International Journal of Multidisciplinary Research and Publications
ISSN (Online): 2581-6187

Robustness and Reliability of Machine Learning Methods under Data Imperfections: A Review

Qin Li[1], Wei Zhang[2]
[1]Department of Engineering Management, Anhui University of Finance and Economics, Bengbu, Anhui, China-233030
[2]Department of Engineering Management, Anhui University of Finance and Economics, Bengbu, Anhui, China-233030
Email address: xuqunzhi@126.com

Abstract*—Robustness and reliability have become central concerns in modern statistical and machine learning methods, particularly when models are deployed under imperfect data conditions. Real-world data are often affected by noise, outliers, contamination, distribution shift, and model misspecification, which can substantially degrade model performance and undermine trust in automated decision systems. This review provides a concise and structured overview of robustness and reliability in statistical and machine learning methods under data imperfections. We first introduce a taxonomy of common data imperfections and clarify the conceptual distinction between robustness and reliability from both statistical and machine learning perspectives. We then review major methodological directions, including classical robust statistical methods, distributionally robust optimization, adversarial robustness in machine learning, and reliability assessment under distribution shift, with an emphasis on their underlying assumptions, guarantees, and limitations. A comparative synthesis highlights shared principles and fundamental differences across these approaches. Finally, we discuss open challenges and future research directions, including unified definitions of robustness and reliability, evaluation standards under complex data modalities, and connections with interpretability and causal reasoning. This review aims to provide a coherent entry point for researchers and practitioners seeking robust and reliable modeling strategies in the presence of imperfect data.*

***Keywords*—** *Robustness; Reliability; Machine learning systems; Adversarial robustness; Dataset shift; Model uncertainty.*

I. INTRODUCTION

Robustness and reliability have emerged as central themes in modern statistical and machine learning research, driven by the increasing deployment of data-driven models in complex and high-stakes real-world environments. Unlike idealized settings assumed in classical theory, real data are often affected by various forms of imperfections, including noise, outliers, contamination, distribution shift, and model misspecification. These imperfections can significantly degrade model performance, invalidate theoretical guarantees, and undermine trust in automated decision-making systems. As a result, developing and understanding robust and reliable modeling strategies has become a critical challenge across statistics, machine learning, and their applications.

The study of robustness has a long history in statistics. Classical robust statistics focuses on constructing estimators that remain stable under small deviations from idealized model assumptions, such as the presence of outliers or contamination in the data [1]–[5]. Concepts such as influence functions, breakdown points, and robust regression provide principled tools to quantify and mitigate the impact of anomalous observations [2]–[5]. More recently, these ideas have been extended to high-dimensional and heavy-tailed settings, where robustness must be achieved under both statistical and computational constraints [6]–[9].

In parallel, robustness has become a major topic in machine learning, though often with a different emphasis. In this literature, robustness is frequently studied in the context of adversarial perturbations, worst-case performance guarantees, and algorithmic stability [14]–[16]. Related developments in distributionally robust optimization aim to protect learning algorithms against uncertainty in the data-generating distribution by optimizing worst-case risks over ambiguity sets [10]–[13]. These approaches provide powerful tools for controlling performance degradation but often rely on strong modeling assumptions or conservative worst-case formulations.

Closely related, but conceptually distinct, is the notion of reliability. While robustness typically concerns sensitivity to data perturbations, reliability focuses on the consistency, calibration, and trustworthiness of model outputs, especially under distribution shift or deployment mismatch. Reliability has been studied through topics such as covariate shift adaptation, out-of-distribution detection, uncertainty quantification, and probabilistic calibration [17]–[23]. From a systems perspective, reliability also encompasses issues arising in the full lifecycle of machine learning models, including data pipelines, model updates, and hidden technical debt [24]–[26].

Despite substantial progress, the literature on robustness and reliability remains fragmented across disciplines and methodologies. Statistical and machine learning communities often adopt different definitions, assumptions, and evaluation criteria, making it difficult to compare approaches or develop unified perspectives. Moreover, many existing surveys focus on narrow subtopics, such as adversarial robustness or uncertainty estimation, without systematically addressing the broader landscape of data imperfections and their implications for both robustness and reliability.

