Beyond Accuracy: An Empirical Study of Uncertainty Estimation in Imputation

Zarin Tahia Hossain
Department of Computer Science
Western University
London, Ontario, Canada
zarin.hossain@uwo.ca

Mostafa Milani
Department of Computer Science
Western University
London, Ontario, Canada
mostafa.milani@uwo.ca

Abstract—Handling missing data is a central challenge in data-driven analysis. Modern imputation methods not only aim for accurate reconstruction but also differ in how they represent and quantify uncertainty. Yet, the reliability and calibration of these uncertainty estimates remain poorly understood. This paper presents a systematic empirical study of uncertainty in imputation, comparing representative methods from three major families: statistical (MICE, SoftImpute), distribution alignment (OT-Impute), and deep generative (GAIN, MIWAE, TabCSDI). Experiments span multiple datasets, missingness mechanisms (MCAR, MAR, MNAR), and missingness rates. Uncertainty is estimated through three complementary routes: multi-run variability, conditional sampling, and predictive-distribution modeling and evaluated using calibration curves and the Expected Calibration Error (ECE). Results show that accuracy and calibration are often misaligned: models with high reconstruction accuracy do not necessarily yield reliable uncertainty. We analyze method specific trade-offs among accuracy, calibration, and runtime, identify stable configurations, and offer guidelines for selecting uncertainty-aware imputers in data cleaning and downstream machine learning pipelines.

Index Terms—Data Imputation, Data Cleaning, Uncertainty, Confidence, Calibration Curve, Expected Calibration Error

I. Introduction

Imputation is the process of estimating missing values in partially observed datasets, forming a foundational step in statistical analysis, machine learning pipelines, and data-centric scientific research. It underpins valid statistical inference, reliable model training, and robust decision systems in domains such as healthcare, finance, and the social sciences. A broad methodological landscape has emerged. Classical statistical approaches include deterministic rules and hot-deck procedures [1], likelihood-based formulations such as EM [2] and Multiple Imputation (MI) [3] with Bayesian MCMC variants [4]. Machine learning methods frame imputation as prediction from observed features, including latent-variable models (probabilistic PCA) [5], [6], instance-based schemes (KNN) [7], [8], chained equations (MICE) [9], and ensemble models such as MissForest [10]. Matrix and optimization-based approaches view it as low-rank recovery or regularized estimation, using nuclear-norm minimization [11], SoftImpute [12], or matrix factorization [13]. Distributional alignment via optimal transport offers an alternative perspective, utilizing entropy-regularized Sinkhorn solvers to match observed and imputed distributions [14]. Recent deep generative models leverage representation learning to capture complex dependencies, including denoising autoencoders (MIDA) [15], variational methods (VAEAC, MIWAE) [16], [17], adversarial training (GAIN) [18], and diffusion-based imputers for tabular data [19], [20].

While recovering plausible values is essential, an equally important question is how *confident* we should be in those imputations. In modern ML, uncertainty is typically divided into aleatory and epistemic [21], [22], [23], [24]. Aleatory uncertainty reflects inherent variability in the data-generating process and remains even with unlimited data. Epistemic uncertainty arises from limited information or model misspecification and can, in principle, be reduced. Both appear in imputation: some missing values are genuinely unpredictable given the observed covariates (aleatory), while others are uncertain due to sparse features, strong modeling assumptions, or stochastic training (epistemic).

Reliable uncertainty is crucial in practice. It indicates how much trust to place in each imputed value and enables *active imputation*, where high-uncertainty entries are prioritized for expert review or new data collection (e.g., remeasuring a patient's blood pressure when the imputation seems unreliable). It also supports *selective imputation* by flagging estimates that should not drive sensitive decisions, such as loan approvals. Under dataset shift, elevated uncertainty warns that the model is operating outside its training distribution, helping prevent overconfident, incorrect imputations. Uncertainty is not only a measure of reliability—it is a practical tool for guiding human oversight, managing risk, and improving data quality.

Several lines of work already produce uncertainty alongside imputations. Classical MI treats uncertainty as a first-class quantity by generating multiple randomized completions and combining within- and between-imputation variability [3], [4], [25]. Likelihood-based and Bayesian implementations provide posterior draws for both parameters and missing data. Deep generative models yield samples from learned conditional distributions: VAE-based methods (e.g., VAEAC, MIWAE) produce stochastic imputations via latent-variable densities [16], [17]; adversarial models such as GAIN support sampling through a generator–discriminator game [18]; and recent diffusion-based approaches instantiate

flexible conditional samplers for complex missingness patterns [19]. In parallel, generic uncertainty tools—bootstrap resampling [26], Bayesian deep learning approximations such as MC dropout [27], and distribution-free conformal prediction [28] have been adapted to imputation pipelines without changing the core imputer.

However, despite this progress, we still lack a comprehensive understanding of *how reliable* these uncertainties are across methods, datasets, and missingness regimes. Most empirical studies emphasize point accuracy (e.g., MAE or MSE) and report uncertainty only informally. As a result, it remains unclear whether the predictive distributions and intervals produced by these models are calibrated, whether nominal probabilities match empirical frequencies [29]. Consequently, practitioners have little guidance on the trade-offs between accuracy and uncertainty calibration. This gap is particularly evident across missingness mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR), where the identifiability of conditional distributions and the influence of model assumptions differ substantially. It also persists across regimes of feature dimension, sample size, correlation structure, and missing rate, all of which can shift the calibration–accuracy balance.

This paper addresses these gaps with a systematic, uncertainty-centric evaluation of representative imputation approaches from three traditions: statistical and model-based methods (e.g., MICE), matrix/optimization-based methods (e.g., SoftImpute, OT-Impute) and deep generative models (e.g., GAIN, MIWAE, diffusion-based TabCSDI). Our study is designed around three principles. First, we standardize *how* uncertainty is extracted across methods, using either repeated runs with independent randomness, posterior or conditional sampling when available, or direct predictive distributions. Second, we evaluate uncertainty *quality* using a calibration curve and the Expected Calibration Error (ECE), alongside accuracy metrics, so that calibration and point performance can be compared on equal footing [29]. Third, we evaluate across missingness mechanisms, missing rates, and diverse datasets to probe method behaviour under realistic variation in data structure and information loss.

Our findings reveal a consistent misalignment between accuracy and calibration. Methods that excel in point prediction are not necessarily those that provide well-calibrated uncertainty. For example, we observe that MICE, despite its simplicity, tends to maintain stable calibration across several datasets and missingness regimes, while SoftImpute achieves competitive accuracy but exhibits poor calibration in many settings. Among deep generative approaches, MIWAE often strikes a favourable balance between accuracy and calibration but at higher computational cost, reflecting the expense of latent-variable training and sampling. These results show that the choice of imputation strategy should depend on the requirements of the downstream task. When decisions rely on trustworthy uncertainty, such as in medical triage, selective automation, or human-in-the-loop review, it is preferable to use approaches that produce well-calibrated uncertainty, even if they sacrifice some point

accuracy. In contrast, when both speed and high point-accuracy are the main priorities, approaches optimized for efficient and accurate imputation may be suitable, provided that their uncertainty estimates are calibrated afterward to ensure they remain reliable.

