Toward Robust Machine Learning under Diverse Incomplete Data Mechanisms in Real-World Applications

Youran Zhou
echo.zhou@deakin.edu.au
Deakin University
Geelong, Australia

Abstract

Incomplete data is a pervasive challenge across a wide range of data types, including tabular, sensor, time-series, image, and textual data. Its presence stems from various real-world factors and gives rise to different missingness mechanisms. While much of the existing research focuses on the Missing Completely At Random (MCAR) assumption, the more complex and realistic mechanisms—Missing At Random (MAR) and Missing Not At Random (MNAR)—remain relatively underexplored despite their prevalence and impact. This PhD project aims to systematically investigate the challenges posed by diverse Incomplete data mechanisms and to develop robust machine learning methods that can perform reliably across MCAR, MAR, and MNAR scenarios. The research spans multiple data modalities and focuses on improving both the theoretical understanding and practical handling of incomplete data. By addressing mechanism-specific imputation challenges and proposing broadly applicable solutions, this work contributes to building more resilient and trustworthy data-driven systems in real-world settings.

CCS Concepts

• **Computing methodologies** → **Machine learning**; **Artificial intelligence**; • **Information systems** → **Data mining**.

Keywords

Incomplete data; Missing Mechanism; Missing data; MNAR; MAR; Tabular data

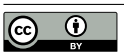
ACM Reference Format:

Youran Zhou. 2025. Toward Robust Machine Learning under Diverse Incomplete Data Mechanisms in Real-World Applications. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3746252.3761662

1 Problem

Incomplete data, where certain entries or attributes are missing from a dataset, is a pervasive and persistent challenge in real-world machine learning. It arises across the entire data lifecycle—from collection and transmission to storage and labeling. Examples include skipped survey responses, sensor failures in IoT systems, occluded regions in medical imaging, and truncated logs in time-series data. These incomplete observations can severely compromise the reliability of downstream analysis, prediction, and decision-making. Although traditionally studied in the context of structured tabular data, incompleteness is also common in text, image, and multimodal datasets. In natural language, named entities may be missing due to extraction errors. In vision, region-level annotations may be absent, or objects may be partially visible. Despite the modality, the presence of incomplete data challenges model robustness and generalizability across a broad range of AI applications. Three intertwined factors make handling incomplete data particularly difficult: **Missing Ratio:** The proportion of missing entries significantly impacts the effectiveness of imputation and learning algorithms. **Missing Mechanism [8, 20]:** The missingness mechanism—Missing Completely At Random (MCAR), Missing At Random (MAR), or Missing Not At Random (MNAR)—affects what information is recoverable, and under which assumptions. **Data Type:** Categorical, numerical, heterogeneous and time series attributes require fundamentally different strategies for recovery and modeling. The statistical theory of missingness mechanisms was originally developed for tabular data, but its conceptual foundations extend naturally to other domains. For instance, whether a label is absent in an image due to random omission, dependency on visible regions, or systematic occlusion mirrors MCAR, MAR, and MNAR respectively. The toy dataset in Table 1 illustrates how different mechanisms yield different missing entries, even within the same observed data.

Although it is unrealistic to expect a single algorithm to perform reliably under all forms of data incompleteness, effective management of incomplete data through well-designed modelling strategies is essential for trustworthy machine learning. This PhD project aims to make machine learning models more robust to incomplete data by systematically exploring three critical dimensions: missingness mechanisms (MCAR, MAR, MNAR), data types (with a focus on tabular data and evaluations on heterogeneous and time-series datasets), and modelling strategies. We study both traditional and generative approaches across two methodological paradigms: direct imputation, where missing values are explicitly filled using statistical or generative models, and representation learning, where incomplete inputs are encoded directly into robust latent features. The goal is to understand the strengths and limitations of each approach, develop new hybrid methods, and offer practical tools that integrate easily into modern ML pipelines. While our primary focus is on tabular data, the proposed methodologies are designed to be extensible to other modalities such as images and text, contributing toward more reliable, generalizable, and mechanism-aware learning in the presence of incomplete data.



CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761662

2 State-of-the-Art

Research on handling *incomplete data* spans classical statistical techniques and modern machine learning. Existing methods are typically grouped into: (i) *explicit imputation*, where missing values are filled before learning, and (ii) *representation-based learning*, which encodes incomplete inputs directly. **Explicit Imputation Methods:** Classical methods such as mean/mode substitution, k-nearest neighbors (KNN) [14], multiple imputation via chained equations (MICE) [24], the expectation-maximization (EM) algorithm [3], matrix factorization [13], and MissForest [21] are simple and effective under low missingness, but often distort data distributions under structured mechanisms (e.g., MAR or MNAR). Generative models—including VAEs [4, 10, 12, 17], GAN-based methods [1, 5, 25], and diffusion-based approaches [18, 23, 26]—learn to model complete data distributions and generate plausible imputations. These methods show promise under MAR and MNAR, yet typically assume continuous numerical inputs, demand large datasets, and are sensitive to hyperparameters. Empirical evaluations [15, 22, 29] suggest these models underperform on small-scale tabular data, especially when feature types are mixed. **Representation-Based Learning:** Representation-based approaches avoid direct imputation by encoding incomplete inputs into latent embeddings. Examples include masked autoencoders [2], GNNs for structured data [11, 27], and cross-modal architectures for multimodal or temporal contexts [7]. While effective in domains like vision and sensor data, these models often struggle with tabular data that is heterogeneous, sparsely observed, or sample-limited [6, 19]. They may also lack interpretability or adaptability to downstream tasks. **Mechanism-Aware vs. Agnostic Learning:** Rubin's taxonomy [9, 20] distinguishes three missingness mechanisms: MCAR, MAR, and MNAR. Mechanism-aware methods attempt to explicitly model the missingness process, e.g., through likelihoods, conditional masking, or two-stage learning [4, 12, 16]. In contrast, agnostic approaches aim for robustness under unknown or mixed patterns, typically through mechanism-agnostic training or regularization [15, 22].

3 Motivation and Approach

Incomplete data—where observations are partially missing across features or modalities—is a pervasive challenge in real-world machine learning. It arises throughout the data lifecycle, from user non-response and sensor malfunction to truncation in logs and annotation gaps. While traditionally studied in tabular datasets, incompleteness also affects text, image, and time-series data, threatening the reliability of downstream analysis and decision-making. Despite decades of work on imputation and learning with missing values, three key challenges persist. First, the **missing ratio** (i.e., proportion of missing entries) significantly influences learning robustness. Second, **missingness patterns** governed by different mechanisms—Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR)—require distinct modeling strategies, yet the true mechanism is often unobservable. Third, **data heterogeneity**, including categorical and mixed-type features, further complicates inference and similarity estimation. Existing approaches frequently assume low missing rates, numerical data, or benign missingness mechanisms (e.g., MCAR), limiting their applicability in real-world scenarios. Moreover, much prior work focuses on imputing missing values, rather than directly learning from incomplete data. As shown in Figure 1, research explicitly targeting MAR/MNAR (Special) settings remains relatively underrepresented.

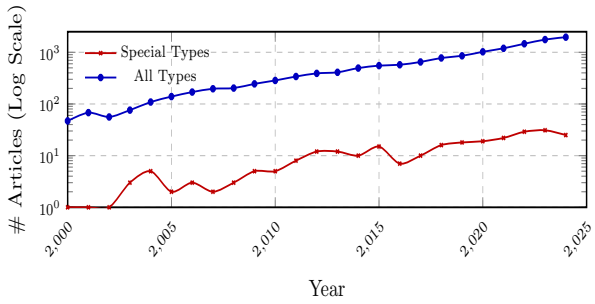


Figure 1: Article counts from Scopus keyword searches (2000–2025) for general vs. special types of incomplete data.

Complete Data		MCAR	MAR	MNAR
IQ	Rating	Rating	Rating	Rating
78	9	?	?	9
84	13	13	?	13
85	8	8	?	?
92	9	9	9	9
96	7	?	7	?
105	11	11	11	11
134	12	?	12	12

Table 1: Example of different missingness mechanisms. MCAR: randomly missing. MAR: dependent on observed IQ. MNAR: dependent on unobserved Rating.

To address the challenges outlined above, this PhD project is structured around the following key objectives, each targeting a specific gap in current approaches to learning with incomplete data. The following objectives guide this work:

(1) To analyze the robustness limitations of existing models under incomplete data for MAR and MNAR patterns, by identifying failure modes across diverse data types.
(2) To develop modeling frameworks that learn directly from incomplete inputs, using generative, kernel-based machine learning methods for reconstruction.
(3) To build a standardized toolkit for simulating, detecting, and benchmarking missingness mechanisms, supporting reproducible and mechanism evaluation for heterogeneous data.
(4) To extend these methods to heterogeneous data, establishing principled approaches that maintain statistical integrity in non-numeric settings.
(5) To unify data-level and modality-level incompleteness in multimodal datasets, enabling robust learning when views are missing in time-series, or multiple tabular datasets.

