross-industry operational excellence, while providing a foundation for ongoing empirical validation and refinement.

8. IMPLICATIONS FOR MANAGERS AND POLICY MAKERS

The DSR framework provides a conceptual lens and actionable guidance for managers and policy makers tasked with deploying and governing AI systems in complex, high-stakes environments. The following recommendations translate the framework into practical steps.

8.1. Prioritise Reliability over Raw Volume

- Managers should establish data acquisition strategies that emphasise quality, contextual relevance, and representativeness over indiscriminate expansion. Routine audits of data pipelines, coupled with reliability indices, can prevent performance losses due to saturation.
- Policy makers should incentivise or regulate standards for dataset provenance, labelling accuracy, and diversity, ensuring reliability is integrated into compliance frameworks.

8.2. Operationalise Saturation Thresholds as Governance Tools

- Managers can embed saturation monitoring into AI lifecycle dashboards to identify when additional inputs cease to add value, thereby optimising compute budgets and preventing redundant storage costs.
- Policy makers can require organisations to report saturation thresholds as part of algorithmic accountability, similar to stress-testing in finance or safety thresholds in healthcare.

8.3. Institutionalise Dynamic Feedback Loops

- Managers should implement adaptive mechanisms that continuously filter, prioritise, or discard data streams in real time. This is particularly vital for healthcare, financial services, and autonomous systems where decision timeliness and trustworthiness are paramount.
- Policy makers can support the development of interoperability standards and real-time monitoring protocols to ensure consistent application of feedback loops across industries.

8.4. Link DSR with Organisational Governance and Ethics

- Managers should integrate DSR principles into MDM practices, ensuring alignment between technical optimisation and organisational ethics. Doing so will reduce risks of bias amplification, cascading errors, and technical debt.
- Policy makers should embed reliability-adjusted data use within broader AI governance frameworks, mandating explainability, oversight structures, and redress mechanisms when failures occur.

8.5. Encourage Cross-Sectoral Application and Collaboration

- Managers can leverage DSR-based metrics to benchmark AI system performance across sectors, fostering innovation in industries with overlapping reliability challenges (e.g., healthcare and finance).
- Policy makers should promote multi-stakeholder collaboration on standards for reliability-adjusted AI deployment, ensuring that best practices diffuse across domains while safeguarding public trust.

9. CONCLUSION

The DSR framework provides a novel perspective on AI input management, highlighting the dynamic interplay between data volume, quality, and contextual relevance. Identifying and managing saturation points enables AI practitioners to sustain high model performance, reduce inefficiencies, and mitigate bias and noise amplification risks. This study contributes a structured approach to understanding data-driven performance limits, guiding organisations in designing reliable, adaptive, and sustainable AI systems.

Acknowledgements and Use of AI: While preparing this manuscript, the authors employed Grammarly and QuillBot to support language refinement, editing, and plagiarism checking. All outputs generated by these tools were carefully reviewed and revised by the author, who assumes full responsibility for the accuracy and integrity of the final content presented in this publication.

Funding Statement: This research and manuscript preparation were conducted without external funding or financial support.

Author Contributions: MMW conceptualised and drafted the initial manuscript, with additional input and contribution from SJ, LAM, and ZJM. All authors reviewed and approved the final version of the manuscript.

Data Availability Statement: This study exclusively used data obtained from secondary sources through a comprehensive literature review. All referenced data are publicly accessible and have been appropriately cited within the manuscript.

Conflict of Interest: There are no conflict of interest or any type of competing interests for this research.

Consent for Publication: Not applicable.

Ethical Approval: This study did not involve human participants or animal subjects and thus did not require ethical approval.