This work presents a unified framework and comprehensive empirical analysis of uncertainty in data imputation. By standardizing how uncertainty is extracted and evaluated under realistic missingness mechanisms, we clarify how different families of imputers trade off accuracy, calibration, and efficiency. The paper is structured as follows. Section II reviews related work on imputation and uncertainty estimation. Section III formalizes the imputation problem, missingness mechanisms, and representative methods. Section IV describes our experimental design, uncertainty extraction strategies, datasets, and evaluation metrics. Section V presents empirical results on accuracy, calibration, and runtime, followed by key takeaways in Section VI. Finally, Section VII concludes with a discussion of implications and future directions.

II. RELATED WORK

Research on data imputation spans statistical heuristics, supervised learning, low-rank and factorization models, distribution-matching formulations, and deep generative methods [30], [9]. Early baselines fill missing entries using variable-wise summaries or donor values (mean/median/mode and hot-deck). These are fast and easy to deploy but fail to preserve multivariate structure, often biasing correlations; they mainly serve as references for more principled approaches.

Supervised learning treats imputation as conditional prediction from observed features. Multiple Imputation (MI) adds stochasticity to generate multiple plausible completions and combine them for inference [3]. A common variant, Multivariate Imputation by Chained Equations (MICE), models each variable given the others and injects randomness through coefficient or residual sampling [9]. Nonparametric and ensemble variants replace regressors with flexible learners: MissForest uses random forests [10], while KNN imputers average over local neighborhoods [7]. These perform well for moderate dimensions and nonlinear relations and, when repeated, yield multiple imputations.

A complementary line assumes a low-dimensional latent structure. Matrix completion methods recover missing entries from latent factors, with SoftImpute performing iterative soft-thresholded SVD under a nuclear-norm penalty [12]. Probabilistic PCA and Bayesian variants sample posteriors for latent factors and missing values, enabling uncertainty-aware imputations [6]. When correlations are captured by a few components, these methods achieve high accuracy with modest computation. Beyond low rank, distribution alignment casts imputation as matching empirical distributions of observed and completed data. Optimal Transport provides a geometric formulation solved via entropy-regularized Sinkhorn iterations [14]; recent OT-based imputers align distributions while controlling reconstruction cost and preserving structure [31].

Deep generative models learn flexible conditional distributions for missing values. VAE-based methods reconstruct masked inputs through encoder–decoder architectures; MIWAE improves training via importance weighting, and VAEAC supports arbitrary missingness patterns [17], [16]. Adversarial models such as GAIN rely on a discriminator to distinguish observed from imputed entries, capturing complex nonlinear dependencies [18]. Diffusion models offer another route by reversing a noise process, yielding high-quality, multimodal imputations for tabular data (e.g., TabDDPM) [20]. Although highly accurate, these neural approaches can be computationally demanding.

Uncertainty has deep roots in statistics and machine learning. Frequentist tools quantify sampling variability via confidence and prediction intervals [32], [33], whereas Bayesian inference represents parameter and predictive uncertainty through posteriors [34]. In ML, concerns about overconfidence revived attention to calibration and to the distinction between aleatoric and epistemic uncertainty [21]. In imputation, MI explicitly targets uncertainty by combining within- and between-imputation variability [3], [25], and generic techniques like bootstrap, MC dropout, and conformal prediction provide uncertainty without modifying the imputer [26], [27], [28]. Recent work integrates uncertainty directly into modern imputers: conditional VAE frameworks estimate epistemic variance via dropout sampling, β-VAE variants trade sharpness for calibrated coverage, and retrieval-augmented Gaussian-process imputers yield posterior predictive variances that guide calibration and neighbour selection [35], [36], [37]. Empirical studies also show that calibrated imputation uncertainty enhances reliability in temporal clinical prediction and downstream ML pipelines [38], [39], [40].

Our work complements these efforts by providing a unified, uncertainty-centric evaluation of classical, optimization-based, and deep generative imputers. We standardize uncertainty extraction (via repeated runs, sampling, or predictive distributions) and assess reliability through calibration curves and Expected Calibration Error alongside point accuracy, clarifying trade-offs among accuracy, calibration, and computation across missingness mechanisms and data regimes.

Beyond imputation, probabilistic data-cleaning methods also model uncertainty explicitly. OTClean [41] treats conditional-independence violations as distributional shifts and repairs them via optimal transport, while CurrentClean [42], [43] addresses stale values by capturing temporal uncertainty. These methods use probabilistic structure to produce more reliable cleaned data, whereas we focus on calibrating uncertainty in missing-value imputation.

III. PRELIMINARIES

This section formalizes the imputation problem and reviews the details of the imputers evaluated in our study.

A. Imputation Problem and Missingness Mechanisms

Let $X \in \mathbb{R}^{n \times d}$ be a data matrix with n rows from an unknown distribution $\mathbb{P}_\theta$ and d attributes. Missingness is represented by a binary mask $M \in \{0, 1\}^{n \times d}$ with $M_{ij} = 1$ if X_{ij} is observed and 0 otherwise. Using the Hadamard product $\odot$ and the all-ones matrix $\mathbf{1}_{n \times d}$,

$$X^{\text{obs}} = X \odot M, \qquad X^{\text{mis}} = X \odot (\mathbf{1}_{n \times d} - M).$$

We model the joint distribution of data and missingness as

$$\mathbb{P}(X, M) = \mathbb{P}_\theta(X) \, \mathbb{P}_\phi(M \mid X),$$

where $\mathbb{P}_\phi(M \mid X)$ is the missingness mechanism. The dataset-level imputation target is the conditional

$$\mathbb{P}\big(X^{\text{mis}} \mid X^{\text{obs}}, M\big) \propto \int \mathbb{P}_\theta(X) \, \mathbb{P}_\phi(M \mid X) \, \mathbb{P}(\theta) \, d\theta, \quad (1)$$

from which samples yield plausible completions and conditional means provide point imputations. Focusing on a single record x_i with mask m_i, while borrowing strength from the entire dataset through the posterior over θ,

$$\mathbb{P}\big(x_i^{\text{mis}} \mid x_i^{\text{obs}}, m_i, X^{\text{obs}}, M\big) =$$
$$\int \mathbb{P}_\theta\big(x_i^{\text{mis}} \mid x_i^{\text{obs}}\big), \mathbb{P}(\theta \mid X^{\text{obs}}, M), d\theta. \quad (2)$$

Under row i.i.d. assumptions, conditioning on θ factorizes across records:

$$\mathbb{P}_\theta\big(X^{\text{mis}} \mid X^{\text{obs}}\big) = \prod_{i=1}^{n} \mathbb{P}_\theta\big(x_i^{\text{mis}} \mid x_i^{\text{obs}}\big),$$

so dependence across rows arises only from uncertainty in θ (or other shared latents) via $\mathbb{P}(\theta \mid X^{\text{obs}}, M)$.