4 Methodology

To accomplish our research objectives, we employ a multi-faceted methodological framework that combines empirical benchmarking,

generative modeling, kernel-based representation learning, and tool development.

Objective 1: Investigating and Benchmarking Existing Methods. We begin by conducting a structured literature review using keyword-based searches across platforms such as Google Scholar and Scopus, categorizing prior work by data modality, experimental design, and missingness assumptions. Building on this foundation, we perform systematic evaluations of representative methods across diverse real-world (e.g., UCI Repository) and synthetic datasets. To ensure controlled comparison, we simulate incomplete data under a wide range of settings—varying missingness mechanisms (MCAR, MAR, MNAR), missing rates, and data types. This empirical analysis allows us to identify which classes of methods (e.g., imputation-based vs. representation-based) demonstrate robustness under realistic and challenging incomplete scenarios.

Objective 2: Developing Robust Models for Incomplete Data. We explore two complementary modeling paradigms. First, we investigate generative models for imputation. While existing methods such as MIWAE [12], Not-MIWAE [4], and diffusion-based models [26] rely on binary mask arrays to encode missingness, such representations often fail to capture global structural dependencies. Inspired by recent work [10], we propose using graph-based representations to model feature interdependencies and guide the imputation process, thereby enhancing robustness in MAR and MNAR settings. Second, we develop representation learning approaches that embed incomplete data directly for downstream tasks (e.g., classification, clustering, regression), avoiding explicit imputation. We employ data-dependent kernel methods to measure similarity from partially observed data. These kernels natively support categorical and discrete attributes, and—by operating directly on incomplete inputs—help reduce risks of information leakage.

Objective 3: Designing a Toolkit for Missingness Simulation and Diagnosis. To support reproducible and mechanism-aware experimentation, we construct a standardized Python toolkit for simulating, detecting, and benchmarking missingness mechanisms. Based on statistical literature and current practice, we formalize and unify common simulation strategies, extending them to ordinal and discrete variables—an underexplored area in existing libraries. We also integrate diagnostic tools such as Little's MCAR test and pairwise dependence checks for mechanism identification. This toolkit serves as a research infrastructure for the community, supporting evaluations in heterogeneous and multimodal tabular data.

Objective 4: Extending to Categorical and Heterogeneous Data. We further extend our modeling techniques to support heterogeneous tabular data encompassing categorical, ordinal, and continuous variables. For kernel-based similarity learning, we adopt a *maximum uncertainty principle*—representing missing categorical values via uniform distributions over all possible categories—thereby avoiding biased imputations and preserving uncertainty. This results in a type-aware and uncertainty-aware similarity estimation framework suitable for mixed data. For graph-based models, we propose a lightweight encoding scheme that avoids the overhead of one-hot encodings, allowing the graph structure to capture semantic relationships more effectively and improving generalization across diverse data types.

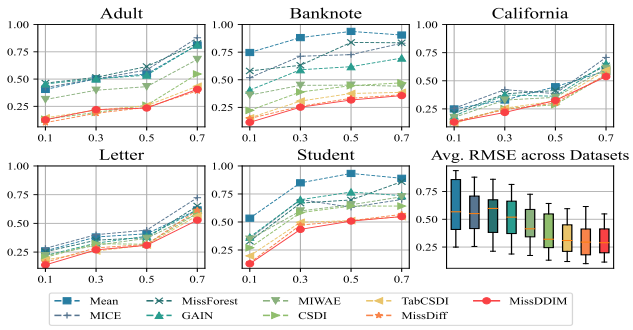


Figure 2: Preliminary results for MissDDIM: average RMSE under MCAR across five benchmark datasets.