REFERENCES

1. Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems, 24*(2), 8–12. https://doi.org/10.1109/MIS.2009.36
2. Kaplan, J., McCandlish, S., Henighan, T., Brown, T., et al. (2020). Scaling laws for neural language models. *arXiv*. https://arxiv.org/abs/2001.08361
3. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., et al. (2022). Training compute-optimal large language models. *arXiv*. https://arxiv.org/abs/2203.15556
4. Atkinson, G., & Metsis, V. (2021). A survey of methods for detection and correction of noisy labels in time series data. In A. F. Gelbukh (Ed.), *Artificial intelligence applications and innovations* (IFIP Advances in Information and Communication Technology, Vol. 625, pp. 447–458). Springer. https://doi.org/10.1007/978-3-030-79150-6_38
5. González-Santoyo, C., Renza, D., & Moya-Albor, E. (2025). Identifying and mitigating label noise in deep learning for image classification. *Technologies, 13*(4), 132. https://doi.org/10.3390/technologies13040132
6. Nigam, N., Dutta, T., & Gupta, H. P. (2020). Impact of noisy labels in learning techniques: A survey. In T. Dutta (Ed.), *Advances in data and information sciences* (pp. 447–458). Springer. https://doi.org/10.1007/978-981-15-0694-9_38
7. Northcutt, C. G., Jiang, L., & Chuang, I. L. (2021). Confident learning: Estimating uncertainty in dataset labels. *arXiv*. https://arxiv.org/abs/1911.00068
8. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Proceedings of the 29th*

International Conference on Neural Information Processing Systems (Vol. 2, pp. 2503–2511). Curran Associates, Inc.
9. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). Everyone wants to do the model work, not the data work: Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). https://doi.org/10.1145/3411764.3445518
10. Banerjee, A., & Compte, O. (2024). Consensus and disagreement: Information aggregation under (not so) naive learning. *Journal of Political Economy, 132*(4). https://doi.org/10.1086/729448
11. Brendel, P., Torres, A., & Arah, O. A. (2023). Simultaneous adjustment of uncontrolled confounding, selection bias and misclassification in multiple-bias modelling. *International Journal of Epidemiology, 52*(4), 1220–1230. https://doi.org/10.1093/ije/dyad001
12. Jarrahi, M. H., Memariani, A., & Guha, S. (2023). The principles of data-centric AI. *Communications of the ACM, 66*(8), 84–92. https://doi.org/10.1145/3571724
13. Settles, B. (2009). Active learning literature survey (Technical Report TR1648). University of Wisconsin–Madison, Department of Computer Sciences. http://digital.library.wisc.edu/1793/60660
14. Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., Burroughs, H., & Jinks, C. (2017). Saturation in qualitative research: Exploring its conceptualization and operationalization. *Quality & Quantity, 52*(4), 1893–1907. https://doi.org/10.1007/s11135-017-0574-8
15. Speed, C., & Metwally, A. A. (2025). The Human–AI Hybrid Delphi Model: A structured framework for context-rich, expert consensus in complex domains. *arXiv*. https://arxiv.org/abs/2508.09349
16. Braun, V., & Clarke, V. (2019). To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales. *Qualitative Research in Sport, Exercise and Health, 13*(1), 1–16. https://doi.org/10.1080/2159676X.2019.1704846
17. Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? *Field Methods, 18*(1), 59–82. https://doi.org/10.1177/1525822X05279903
18. Aldoseri, A., Al-Khalifa, K. N., & Hamouda, A. M. (2023). Re-thinking data strategy and integration for artificial intelligence: Concepts, opportunities, and challenges. *Applied Sciences, 13*(12), 7082. https://doi.org/10.3390/app13127082
19. Qin, Z., Zhaopan, X., Zhou, Y., & You, Y. (2024). Dataset growth. *arXiv*. https://doi.org/10.48550/arXiv.2405.18347
20. Okeleke, P. A., Ajiga, D., Folorunsho, S., & Ezeigweneme, C. (2024). Predictive analytics for market trends using AI: A study in consumer behavior. *International Journal of Engineering Research Updates, 7*(1). https://doi.org/10.53430/ijeru.2024.7.1.0032
21. Ali, O., Murray, P. A., Momin, M., Dwivedi, Y. K., & Malik, T. (2024). The effects of artificial intelligence applications in educational settings: Challenges and strategies. *Technological Forecasting and Social Change, 199*, 123076. https://doi.org/10.1016/j.techfore.2023.123076
22. National Academy of Medicine. (2023). Artificial intelligence in health care: The hope, the hype, the promise, the peril (D. Whicher, M. Ahmed, S. T. Israni, & M. Matheny, Eds.). National Academies Press. https://www.ncbi.nlm.nih.gov/pubmed/39146448
23. Nivedhaa, N. (2024). A comprehensive review of AI's dependence on data. *ResearchGate*. https://doi.org/10.13140/RG.2.2.27033.63840
24. Barddal, J. P., Enembreck, F., Gomes, H. M., & Pfahringer, B. (2018). Merit-guided dynamic feature selection filter for data streams. *Expert Systems with Applications, 116*, 311–326. https://doi.org/10.1016/j.eswa.2018.09.031
25. Bellavista, P., Berrocal, J., Corradi, A., Das, S. K., Foschini, L., & Zanni, A. (2019). A survey on fog computing for the Internet of Things. Pervasive and Mobile Computing, 52, 71-99. https://doi.org/10.1016/j.pmcj.2018.12.007
26. Mortaji, S. T. H., & Sadeghi, M. E. (2024). Assessing the reliability of artificial intelligence systems: Challenges, metrics, and future directions. *International Journal of Innovation in Management Economics and Social Sciences, 4*(2), 1–13. https://doi.org/10.59615/ijimes.4.2.1
27. Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