We use Rubin's taxonomy [30]. In MCAR, missingness is independent of the data, $\mathbb{P}_\phi(M \mid X) = \mathbb{P}_\phi(M)$; imputations can be accurate with well-calibrated uncertainty. In MAR, missingness depends only on observed values, $\mathbb{P}_\phi(M \mid X) = \mathbb{P}_\phi(M \mid X^{\text{obs}})$; accurate imputation remains feasible if predictors of missingness are modeled, though uncertainty typically increases. In MNAR, $\mathbb{P}_\phi(M \mid X) \neq \mathbb{P}_\phi(M \mid X^{\text{obs}})$; missingness depends on unobserved information (potentially the missing value itself), and both accuracy and calibration degrade unless the mechanism is explicitly modeled.

B. Imputation Methods

We evaluate representative methods spanning classical regression-based multiple imputation, convex low-rank recovery, distribution matching via optimal transport, adversarial learning, variational inference, and diffusion-based conditional generation. Beyond point accuracy, we standardize how uncertainty is obtained (repeated runs, conditional/posterior sampling, or direct predictive distributions).

1) MICE is based on Multiple Imputation, MI, [3] that treats missing values as random draws from their predictive distribution, producing several (k) completed datasets to reflect uncertainty due to missingness. Each dataset is analyzed separately, and results are pooled using Rubin's rules, which combine within-imputation and between-imputation variances to yield valid statistical inference. While traditional MI relied on joint parametric models (e.g., multivariate normal), these

approaches are infeasible for large, mixed-type datasets. Multiple Imputation by Chained Equations, MICE, addresses this by modeling each incomplete variable conditionally on others using regression models suitable for their data type (linear, logistic, Poisson, etc.). The algorithm iteratively imputes missing values until convergence, generating multiple completed datasets through stochastic draws. MICE assumes data are MAR and performs best when strong auxiliary predictors are included. It allows users to enforce constraints and tailor models to individual variables. Although it does not always correspond to a coherent joint model and may be biased under MNAR conditions, MICE remains widely used because of its practicality, flexibility, and solid empirical performance in many real-world applications.

2) SoftImpute [12] is a low-rank matrix completion method designed to efficiently recover missing entries in large data matrices. It assumes that the complete data can be well-approximated by a low-rank structure, where most variability is captured by a few latent factors (for example, user and item preferences in recommendation systems). The method formulates the imputation task as a convex optimization problem that balances reconstruction accuracy on observed entries with a nuclear-norm penalty on the estimated matrix. The nuclear norm serves as a convex surrogate for matrix rank, and the regularization parameter controls the effective dimensionality of the solution, similar to how the ℓ_1 norm regularizes sparsity in Lasso regression. Algorithmically, SoftImpute iteratively fills in the missing values and applies soft-thresholded singular value decomposition (SVD) to update the estimate, shrinking singular values according to the regularization strength. Using warm starts and sparse matrix operations, the algorithm scales efficiently to large datasets. Although newer deep generative and diffusion-based methods often achieve higher accuracy and better uncertainty calibration, SoftImpute remains a foundational, interpretable, and computationally efficient baseline for large-scale imputation tasks based on low-rank modeling.

3) OT-Impute [31] performs imputation through distribution matching using optimal transport (OT). The method assumes that random subsets of the dataset should share the same distribution, encouraging imputations that preserve both local structure and global data geometry. It minimizes the Sinkhorn divergence between pairs of mini-batches sampled from the imputed matrix, which measures the transport cost required to align their empirical distributions. The Sinkhorn divergence is a differentiable, entropic-regularized approximation of the Wasserstein distance that can be efficiently optimized using matrix scaling algorithms. In practice, missing entries are initialized with noisy column means and updated iteratively via stochastic gradient descent on the Sinkhorn loss. This non-parametric approach optimizes imputed values directly without assuming a specific generative model. A parametric extension, trained with the same loss in a round-robin fashion, enables out-of-sample imputation, though our experiments focus on the nonparametric variant. Empirically, OT-Impute achieves strong performance under MCAR, MAR, and even MNAR

mechanisms, offering a flexible alternative to low-rank and deep generative imputers by directly enforcing distributional alignment rather than relying on structural assumptions.

4) GAIN [18] adapts the generative adversarial network (GAN) framework for missing data imputation. It consists of a generator that proposes imputations for missing values and a discriminator that attempts to distinguish between observed and imputed entries. The generator is trained through an adversarial loss to fool the discriminator while maintaining reconstruction accuracy on observed values through an additional loss term. By sampling different noise vectors, GAIN naturally produces multiple imputations for the same partially observed instance. Empirically, GAIN achieves strong accuracy on mixed-type datasets and moderate to high missingness rates. However, it inherits typical GAN challenges such as instability, sensitivity to hyperparameters, and limited theoretical guarantees beyond MCAR. Moreover, while it generates diverse imputations, it lacks calibrated uncertainty estimates. To address this, we extended GAIN to a heteroscedastic version (denoted by GAIN-U in Section V) where the generator outputs both means and variances, enabling uncertainty quantification while retaining the adversarial objective. It replaces the reconstruction MSE with a Gaussian negative log-likelihood on observed coordinates, while retaining the adversarial term.

MIWAE [17] extends the importance-weighted autoencoder (IWAE) [44] to perform variational inference directly on incomplete data under the MAR assumption. It models each record using latent variables drawn from a prior and reconstructs the observed features through a neural decoder. An encoder network parameterized by γ provides an amortized approximation to the posterior, producing distributional parameters (e.g., mean and variance) from partially observed inputs. Training maximizes a missing-data importance-weighted bound, a tighter version of the evidence lower bound (ELBO) that uses multiple importance samples to approximate the observed-data log likelihood. Only observed entries contribute to the loss, enabling the model to train on incomplete records without discarding data. The bound approaches the true likelihood as the number of samples increases. MIWAE combines probabilistic a approache with deep-learning scalability. It is theoretically sound under MAR, easy to train, and produces both point and multiple imputations that capture uncertainty.

5) TabCSDI [45] extends conditional score-based diffusion models to tabular data with mixed numerical and categorical features. The method learns to generate plausible values for the missing part x_i^{mis} conditioned on the observed part x_i^{obs} through a self-supervised denoising objective. During training, some features are randomly masked and reconstructed from noisy versions, while observed values remain fixed to provide context. The model performs a standard forward noising process and trains a neural network to predict the injected noise, effectively learning the reverse diffusion dynamics that recover clean samples during inference. Architecturally, TabCSDI uses a transformer-based encoder adapted for non-temporal data. Multiple imputations are obtained by sampling different re-

verse diffusion trajectories, though the process is computationally expensive due to many reverse steps. Its main limitations include high inference cost, difficulty with very high-cardinality categoricals, lack of explicit handling of MNAR mechanisms, and limited uncertainty calibration. Nonetheless, TabCSDI offers a powerful likelihood-free framework for coherent, distribution-aware imputation.

IV. Experimental Methodology for Imputation Uncertainty

The purpose of this study is to evaluate imputation methods not only by their accuracy but also by the reliability of the uncertainty they provide. Most prior work has treated imputations as point estimates, focusing on minimizing reconstruction error. Yet missing values rarely have a single correct completion, and variability across plausible imputations can be as important as mean accuracy. Without uncertainty estimates, users risk overconfidence in the filled data, which can mislead downstream analyses and decision-making. This motivates a systematic comparison of state-of-the-art imputation methods under a controlled experimental framework. We aim to assess how uncertainty is captured, how well it is calibrated, and what trade-offs arise between accuracy, runtime, and uncertainty quality. The broader goal is to establish imputation methods as tools that provide not only completed datasets but also trustworthy information about the confidence of those completions.