5 Preliminary Results

Literature Review and Empirical Benchmarking: We surveyed incomplete data modeling techniques with emphasis on missingness mechanisms (MCAR, MAR, MNAR) [28], identifying gaps such as the absence of standardized missingness generation and limited support for categorical or heterogeneous data. Building on this, we benchmarked ten real-world datasets under controlled missingness [29]. Results show that many deep learning imputers achieve low reconstruction error (e.g., RMSE) but degrade on downstream tasks, revealing a mismatch between imputation fidelity and task utility. **Framework Contributions:** We developed three complementary modeling frameworks to address incomplete heterogeneous data. The first, **HI-PMK** (Heterogeneous Incomplete Probability Mass Kernel) [32], is a data-dependent kernel method that avoids imputation by directly computing similarity under uncertainty. It supports numerical, ordinal, and categorical data types, and introduces a conservative maximum uncertainty strategy for missing values. This work has been accepted at **ECAI 2025**. The second, **MissDDIM** [30], is a diffusion-based generative model that captures complex data distributions and produces consistent imputations across different missingness mechanisms. This work has been accepted as a **short paper at CIKM 2025**. Figure 2 illustrates its performance under MCAR, where MissDDIM achieves lower average RMSE across five benchmark datasets compared to classical methods. The third, **IVGAE** (Imputation via Variational Graph Autoencoder), integrates a graph-based representation with a dual-decoder architecture and Transformer-style heterogeneous embeddings to improve robustness under structured missingness. This work is currently **under review**. **Toolkit Development:** We have developed `MissMecha` [31], an open-source Python package for simulating, diagnosing, and benchmarking missingness mechanisms in tabular data. MissMecha supports MCAR, MAR, and MNAR mechanisms, and extends simulation capabilities to categorical and ordinal variables—addressing a key limitation in existing tools. It also includes statistical diagnostics such as Little's MCAR test and pairwise dependence analysis. A demo paper describing MissMecha has been submitted to the CIKM 2025 Demo Track. The toolkit is publicly available on GitHub and PyPI, with documentation and usage examples at: https://echoid.github.io/MissMecha/.

6 Conclusion and Future Work

This proposal addresses the fundamental challenge of learning from Incomplete data—a ubiquitous issue in real-world machine learning. We have systematically characterized the different mechanisms of incompleteness (MCAR, MAR, MNAR) and highlighted their theoretical implications and practical obstacles across diverse data types. Our preliminary studies, combining a literature review and empirical benchmarking, reveal that many existing methods lack robustness under structured missingness and often fail to generalize across heterogeneous feature types. To tackle these challenges, we propose a multi-faceted research agenda that includes: (1) developing generative models enhanced by graph-based structural representations, (2) designing kernel-based approaches that directly embed incomplete data without requiring imputation, and (3) releasing a standardized open-source toolkit for missingness simulation and diagnosis. Collectively, these contributions aim to advance both the theoretical foundation and the practical tools available for handling Incomplete data. While this project primarily focuses on incomplete tabular data, a natural next step is to extend our framework to more complex multimodal datasets, where missingness may occur at both feature- and modality-level. Real-world applications, such as healthcare records, sensor networks, or human activity data, often contain missing entire views (e.g., missing time series, absent image scans, or unrecorded textual notes). Future work will explore adapting our mechanism-aware and agnostic strategies to such scenarios, emphasizing flexible architectures that can accommodate modality-specific patterns and align heterogeneous inputs for robust learning under incomplete observations.

Acknowledgment

This work is supported by the Air Force Office of Scientific Research under award number FA2386-23-1-4003 and Deakin University.

Generative AI Disclosure

This paper was written entirely by the author. A generative AI tool ChatGPT was used for language polishing and minor improvements in expression clarity. The author affirms that all research ideas, methodology, analyses, and results are original and independently developed.

References

[1] Mohammed Ali Al-taezi, Yu Wang, Pengfei Zhu, Qinghua Hu, and Abdulrahman Al-Badwi. 2024. Improved generative adversarial network with deep metric learning for missing data imputation. *Neurocomputing* 570 (2024), 127062.

[2] Filippo Maria Bianchi, Lorenzo Livi, Karl Øyvind Mikalsen, Michael Kampffmeyer, and Robert Jenssen. 2019. Learning representations of multivariate time series with missing data. *Pattern Recognition* 96 (2019), 106973. doi:10.1016/j.patcog.2019.106973

[3] Zoubin Ghahramani and Michael Jordan. 1993. Supervised learning from incomplete data via an EM approach. *Advances in neural information processing systems* 6 (1993).

[4] Niels Bruun Ipsen, Pierre-Alexandre Mattei, and Jes Frellsen. 2021. not-MIWAE: Deep Generative Modelling with Missing not at Random Data. In *ICLR 2021-International Conference on Learning Representations*.

[5] Jaeyoon Kim, Donghyun Tae, and Junhee Seok. 2020. A survey of missing data imputation using generative adversarial networks. In *2020 International conference on artificial intelligence in information and communication (ICAIIC)*. IEEE, 454–456.

[6] Jae Kwang Kim and Jun Shao. 2021. *Statistical methods for handling incomplete data*. Chapman and Hall/CRC.

[7] Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. 2023. GCNet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on pattern analysis and machine intelligence* 45, 7 (2023), 8419–8432.

[8] Roderick JA Little and Donald B Rubin. 2002. Bayes and multiple imputation. *Statistical analysis with missing data* (2002), 200–220.