The goal of this review is to provide a concise and structured overview of robustness and reliability in statistical and machine learning methods under data imperfections. Rather than offering an exhaustive survey, we focus on identifying common types of data imperfections, summarizing major methodological directions, and highlighting shared principles and fundamental differences across approaches. We

Qin Li and Wei Zhang, "Robustness and Reliability of Machine Learning Methods under Data Imperfections: A Review," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 8, Issue 7, pp. 18-22, 2025.

place particular emphasis on the assumptions underlying different methods, their theoretical or empirical guarantees, and their limitations in practice. Finally, we discuss open challenges and future research directions, including unified conceptual frameworks, evaluation standards under complex data modalities, and connections with interpretability and causal reasoning [27]–[33].

The remainder of this paper is organized as follows. Section 2 introduces a taxonomy of common data imperfections and clarifies their implications for robustness and reliability. Section 3 reviews major methodological directions in statistical and machine learning research, including robust statistical methods, distributionally robust and adversarial learning approaches, and reliability assessment under distribution shift. Section 4 discusses open challenges and future research directions. Finally, Section 6 concludes the paper.

II. Type of data Imperfections

Real-world data imperfections arise from multiple sources and affect statistical and machine learning methods in fundamentally different ways. A clear taxonomy of these imperfections is essential for understanding what types of robustness and reliability guarantees can be expected, and under which assumptions existing methods operate. In this review, we focus on five broad categories of data imperfections that are most commonly encountered in practice.

A. Noise and Outliers

Noise refers to random perturbations in measurements, while outliers correspond to atypical observations that deviate substantially from the bulk of the data. These imperfections are central to classical robust statistics and motivate estimators with bounded influence and high breakdown points [1]–[5]. Noise and outliers primarily affect estimation accuracy and stability, and robustness is typically defined in terms of insensitivity to a small fraction of anomalous observations.

B. Data Contamination and Adversarial Corruption

Beyond isolated outliers, contamination models allow a non-negligible fraction of data to be arbitrarily corrupted. In modern settings, such corruption may be adversarial rather than random, posing stronger challenges for learning algorithms. This type of imperfection motivates minimax formulations, robust estimation under contamination, and adversarial robustness in machine learning [6]–[9], [14]–[16]. Here, robustness is often framed as performance guarantees under worst-case perturbations

C. Distributional Shift

Distribution shift occurs when the training and deployment data are drawn from different distributions, due to changes in population, environment, or data collection mechanisms. Common forms include covariate shift, label shift, and concept drift. Distribution shift primarily impacts model reliability rather than pointwise robustness, as predictive performance and uncertainty estimates may degrade even in the absence of explicit contamination [17]–[20].

D. Model Misspecification

Model misspecification arises when the assumed statistical or algorithmic model fails to capture the true data-generating process. This includes incorrect parametric assumptions, simplified likelihoods, or mismatched inductive biases in machine learning models. Misspecification can lead to systematic errors that are not mitigated by classical robustness techniques and often requires broader notions of reliability and model assessment [10]–[13].

E. Limited, Biased, or Incomplete Data

Practical datasets are often small, imbalanced, or subject to selection bias. These imperfections can amplify sensitivity to noise and shift, and they interact strongly with uncertainty quantification and calibration. Reliability under limited or biased data is closely tied to generalization, stability, and the interpretability of model outputs [21]–[23]

While these categories are conceptually distinct, they frequently co-occur in real applications. For example, distribution shift may exacerbate the impact of contamination, and model misspecification may undermine uncertainty estimates under noisy data. This taxonomy provides a unifying lens for organizing existing methods and clarifies that robustness and reliability are inherently context-dependent notions. In the next section, we review major methodological directions that address these data imperfections from statistical and machine learning perspectives.

III. Major Methodological directions

This section reviews major methodological directions that address robustness and reliability under data imperfections. Rather than providing exhaustive technical details, we emphasize the core ideas, assumptions, and limitations that distinguish different approaches.

A. Robust Statistical Methods

Robust statistical methods aim to construct estimators whose performance degrades gracefully under deviations from idealized assumptions. Classical approaches focus on bounding the influence of individual observations and ensuring stability under a small fraction of anomalous data. Foundational tools include M-estimators, influence functions, and breakdown points, which provide interpretable measures of robustness against noise and outliers [1]–[5]. Robust regression and covariance estimation extend these ideas to multivariate settings and are widely used in practice [4], [5].

Recent work has significantly advanced robust estimation in high-dimensional and heavy-tailed regimes. These methods address settings where both classical asymptotics and computational tractability pose challenges. Examples include robust mean and covariance estimators with provable guarantees under contamination models and heavy-tailed distributions [6]–[9]. A common feature of these approaches is the explicit specification of a contamination model or moment condition, which enables finite-sample performance bounds but also limits applicability when assumptions are violated.