A. Approaches to Computing Uncertainty

There are many ways to quantify uncertainty in predictive modeling, but three approaches have emerged as the most natural and widely used in the context of data imputation. These approaches are consistent with common practices in statistics, ML, and generative modeling, and they align with how uncertainty is typically approximated when the true posterior distribution is intractable. Each reflects a different point of entry for stochasticity and provides a complementary view on uncertainty.

1) *Repeated model runs.* This approach quantifies uncertainty by running the imputer multiple times with different seeds, bootstrapped data, or initialization noise, and measuring the variability of the outputs. We refer to this as the vanilla approach and use the plain method name to denote it (e.g., in the experiments, GAIN refers to running the GAIN imputer method multiple times). The idea mirrors MI in classical missing-data analysis. The main challenge is computational cost: repeated training can be expensive for complex models such as GANs, VAEs, or diffusion models. Another practical issue is deciding how many runs are sufficient to obtain stable variance estimates. Too few runs may underestimate epistemic variability, while too many runs can be prohibitively expensive.

2) *Sampling from the conditional distribution.* Probabilistic generative models allow multiple imputations to be drawn from a fixed, trained model. This probes aleatoric uncertainty, as the parameters are held fixed and randomness enters through the latent space or injected noise. Compared to repeated runs, this approach is more efficient because training is done only once. However, it requires models that explicitly support conditional sampling (e.g., GAIN, MIWAE, diffusion-based imputers). The technical challenge is deciding how many samples to draw: too few gives noisy estimates of variability, while too many increases the runtime without much benefit. For a model X, we use X-S to denote the sampling-based variant.

3) *Predictive distribution modeling.* Some models are trained to output not only point estimates but also parameters of a predictive distribution (e.g., mean and variance for a Gaussian likelihood). This approach provides per-cell parametric uncertainty directly as part of the model output. Its main challenge lies in model design and training: the likelihood must be specified correctly, and the loss function must encourage meaningful variance estimation. Poorly specified likelihoods can lead to overconfident or underconfident uncertainty estimates. Moreover, these models can be harder to train and tune, since variance parameters are more sensitive to optimization instabilities. For a model X, we use X-U to denote the uncertainty-output variant.

Together, these strategies approximate the posterior predictive distribution in Eq 1 with different trade-offs between computational efficiency, modeling flexibility, and uncertainty calibration quality.

Each imputation algorithm in our study implements one or more of these strategies according to its modeling structure:

- *MICE:* Supports only repeated runs (multiple chains); no sampling or uncertainty-output variant.
- *OT-Impute:* Provides uncertainty from repeated optimization with different initializations or minibatch orders.
- *SoftImpute:* Deterministic by design; any minor stochasticity from randomized SVD is treated as pseudo-uncertainty.
- *GAIN:* Supports all three approaches: multi-run (GAIN), conditional sampling (GAIN-S), and heteroscedastic predictive modeling (GAIN-U).
- *MIWAE:* Likewise supports multi-run (MIWAE), conditional sampling (MIWAE-S), and decoder-based predictive variances (MIWAE-U).
- *TabCSDI:* Supports multi-run (TabCSDI) and sampling through multiple reverse-diffusion trajectories (TabCSDI-S); no explicit uncertainty-output variant.

Overall, our evaluation includes six vanilla imputers (MICE, OT-Impute, SoftImpute, GAIN, MIWAE, TabCSDI), three sampling-based variants (GAIN-S, MIWAE-S, TabCSDI-S), and two uncertainty-output models (GAIN-U, MIWAE-U). Each yields an estimated predictive distribution $\widehat{\mathbb{P}}(x_{ij}^{\text{mis}} \mid X^{\text{obs}}, M)$ for every missing cell—either empirically from multiple imputations or parametrically from model outputs—which we evaluate in terms of accuracy and calibration.

B. Benchmarking Data

We use five *numerical tabular* datasets spanning sizes and dimensionalities (Table I). All features are z-scored to stabilize training and ensure comparable error scales.

Dataset	#Rec.	#Attr.
housing	20,640	8
biodegradation	1055	41
cancer	569	30
energy	768	8
wine	178	13

TABLE I: Dataset statistics. All attributes are numerical.

All datasets used in this study are complete and therefore treated as ground truth. We adopt a semi-synthetic setup in which we start with fully observed data and inject missingness according to controlled MCAR, MAR, and MNAR mechanisms. Because the original datasets contain no missing values, we know the true values of every masked entry, enabling direct evaluation of both imputation accuracy and uncertainty calibration. Artificially injecting missingness also gives us full control over which features are affected, the conditions under which values become missing, and the overall missing rate, allowing systematic experimentation while preserving the real data distribution.

We injected missing values synthetically according to the three mechanisms (MCAR, MAR, MNAR). For each dataset, we fixed a target missingness rate (e.g., 10–15%) and masked the corresponding number of entries. Under MCAR, cells were sampled uniformly at random across all rows and columns, with each position selected only once. Under MAR, we defined a dependency condition (such as another feature exceeding its mean), assigned higher masking probabilities to rows satisfying the condition, and sampled cells according to these probabilities until the desired rate was reached. Under MNAR, we specified a condition on the variable itself and assigned higher masking probabilities to cells meeting that condition, producing biased patterns in which certain values were more likely to be removed. Because sampling is probabilistic in MAR and MNAR, the final missing rate may differ slightly from the target. The normalized dataset prior to masking serves as the ground truth, and the corrupted dataset is used as input to the imputation algorithms. Full implementation details are available in our repository [46].

C. Evaluation Measures

We report runtime, *accuracy* via MAE, and *calibration* via calibration curves and ECE. Let $S = \{(i,j) : M_{ij} = 0\}$ denote masked cells. With ground truth x_{ij} and imputed mean $\hat{x}_{ij}$,

$$MAE = \frac{1}{|S|} \sum_{(i,j) \in S} |x_{ij} - \hat{x}_{ij}|.$$

Lower MAE indicates better reconstruction on the same masked set. Calibration evaluates whether the model's stated confidence matches empirical accuracy. For a grid of nominal coverage levels $q \in \{0, 0.1, \ldots, 1.0\}$, where each q represents the confidence the model claims (e.g., $q = 0.9$ corresponds to a 90% prediction interval), we compute the model's q-level intervals and measure the fraction of true values they contain. Plotting empirical coverage (y–axis) against q (x–axis) yields
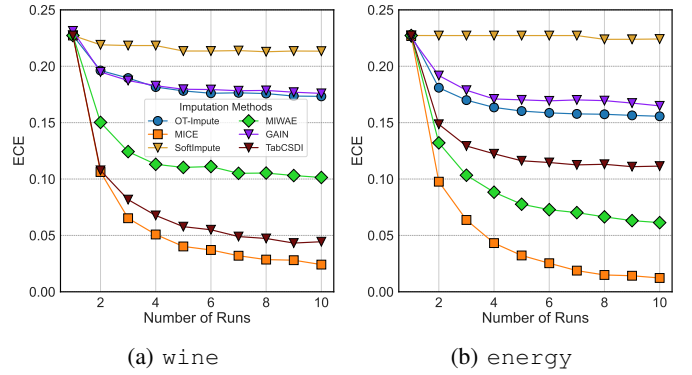

(a) wine   (b) energy

Fig. 1: ECE vs. n-runs at 30% MCAR.

the calibration curve: perfect calibration lies on $y = x$, while deviations indicate over- or under-confidence. We measure miscalibration with the expected calibration error

$$ECE = \frac{1}{|Q|} \sum_{q \in Q} |\text{Cov}(q) - q|,$$

where $\text{Cov}(q)$ is the observed coverage at level q. Lower values indicate better calibration. For continuous targets, we compute calibration using the CDF-based approach.