[9] Roderick JA Little and Donald B Rubin. 2019. *Statistical analysis with missing data*. John Wiley & Sons.

[10] Chao Ma and Cheng Zhang. 2021. Identifiable generative models for missing not at random data imputation. *Advances in Neural Information Processing Systems* 34 (2021), 27645–27658.

[11] Ivan Marisca, Andrea Cini, and Cesare Alippi. 2022. Learning to reconstruct missing data from spatiotemporal graphs with sparse observations. *Advances in Neural Information Processing Systems* 35 (2022), 32069–32082.

[12] Pierre-Alexandre Mattei and Jes Frellsen. 2019. MIWAE: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*. PMLR, 4413–4423.

[13] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. 2010. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research* 11 (2010), 2287–2322.

[14] Xiaoye Miao, Yunjun Gao, Gang Chen, Baihua Zheng, and Huiyong Cui. 2016. Processing Incomplete k Nearest Neighbor Search. *IEEE Transactions on Fuzzy Systems* 24, 6 (2016), 1349–1363. doi:10.1109/TFUZZ.2016.2516562

[15] Xiaoye Miao, Yangyang Wu, Lu Chen, Yunjun Gao, and Jianwei Yin. 2022. An experimental survey of missing data imputation algorithms. *IEEE Transactions on Knowledge and Data Engineering* 35, 7 (2022), 6630–6650.

[16] Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. 2020. Missing data imputation using optimal transport. In *International Conference on Machine Learning*. PMLR, 7130–7140.

[17] Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. 2020. Handling incomplete heterogeneous data using vaes. *Pattern Recognition* 107 (2020), 107501.

[18] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*. PMLR, 8599–8608.

[19] Roozbeh Razavi-Far, Boyuan Cheng, Mehrdad Saif, and Majid Ahmadi. 2020. Similarity-learning information-fusion schemes for missing data imputation. *Knowledge-Based Systems* 187 (2020), 104805.

[20] Donald B Rubin. 1976. Inference and missing data. *Biometrika* 63, 3 (1976), 581–592.

[21] Daniel J Stekhoven and Peter Bühlmann. 2012. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 1 (2012), 112–118.

[22] Yige Sun, Jing Li, Yifan Xu, Tingting Zhang, and Xiaofeng Wang. 2023. Deep learning versus conventional methods for missing data imputation: A review and comparative study. *Expert Systems with Applications* 227 (2023), 120201.

[23] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. 2021. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems* 34 (2021), 24804–24816.

[24] Stef Van Buuren and Karin Groothuis-Oudshoorn. 2011. mice: Multivariate imputation by chained equations in R. *Journal of statistical software* 45 (2011), 1–67.

[25] Jinsung Yoon, James Jordon, and Mihaela Schaar. 2018. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*. PMLR, 5689–5698.

[26] Shuhan Zheng and Nontawat Charoenphakdee. [n. d.]. Diffusion models for missing value imputation in tabular data. In *NeurIPS 2022 First Table Representation Workshop*.

[27] Wenqing Zheng, Edward W Huang, Nikhil Rao, Sumeet Katariya, Zhangyang Wang, and Karthik Subbian. 2022. Cold Brew: Distilling Graph Node Representations with Incomplete or Missing Neighborhoods. In *International Conference on Learning Representations*.

[28] Youran Zhou, Sunil Aryal, and Mohamed Reda Bouadjenek. 2024. Review for Handling Missing Data with special missing mechanism. arXiv:2404.04905

[29] Youran Zhou, Mohamed Reda Bouadjenek, and Sunil Aryal. 2024. Missing Data Imputation: Do Advanced ML/DL Techniques Outperform Traditional Approaches?. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 100–115.

[30] Youran Zhou, Mohamed Reda Bouadjenek, and Sunil Aryal. 2025. MissDDIM: Deterministic and Efficient Conditional Diffusion for Tabular Data Imputation. arXiv:2508.03083 [cs.AI] https://arxiv.org/abs/2508.03083

[31] Youran Zhou, Mohamed Reda Bouadjenek, and Sunil Aryal. 2025. MissMecha: An All-in-One Python Package for Studying Missing Data Mechanisms. *arXiv preprint arXiv:2508.04740* (2025).

[32] Youran Zhou, Mohamed Reda Bouadjenek, Jonathan Wells, and Sunil Aryal. 2025. HI-PMK: A Data-Dependent Kernel for Incomplete Heterogeneous Data Representation. arXiv:2501.04300 [cs.LG] https://arxiv.org/abs/2501.04300