28. Kedziora, D., & Marciniak, P. (2025). Design principles for AI-enhanced process automation: An eDSR approach to intelligent data validation in financial decisions [Preprint]. *Research Square*. https://doi.org/10.21203/rs.3.rs-7206919/v1
29. Budnikov, M., Bykova, A., & Yamshchikov, I. P. (2025). Generalization potential of large language models. Neural Computing and Applications, 37, 1973–1997. https://doi.org/10.1007/s00521-024-10827-6
30. Ajiboye, A., Arshah, R. A., & Qin, H. (2015). Evaluating the effect of dataset size on predictive models using supervised learning techniques. *International Journal of Computer Systems & Software Engineering, 1*, 75–84. https://doi.org/10.15282/ijsecs.1.2015.6.0006
31. Scannapieco, M., & Catarci, T. (2002, May). Data quality under the computer science perspective. Rome, Italy.
32. Tamm, H. C., & Nikiforova, A. (2025). From data quality for AI to AI for data quality: A systematic review of tools for AI-augmented data quality management in data warehouses. In International Conference on Business Informatics Research. Springer International Publishing.
33. Heringa, M. B., Cnubben, N. H. P., Slob, W., & Hakkert, B. C. (2020). Use of the kinetically-derived maximum dose concept in selection of top doses for toxicity studies hampers proper hazard assessment and risk management. *Regulatory Toxicology and Pharmacology, 114*, 104659. https://doi.org/10.1016/j.yrtph.2020.104659
34. Mishra, S., Rao, A., Krishnan, R., & Zio, E. (2024). Reliability, resilience and human factors engineering for trustworthy AI systems [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2411.08981
35. Cho, J.-H., Xu, S., Hurley, P., & Beaumont, M. (2019). STRAM: Measuring the trustworthiness of computer-based systems. *ACM Computing Surveys, 51*(6), 1–47. https://doi.org/10.1145/3277666
36. Smith, J., & Ethan, A. (2019). Intelligent user feedback loops: AI for continuous product innovation. *International Journal of Advanced Engineering Technologies and Innovations, 1*(5), 21–XX.
37. Tadi, V. (2020). Optimizing data governance: Enhancing quality through AI-integrated master data management across industries. *American Journal of Engineering Research, 1*(3), 1–15.
38. Novelli, C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: What it is and how it works. *AI & Society, 39*(4), 1–12. https://doi.org/10.1007/s00146-023-01635-y
39. Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly, 30*(3), 611–642. https://doi.org/10.2307/25148742
40. Naeem, M., Ozuem, W., Howell, K., & Ranfagni, S. (2023). A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. *International Journal of Qualitative Methods, 22*(11). https://doi.org/10.1177/16094069231205789
41. Arshed, N., Rehman, H. U., Nazim, M., & Saher, A. (2021). Evading law of diminishing returns, a case of human capital development. Journal of Contemporary Issues in Business and Government, 27(5), 2569–2584. https://doi.org/10.47750/cibg.2021.27.05.133
42. Wellmann, F. (2013). Information theory for correlation analysis and estimation of uncertainty reduction in maps and models. Entropy, 15(4), 1464–1485. https://doi.org/10.3390/e15041464
43. Jensen, G., Ward, R. D., & Balsam, P. D. (2013). Information: Theory, brain, and behavior. Journal of the Experimental Analysis of Behavior, 100(3), 408–431. https://doi.org/10.1002/jeab.49