V. EXPERIMENTAL RESULTS

This chapter reports empirical results and analyzes uncertainty estimates across imputation methods. We first tune key parameters (Section V-A), then summarize runtime (Section V-B) and accuracy (Section V-C), and finally evaluate uncertainty and calibration (Section V-D). All code, datasets, and additional figures are available in our repository [46].

A. Tuning and Default Parameter Setting

We tune these hyperparameters to balance accuracy and runtime:

- MICE: 20–80 iterations depending on dataset size.
- OT-Impute: batch size 64–128; ∼300 iterations (small) to ∼500 (large).
- SoftImpute: shrinkage λ via CV on observed entries (log grid; grid_len= 15 small/medium, = 25 large).
- MIWAE: importance samples $K = 10$; 1500–2500 epochs.
- GAIN: 1500–2500 epochs; extra generator updates on high-d data for stable training.
- TabCSDI: epochs and reverse-diffusion steps capped for cost; e.g., wine: epochs= 500, num_steps= 600; energy: epochs= 400, num_steps= 1000.

Figure 1 shows that ECE steadily improves with increasing n-runs, but the gains plateau around five runs. Accordingly, we set n-runs = 5 as the default for all multi-run experiments. For sampling-based variants (Figure 2), both MIWAE-S and GAIN-S reach stable calibration by n-samples ≈ 20, while TabCSDI-S exhibits its best calibration performance between 50 to 70 samples. We therefore adopt n-samples=20 for MIWAE-S and GAIN-S, and n-samples=50 for TabCSDI-S.
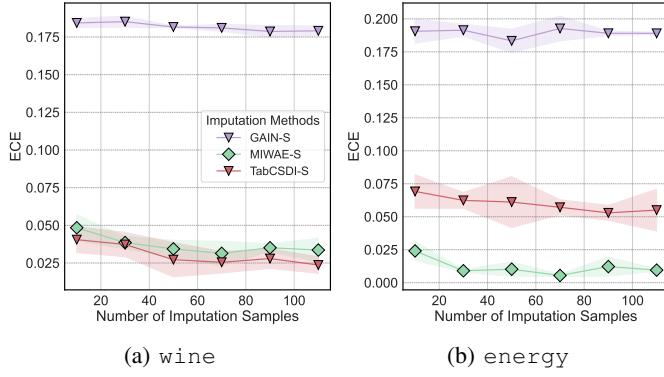
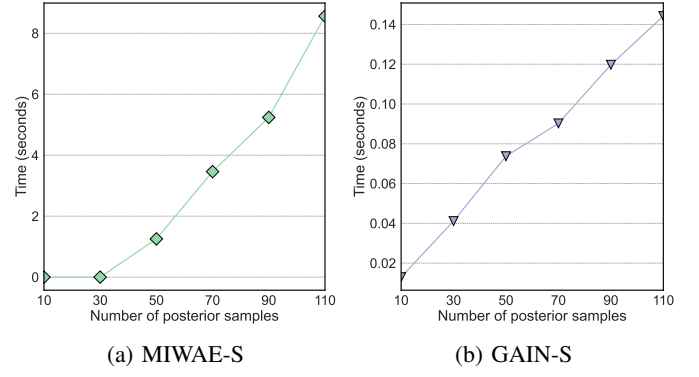(a) wine   (b) energy

Fig. 2: ECE vs. number of samples at 30% MCAR.

(a) MIWAE-S   (b) GAIN-S

Fig. 4: Runtime vs. `n-samples` at 30% MCAR in `wine`.
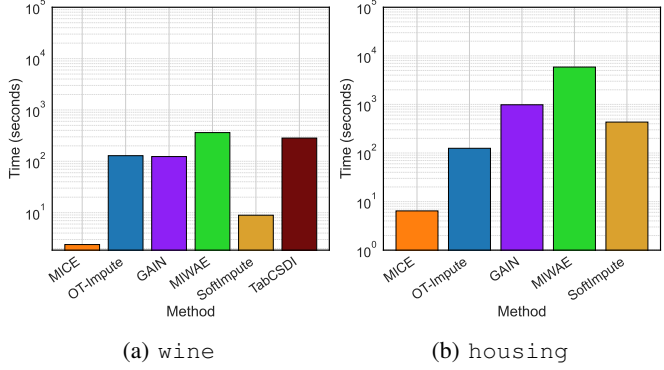

(a) wine   (b) housing

Fig. 3: Time per run at 30% missingness vs Classical methods report total time; deep models report train+single imputation.
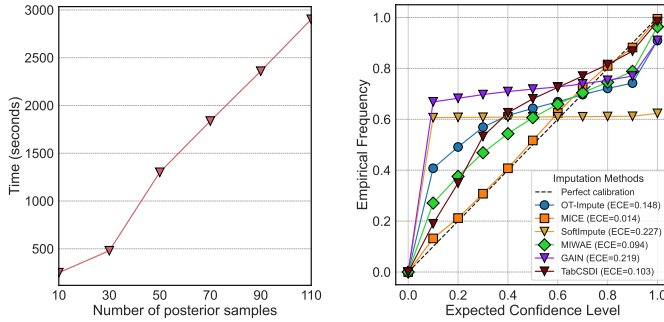

Fig. 5: Runtime vs. `n-samples` at 30% MCAR on `wine` for TabCSDI-S.

Fig. 6: Calibration curves for 30% MCAR on `biodegradation`.

B. Runtime per Method

Classical imputers such as MICE, OT-Impute, and Soft-Impute operate directly on masked data without training, so their total runtime corresponds to iterative convergence. Learning-based models (GAIN, MIWAE, TabCSDI) include both training and inference time.

MICE is generally faster than other methods in Figure 3a because it iteratively fits simple per-feature regressions on the currently observed data rather than training complex models or performing extensive sampling. However, in `housing`, the large number of attributes increases the per-iteration workload (more predictors per regression). OT-Impute involves repeated

Sinkhorn computations over many pairs and iterations, which can be moderately costly, yet it remains faster than deep learning and diffusion models since it avoids lengthy neural-network training and sampling. Its runtime also appears nearly constant across datasets because a fixed batch size and iteration count are used. SoftImpute is slower than MICE and can even be slower than OT-Impute and deep learning models on larger datasets due to repeated SVD operations. Among the generative models, GAIN is generally faster than MIWAE because it avoids multiple decoder evaluations per sample. Diffusion-based TabCSDI is the most computationally expensive method, as both training and inference require numerous sequential denoising steps.