Overall, robust statistical methods provide principled and interpretable guarantees under well-defined imperfection

Qin Li and Wei Zhang, “Robustness and Reliability of Machine Learning Methods under Data Imperfections: A Review,” *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 8, Issue 7, pp. 18-22, 2025.

models. However, they are often tailored to specific estimands and may not directly extend to complex predictive models or highly nonparametric settings.

B. Distributionally Robust and Adversarial Learning

In machine learning, robustness is frequently studied through worst-case formulations. Distributionally robust optimization (DRO) seeks models that perform well against adversarial perturbations of the data-generating distribution within a prescribed ambiguity set [10]–[13]. By framing learning as a minimax problem, DRO provides a unifying perspective that connects classical robustness with modern optimization-based methods.

Adversarial robustness in deep learning represents a related but distinct line of work. Here, robustness is defined with respect to small, often norm-bounded, perturbations of input data that are designed to induce misclassification [14]–[16]. These methods have revealed fundamental trade-offs between robustness and nominal accuracy and have motivated new training procedures, such as adversarial training, that explicitly incorporate worst-case perturbations.

While DRO and adversarial learning offer strong protection against specific classes of perturbations, they can be overly conservative and computationally demanding. Moreover, guarantees are typically tied to carefully chosen uncertainty sets or perturbation models, which may not accurately reflect real-world data imperfections.

C. Reliability under Distribution Shift

Reliability addresses the consistency and trustworthiness of model predictions, particularly when training and deployment conditions differ. Distribution shift is a primary driver of reliability failures in practice and has been studied extensively through covariate shift adaptation, dataset shift analysis, and out-of-distribution detection [17]–[20].

Complementary work focuses on uncertainty quantification and calibration, ensuring that predictive probabilities or confidence measures remain meaningful under shift [21]–[23]. From a systems perspective, reliability also depends on factors beyond model training, including data pipelines, monitoring, and model maintenance, as highlighted in studies of technical debt and system-level failures in machine learning [24]–[26].

Compared to robustness, reliability emphasizes evaluation and diagnostics rather than worst-case guarantees. Many reliability methods are empirical and application-driven, providing practical tools but limited theoretical assurances.

As a summary, across these directions, robustness and reliability are addressed through different lenses: estimator stability, worst-case optimization, and predictive trustworthiness. Each approach targets specific data imperfections and relies on distinct assumptions. The next section synthesizes these methods, highlighting common principles, key differences, and practical trade-offs.

IV. Open Challenges and future directions

Despite extensive progress in robustness and reliability research, several fundamental challenges remain open. These challenges arise not only from methodological limitations, but also from conceptual fragmentation, evaluation ambiguity, and the growing complexity of real-world data and deployment settings. In this section, we highlight key directions that warrant further investigation.

A. Unifying Definitions of Robustness and Reliability

One persistent challenge is the lack of unified definitions for robustness and reliability across statistics and machine learning. In classical robust statistics, robustness is typically defined through explicit contamination models, influence functions, or breakdown points [1]–[5]. In contrast, machine learning often adopts operational definitions based on adversarial perturbations, worst-case risks, or empirical performance under stress tests [14]–[16]. Reliability, meanwhile, is frequently associated with calibration, uncertainty quantification, and system-level consistency rather than estimator stability [21]–[26].

This conceptual divergence makes it difficult to compare methods across communities or to assess whether different approaches address the same underlying problem. Developing unified frameworks that clarify the relationships among robustness, reliability, generalization, and stability remains an important open direction. Such frameworks would facilitate clearer communication and more principled method selection in practice.

B. Robustness under Complex and Multimodal Data

Much of the existing robustness literature focuses on relatively simple data modalities, such as tabular data or fixed-dimensional feature vectors. However, modern applications increasingly involve complex, high-dimensional, and multimodal data, including images, text, graphs, and heterogeneous biomedical measurements. In these settings, data imperfections may manifest in structured, correlated, or modality-specific ways that are poorly captured by existing contamination or perturbation models.

Extending robust statistical principles and reliability assessment tools to such complex data types remains challenging. In particular, it is unclear how classical notions of robustness, such as bounded influence, should be generalized to deep or multimodal architectures, or how uncertainty estimates should be interpreted under interacting sources of imperfection. Addressing these questions is critical for robust deployment in real-world systems.

C. Evaluation Standards and Benchmarking

Another major challenge concerns the evaluation of robustness and reliability. Current practice relies heavily on task-specific benchmarks, stress tests, or empirical performance metrics, which vary widely across studies. While such evaluations provide valuable insights, they often lack standardization and may not reflect realistic deployment conditions.