For uncertainty estimation, running five independent imputations increases the total runtime roughly fivefold. The sampling-based variants MIWAE-S and GAIN-S are faster overall because they train once and generate multiple imputations from the same model; however, sampling increases inference time which makes `-S` slower than `-U`. Diffusion-based TabCSDI-S remains the most time-consuming due to its repeated reverse chains during sampling. The plots in Figure 4 and Figure 5 show runtime versus the number of posterior samples for the `S`-variant methods. As expected, taking more posterior samples increases runtime roughly linearly, since inference repeats the sampling loop more times.

C. Imputation Accuracy

Figure 7 shows the MAE across varying missingness rates for the `wine` and `energy` datasets. As expected, higher missingness consistently leads to higher MAE across all methods, because with fewer observed entries, models have less information to infer dependencies between attributes, reducing the accuracy of the reconstructed values.

At 30% missingness (Tables II–III), MAE generally increases from MCAR to MNAR. Exceptions appear in `housing`, where strong feature collinearity allows MAR/M-NAR imputations to outperform MCAR. MIWAE achieves the best accuracy because its multiple-sampling and importance-weighting approach allows it to better approximate the true data distribution and produce more reliable imputations. Another key observation is that for datasets like `wine` and

Dataset / Mechanism		MICE	OT-Impute	SoftImpute	MIWAE	GAIN	TabCSDI
wine	MCAR	0.784 ± 0.071	0.570 ± 0.004	0.603 ± 0.003	0.567 ± 0.016	0.788 ± 0.010	1.046 ± 0.033
	MAR	0.801 ± 0.073	0.628 ± 0.004	0.639 ± 0.016	0.582 ± 0.015	0.800 ± 0.010	1.103 ± 0.042
	MNAR	0.818 ± 0.084	0.655 ± 0.003	0.587 ± 0.001	0.621 ± 0.015	0.807 ± 0.014	1.076 ± 0.021
energy	MCAR	0.541 ± 0.044	0.553 ± 0.005	0.482 ± 0.009	0.431 ± 0.008	0.803 ± 0.011	0.958 ± 0.037
	MAR	0.579 ± 0.046	0.619 ± 0.006	0.472 ± 0.003	0.425 ± 0.010	0.789 ± 0.009	0.977 ± 0.032
	MNAR	0.657 ± 0.039	0.796 ± 0.006	0.543 ± 0.000	0.578 ± 0.011	0.868 ± 0.016	1.151 ± 0.047
housing	MCAR	0.779 ± 0.003	0.568 ± 0.001	0.472± 0.003	0.406 ± 0.005	0.654 ± 0.011	–
	MAR	0.763 ± 0.004	0.499 ± 0.001	0.403 ± 0.001	0.306 ± 0.003	0.546 ± 0.011	–
	MNAR	0.757 ± 0.001	0.553	0.482	0.391 ± 0.004	0.624 ± 0.027	–
biodegradation	MCAR	0.3307 ± 0.0002	0.3611 ± 0.0020	0.4249 ± 0.0174	0.5145 ± 0.0661	0.5439 ± 0.0044	0.9137 ± 0.1421
	MAR	0.3406 ± 0.0023	0.3798 ± 0.0021	0.4533 ± 0.0151	0.5286 ± 0.0625	0.5393 ± 0.0047	-
	MNAR	0.5243 ± 0.0090	0.6780 ± 0.0015	0.7377 ± 0.0264	0.5959 ± 0.0464	0.6606 ± 0.0118	-
cancer	MCAR	0.911 ± 0.002	0.691	0.607	0.497 ± 0.010	0.735 ± 0.016	–
	MAR	0.966 ± 0.002	0.705	0.601	0.506 ± 0.009	0.662 ± 0.011	–
	MNAR	0.964 ± 0.001	0.824	0.632	0.528 ± 0.014	0.817 ± 0.036	–

TABLE II: MAE at 30% missingness (mean ± std).

Dataset/Mech.		MIWAE-S	MIWAE-U	GAIN-S	GAIN-U	TabCSDI-S
wine	MCAR	0.568	0.587	0.773	0.797	0.805
	MAR	0.603	0.601	0.781	0.791	0.868
	MNAR	0.620	0.643	0.790	0.826	0.890
energy	MCAR	0.434	0.462	0.770	0.769	0.883
	MAR	0.434	0.443	0.771	0.742	0.919
	MNAR	0.585	0.608	0.867	0.797	1.107
housing	MCAR	0.412	0.404	0.660	0.603	–
	MAR	0.311	0.308	0.529	0.522	–
	MNAR	0.397	0.385	0.605	0.597	–
biodegradation	MCAR	0.4235	0.4135	0.5445	0.5136	0.6835
	MAR	0.4515	0.4319	0.5362	0.4983	–
	MNAR	0.7360	0.7421	0.6550	0.6092	–
cancer	MCAR	0.523	0.515	0.705	0.693	–
	MAR	0.537	0.501	0.649	0.688	–
	MNAR	0.515	0.541	0.845	0.730	–

TABLE III: MAE at 30% missingness for -S/-U variants.

(a) wine

(b) housing

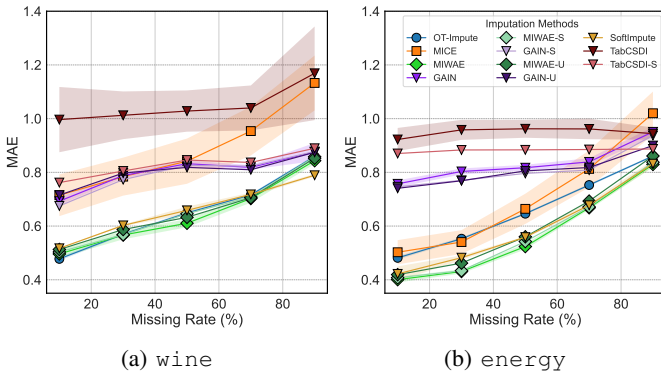Fig. 8: Calibration curves for 30% MCAR.

(a) wine

(b) energy

Fig. 7: MAE vs. missing rate for MCAR.

biodegradation the OT-Impute performs best due to its effective distribution-matching. SoftImpute often ranks second best overall, where its nuclear-norm regularization effectively captures latent structure.
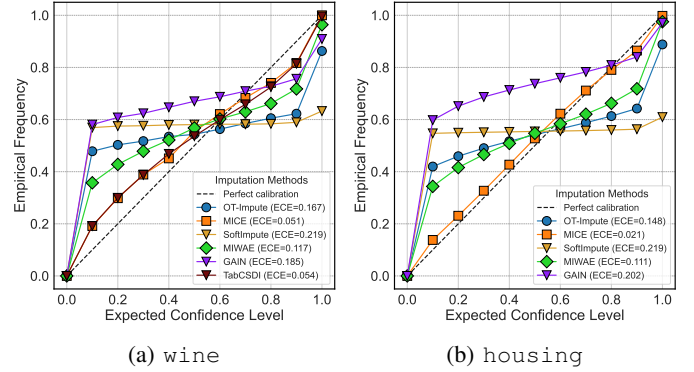
D. Calibration Curves and ECE

Figures 6 and 8 present calibration curves across imputation methods and datasets at a 30% MCAR. Overall, MICE achieves the most reliable calibration, producing curves near the ideal diagonal and the lowest ECE by incorporating realistic residual noise and averaging variability across multiple runs. Across datasets and mechanisms, SoftImpute is the least calibrated. As a deterministic low-rank method, it provides only point estimates without modeling uncertainty. Post-hoc proxies yield overly narrow, flat calibration curves, reflecting constant and unreliable uncertainty across confidence levels. Unlike MICE or generative models, OT-Impute lacks per-cell predictive distributions and instead aligns global feature distributions. This yields accurate point imputations but poorly sized uncertainty ranges, leading to higher ECE and mixed calibration—over-confident when transport mass is concentrated and under-confident when dispersed. Still, it calibrates slightly better than SoftImpute and GAIN due to modest stochasticity from mini-batching.

GAIN focuses on producing realistic imputations for the discriminator rather than calibrated uncertainty, leading to
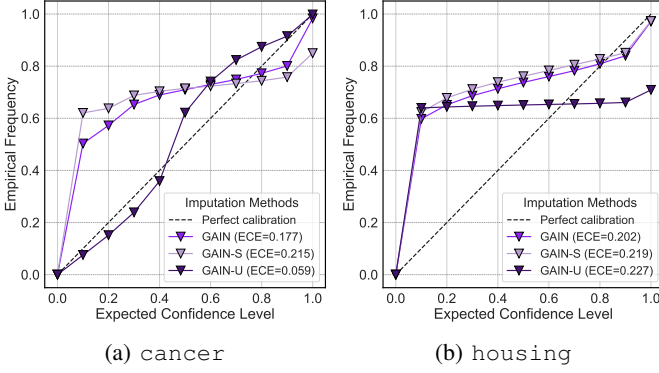
(a) cancer   (b) housing

Fig. 9: Calibration curves for GAIN family in 30% MCAR.

(a) wine   (b) energy

Fig. 11: Calibration curves for TabCSDI in 30% MCAR.

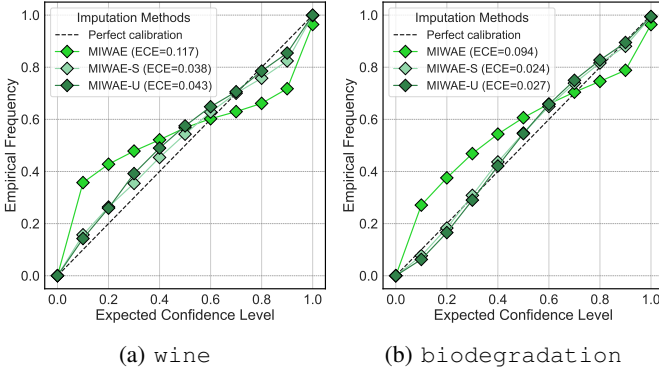(a) wine   (b) biodegradation

Fig. 10: Calibration curves for MIWAE family in 30% MCAR.

over-dispersed (under-confident) predictions and higher ECE (Figure 9). The GAIN-U variant introduces a variance head to estimate per-cell mean and variance, enabling multiple imputations and improving calibration in some datasets like cancer (Figure 9a), though the added complexity can cause unstable training and noisy variance estimates, as seen in housing (Figure 9b). The sampling-only GAIN-S variant, which generates multiple noisy samples without learning variance, performs slightly worse than GAIN because averaging samples tends to smooth out meaningful variability.

MIWAE is generally well-calibrated, it models each missing cell with a full predictive distribution (mean and variance) and averages multiple imputations at inference. This stochastic approach captures genuine variability in the data, preventing over-tight or overly diffuse predictions and thus lowering ECE. Its variants, MIWAE-S and MIWAE-U in Figure 10, further refine the predictive spread, MIWAE-S by adjusting variance to correct over/under-confidence, and MIWAE-U by directly using model-predicted uncertainty, resulting in confidence intervals that better match empirical coverage and occasionally outperform MICE (comparison between Figures 8a and 10a)

TabCSDI shows low ECE across all mechanisms in wine (Figure 11a) but shows mixed confidence in energy (Figure 11b). In wine, strong feature correlations and limited samples cause wide intervals, while in energy, varying feature difficulty leads to over-confidence on easy attributes and
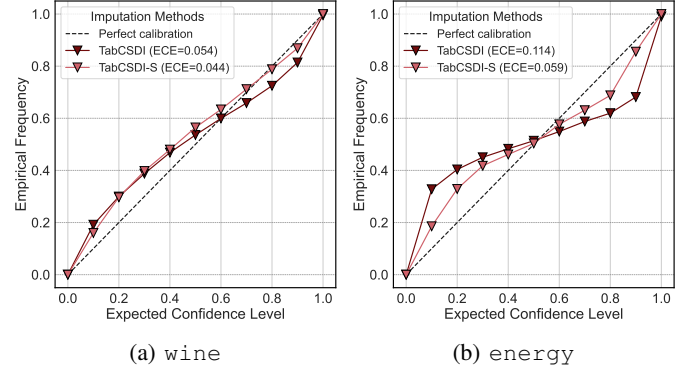
under-confidence on hard ones. The sampling-based TabCSDI-S further improves calibration by averaging multiple diffusion trajectories, producing more accurate uncertainty estimates and reducing both over- and under-confidence.

VI. ANALYSIS AND TAKEAWAYS

No single method dominates across datasets in accuracy, calibrated uncertainty, or runtime; the choice should weigh these trade-offs. Across mechanisms, we observe a consistent degradation from MCAR to MAR to MNAR in both accuracy and calibration, with a few dataset-specific exceptions. Methods also differ in run-to-run robustness: variance across seeds is generally modest for MICE/SoftImpute/MIWAE and higher for diffusion-based models (TabCSDI/TabCSDI-S), motivating the reporting of mean $\pm$ std over multiple runs. In our experiments, SoftImpute delivers strong accuracy across most datasets but remains consistently poorly calibrated, making it unsuitable when trustworthy uncertainty estimates are needed. MICE provides the most dependable calibration overall, though its point accuracy is generally lower than leading alternatives. MIWAE strikes a favorable middle ground, achieving both good accuracy and reasonably strong calibration, albeit with the highest computational cost. Therefore, the choice of method should depend on the main priority, whether accuracy, calibrated uncertainty, or runtime efficiency, and the characteristics of the dataset.

VII. CONCLUSION AND FUTURE WORK

This work evaluated uncertainty *calibration* in imputation rather than accuracy alone. Across six representative methods and multiple datasets, rates, and mechanisms, we quantified uncertainty using three complementary strategies: repeated runs, sampling from a trained model, and direct predictive distributions. Our results show that accuracy and calibration are distinct: methods with strong point error may be poorly calibrated, while methods that explicitly model uncertainty tend to yield more reliable coverage. In short, imputers should be judged not only by how close they get on average, but by whether their stated confidence matches observed frequencies.