Developing principled evaluation protocols that align with clearly defined notions of robustness and reliability is an open problem. This includes designing benchmarks that capture realistic data imperfections, establishing metrics that reflect meaningful performance degradation, and clarifying the trade-offs between robustness, accuracy, and computational cost.

Qin Li and Wei Zhang, "Robustness and Reliability of Machine Learning Methods under Data Imperfections: A Review," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 8, Issue 7, pp. 18-22, 2025.

Improved evaluation standards would enable more transparent comparison of methods and support more reliable model selection.

D. Robustness and Reliability across the Model Lifecycle

Robustness and reliability are not static properties of a trained model, but evolve throughout the model lifecycle, from data collection and preprocessing to deployment, monitoring, and updating. System-level studies have shown that failures often arise from interactions among components rather than from isolated modeling choices [24]–[26].

Future research should more explicitly account for lifecycle considerations, including online monitoring, drift detection, model retraining, and human-in-the-loop interventions. Integrating robustness and reliability analysis into the full lifecycle perspective may help bridge the gap between theoretical guarantees and practical performance.

E. Connections with Interpretability and Causal Reasoninge

Finally, robustness and reliability are closely related to interpretability and causality, though these connections remain underexplored. Interpretable models may facilitate diagnosis of robustness failures, while causal reasoning offers principled tools for understanding distribution shift and model misspecification [31]. Conversely, unreliable uncertainty estimates or unstable predictions can undermine interpretability and trust.

Exploring these interactions may lead to more holistic approaches that combine robust estimation, reliable prediction, and causal understanding. Such integration represents a promising direction for developing trustworthy statistical and machine learning systems under imperfect data.

F. Robustness under Resource and Cost Constraints

Most existing robustness and reliability methods implicitly assume that sufficient computational, sampling, or labeling resources are available. In practice, however, robustness must often be achieved under explicit resource constraints, such as limited sample sizes, restricted labeling budgets, or real-time computational requirements. These constraints are particularly relevant in streaming, online, and large-scale systems, where trade-offs between robustness, accuracy, and cost become unavoidable.

Developing robustness-aware methods that explicitly incorporate resource constraints remains an open challenge. This includes adaptive strategies that allocate resources dynamically, as well as principled frameworks for quantifying the cost of robustness and reliability improvements. Addressing robustness under resource constraints is essential for bridging the gap between theoretical guarantees and practical deployment.

G. Robustness and Reliability in Human-in-the-Loop Systems

Many real-world decision systems involve human oversight, intervention, or interaction, yet most robustness and reliability analyses treat models as fully autonomous components. In practice, human-in-the-loop settings introduce additional sources of uncertainty, including subjective judgment, delayed feedback, and evolving decision policies. These factors can significantly influence system reliability, particularly under imperfect or shifting data conditions.

Future research should more explicitly account for the interaction between algorithmic robustness and human decision-making. This includes understanding how uncertainty estimates are communicated to users, how human feedback affects robustness over time, and how responsibility is shared between models and operators. Integrating human-in-the-loop considerations may lead to more realistic and trustworthy notions of robustness and reliability.

V. Conclusion

This review has provided a concise overview of robustness and reliability in statistical and machine learning methods under data imperfections. By introducing a taxonomy of common data imperfections, we clarified the contexts in which robustness and reliability are required and highlighted that these notions are inherently problem dependent. We reviewed major methodological directions spanning classical robust statistics, distributionally robust and adversarial learning, and reliability assessment under distribution shift, emphasizing their underlying assumptions, guarantees, and limitations.

A central message of this review is that robustness and reliability are addressed through complementary, rather than competing, perspectives. Statistical approaches offer interpretable guarantees under explicit imperfection models, while machine learning methods emphasize worst-case protection and empirical performance in complex settings. Reliability-focused techniques further broaden the scope by addressing predictive trustworthiness and system-level behavior under deployment mismatch. No single approach provides a universal solution, underscoring the importance of aligning methodological choices with the specific types of data imperfections encountered in practice.

Looking forward, progress in this area will require greater conceptual clarity, improved evaluation standards, and closer integration between methodological development and real-world deployment considerations. Bridging statistical and machine learning perspectives, and connecting robustness and reliability with interpretability and causal reasoning, may offer promising pathways toward more trustworthy data-driven systems. We hope this review serves as a useful entry point for researchers and practitioners seeking principled strategies to address data imperfections in modern statistical and machine learning applications.