There remains substantial room for future work. Our experiments focused on numeric tabular datasets, and an important

next step is to extend the framework to categorical and mixed-type data. Handling heterogeneous attributes requires different forms of uncertainty representation and calibration, and developing a unified approach for such data remains challenging. Because TabCSDI is computationally intensive, we were unable to evaluate it on all datasets; completing this analysis is part of future work. Another direction is to separate aleatoric and epistemic uncertainty and study how each behaves. Simulation can help distinguish them: epistemic uncertainty should decrease with more data or stronger models, while aleatoric uncertainty reflects inherent noise that persists. A final avenue is to examine how calibrated uncertainty interacts with downstream tasks such as fairness-sensitive prediction, risk assessment, and human-in-the-loop decision making. Integrating calibrated uncertainty into interactive workflows may improve reliability, interpretability, and decision support.

References

[1] R. R. Andridge and R. J. A. Little, "A review of hot deck imputation for survey non-response," *Int. Stat. Rev.*, vol. 78, no. 1, pp. 40–64, 2010.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. R. Stat. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.

[3] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, 1987.

[4] J. L. Schafer, *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, 1997.

[5] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *J. R. Stat. Soc. B*, vol. 61, no. 3, pp. 611–622, 1999.

[6] S. Oba, M.-a. Sato, I. Takemasa, M. Monden, K.-i. Matsubara, and S. Ishii, "A bayesian missing value estimation method for gene expression profile data," *Bioinformatics*, vol. 19, no. 16, pp. 2088–2096, 2003.

[7] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[8] G. E. A. P. A. Batista and M. C. Monard, "A study of k-nearest neighbour as an imputation method," *HST*, vol. 87, pp. 251–260, 2002.

[9] S. Van Buuren and K. Groothuis-Oudshoorn, "Mice: Multivariate imputation by chained equations in r," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011.

[10] D. J. Stekhoven and P. Bühlmann, "Missforest—non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

[11] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[12] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *JMLR*, vol. 11, pp. 2287–2322, 2010.

[13] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[14] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *NeurIPS*, pp. 2292–2300, 2013.

[15] L. Gondara and K. Wang, "Mida: Multiple imputation using denoising autoencoders," in *PAKDD*, pp. 260–272, 2018.

[16] O. Ivanov, M. Figurnov, and D. Vetrov, "Variational autoencoder with arbitrary conditioning," in *ICLR*, 2018.

[17] P.-A. Mattei and J. Frellsen, "Miwae: Deep generative modelling and imputation of incomplete data sets," in *ICML*, pp. 4413–4423, PMLR, 2019.

[18] J. Yoon, J. Jordon, and M. van der Schaar, "Gain: Missing data imputation using generative adversarial nets," in *ICML*, pp. 5689–5698, 2018.

[19] Y. Tashiro, J. Song, Y. Song, and S. Ermon, "Csdi: Conditional score-based diffusion models for probabilistic time series imputation," in *NeurIPS*, 2021.

[20] A. Kotelnikov, D. Baranchuk, A. Fenus, D. Vetrov, and S. Ivanov, "Tabddpm: Modelling tabular data with diffusion models," in *NeurIPS*, 2023.

[21] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Mach. Learn.*, vol. 110, no. 3, pp. 457–506, 2021.

[22] L. Wimmer, Y. Sale, P. Hofman, B. Bischl, and E. Hüllermeier, "Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures?," *Mach. Learn.*, vol. 112, pp. 2835–2867, 2023.

[23] F. Bickford Smith, J. Kossen, E. Trollope, M. Van Der Wilk, A. Foster, and T. Rainforth, "Rethinking aleatoric and epistemic uncertainty," in *ICML*, vol. 267 of *ICML*, pp. 4345–4359, 2025.

[24] A. Thawani, B. Ramsundar, and P. Baldi, "Characterizing uncertainty in machine learning for chemistry," *Acc. Chem. Res.*, vol. 56, no. 7, pp. 871–883, 2023.

[25] S. van Buuren, *Flexible Imputation of Missing Data*. Chapman & Hall/CRC, 2 ed., 2018.

[26] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1994.

[27] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML*, pp. 1050–1059, 2016.

[28] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *Foundations and Trends in Machine Learning*, vol. 16, no. 2, pp. 174–246, 2023.

[29] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *JASA*, vol. 102, no. 477, pp. 359–378, 2007.

[30] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[31] B. Muzellec, J. Josse, C. Boyer, and M. Cuturi, "Missing data imputation using optimal transport," in *ICML*, pp. 7130–7140, 2020.

[32] E. L. Lehmann and G. Casella, *Theory of Point Estimation*. Springer, 1998.

[33] G. Casella and R. L. Berger, *Statistical Inference*. Duxbury Press, 2002.

[34] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*. CRC Press, 3 ed., 2013.

[35] S. Hwang and D.-K. Chae, "An uncertainty-aware imputation framework for alleviating the sparsity problem in collaborative filtering," in *CIKM*, pp. 802–811, 2022.

[36] B. Roskams-Hieter, J. Wells, and S. Wade, "Leveraging variational autoencoders for multiple data imputation," in *ECML PKDD*, vol. 14169 of *LNCS*, pp. 491–506, Springer, 2023.

[37] J. Wang, Y. Zhang, K. Wang, X. Lin, and W. Zhang, "Missing data imputation with uncertainty-driven network," *SIGMOD*, vol. 2, no. 3, pp. 1–25, 2024.

[38] A. W. Mulyadi, E. Jun, and H.-I. Suk, "Uncertainty-aware variational-recurrent imputation network for clinical time series," *IEEE Trans. Cybern.*, vol. 52, no. 9, pp. 9684–9694, 2021.

[39] C. Cappiello, F. Cerutti, C. Sancricca, R. Zanelli, *et al.*, "About the effects of data imputation techniques on ml uncertainty.," in *VLDB Workshops*, 2023.

[40] M. H. Moslemi and M. Milani, "Threshold-independent fair matching through score calibration," in *GUIDE-AI @ SIGMOD*, 2024.

[41] A. Pirhadi, M. H. Moslemi, A. Cloninger, M. Milani, and B. Salimi, "Otclean: Data cleaning for conditional independence violations using optimal transport," *SIGMOD*, vol. 2, no. 3, p. 160, 2024.

[42] Z. Zheng, T. M. Quach, Z. Jin, F. Chiang, and M. Milani, "Currentclean: Interactive change exploration and cleaning of stale data," in *CIKM*, pp. 2917–2920, 2019.

[43] M. Milani, Z. Zheng, and F. Chiang, "Currentclean: Spatio-temporal cleaning of stale data," in *ICDE*, pp. 172–183, 2019.

[44] Y. Burda, R. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," in *ICLR*, 2016.

[45] S. Zheng and N. Charoenphakdee, "Diffusion models for missing value imputation in tabular data," in *TRL @ NeurIPS*, 2022.

[46] H. Zarin, "Imputation_uncertainty." https://github.com/ZarinTahia/Imputation_Uncertainty, 2025. GitHub repository; accessed August 2025.