References

[1] P. J. Huber, “Robust estimation of a location parameter,” *Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[2] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, 2nd ed. Hoboken, NJ:*:*

[3] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley, 1986.

[4] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. New York: Wiley, 1987.

[5] D. L. Donoho and P. J. Huber, “The notion of breakdown point,” in *A Festschrift for Erich L. Lehmann*, New York: Springer, pp. 157–184, 1983.

Qin Li and Wei Zhang, “Robustness and Reliability of Machine Learning Methods under Data Imperfections: A Review,” *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 8, Issue 7, pp. 18-22, 2025.

[6] Y. Chen, C. Gao, and A. J. Goldsmith, "Robust covariance and scatter matrix estimation under Huber's contamination model," *Annals of Statistics*, vol. 46, no. 5, pp. 1932–1960, 2018.
[7] I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, and A. Stewart, "Robust estimators in high dimensions without the computational intractability," *SIAM Journal on Computing*, vol. 48, no. 2, pp. 742–864, 2019.
[8] E. Lugosi and S. Mendelson, "Sub-Gaussian estimators of the mean of a random vector," *Annals of Statistics*, vol. 47, no. 2, pp. 783–794, 2019.
[9] R. Vershynin, "Estimation in high dimensions: A geometric perspective," in *Sampling Theory, a Renaissance*, New York: Springer, pp. 3–66, 2015.*:*
[10] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust Optimization*. Princeton, NJ: Princeton University Press, 2009.
[11] J. Blanchet and K. Murthy, "Quantifying distributional model risk via optimal transport," *Mathematics of Operations Research*, vol. 44, no. 2, pp. 565–600, 2019.
[12] R. Gao and A. J. Kleywegt, "Distributionally robust stochastic optimization with Wasserstein distance," *Mathematics of Operations Research*, vol. 41, no. 4, pp. 1548–1573, 2016.
[13] S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani, "Regularization via mass transportation," *Journal of Machine Learning Research*, vol. 18, pp. 1–68, 2017.
[14] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proceedings of the International Conference on Learning Representations*, 2015.
[15] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proceedings of the International Conference on Learning Representations*, 2018.
[16] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *Proceedings of the International Conference on Learning Representations*, 2019.
[17] S. Sugiyama, M. Krauledat, and K. R. Müller, "Covariate shift adaptation by importance weighted cross validation," *Journal of Machine Learning Research*, vol. 8, pp. 985–1005, 2007.
[18] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*. Cambridge, MA: MIT Press, 2009.
[19] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proceedings of the International Conference on Learning Representations*, 2017.
[20] B. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, B. Snoek, and J. Dillon, "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems*, pp. 13991–14002, 2019.
[21] A. P. Dawid, "The well-calibrated Bayesian," *Journal of the American Statistical Association*, vol. 77, no. 379, pp. 605–610, 1982.
[22] A. Guo, C. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the International Conference on Machine Learning*, pp. 1321–1330, 2017.
[23] J. Kuleshov, N. Fenner, and S. Ermon, "Accurate uncertainties for deep learning using calibrated regression," in *Proceedings of the International Conference on Machine Learning*, pp. 2796–2804, 2018.
[24] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, and M. Young, "Hidden technical debt in machine learning systems," in *Advances in Neural Information Processing Systems*, pp. 2503–2511, 2015.
[25] M. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," *arXiv preprint arXiv:1606.06565*, unpublished, 2016.
[26] Z. Zhang, M. Sabuncu, and D. Sontag, "A survey on reliability of machine learning systems," *IEEE Signal Processing Magazine*, vol. 40, no. 2, pp. 92–109, 2023.
[27] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.
[28] O. Bousquet and A. Elisseeff, "Stability and generalization," *Journal of Machine Learning Research*, vol. 2, pp. 499–526, 2002.
[29] G. Candes, Y. Fan, L. Janson, and J. Lv, "Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection," *Journal of the Royal Statistical Society, Series B*, vol. 80, no. 3, pp. 551–577, 2018.
[30] G. Shafer, A. Shen, N. Vereshchagin, and V. Vovk, "Test martingales, Bayes factors and p-values," *Statistical Science*, vol. 26, no. 1, pp. 84–101, 2011.
[31] J. Pearl and D. Mackenzie, *The Book of Why*. New York: Basic Books, 2018.
[32] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of the International Conference on Machine Learning*, pp. 1050–1059, 2016.
[33] R. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani, "Exact post-selection inference for sequential regression procedures," *Journal of the American Statistical Association*, vol. 111, no. 514, pp. 600–620, 2016.