ECAI 2025
I. Lynce et al. (Eds.)

doi:10.3233/FAIA251120

HI-PMK: A Data-Dependent Kernel for Incomplete Heterogeneous Data Representation

Youran Zhou,[*], Mohamed Reda Bouadjenek, Jonathan Wells and Sunil Aryal

School of Information Technology, Deakin University, Geelong, Australia

Abstract. Handling incomplete and heterogeneous data remains a central challenge in real-world machine learning, where missing values may follow complex mechanisms (MCAR, MAR, MNAR) and features can be of mixed types (numerical and categorical). Existing methods often rely on imputation, which may introduce bias or privacy risks, or fail to jointly address data heterogeneity and structured missingness. We propose the **H**eterogeneous **I**ncomplete **P**robability **M**ass **K**ernel (**HI-PMK**), a novel data-dependent representation learning approach that eliminates the need for imputation. HI-PMK introduces two key innovations: (1) a probability mass-based dissimilarity measure that adapts to local data distributions across heterogeneous features (numerical, ordinal, nominal), and (2) a missingness-aware uncertainty strategy (MaxU) that conservatively handles all three missingness mechanisms by assigning maximal plausible dissimilarity to unobserved entries. Our approach is privacy-preserving, scalable, and readily applicable to downstream tasks such as classification and clustering. Extensive experiments on over 15 benchmark datasets demonstrate that HI-PMK consistently outperforms traditional imputation-based pipelines and kernel methods across a wide range of missing data settings. Code is available at: github.com/echoid/Incomplete-Heter-Kernel

1 Introduction

Missing data is a common challenge in real-world, data-driven applications, caused by factors such as data collection errors, survey non-responses, or system malfunctions [3]. These missing values can occur across all data types—numerical, categorical, or heterogeneous—making data analysis more complex and degrading machine learning model performance, often leading to biased or sub-optimal outcomes. Existing methods for handling missing data fall broadly into two categories: imputation-based approaches and representation learning approaches (see Figure 1).

Imputation-based methods estimate missing values using observed data, creating imputed complete datasets for downstream analysis. These methods are often preferred due to their intuitive evaluation process, where imputed data can be directly compared to the original complete dataset. However, they face two significant limitations: (i) revealing original data during imputation raises privacy concerns, and (ii) access to complete datasets is often impractical, necessitating reliance on indirect downstream metrics, which can obscure the evaluation of imputation quality.

Representation learning methods, in contrast, bypass imputation entirely by learning robust representations directly from incomplete data for downstream tasks. This approach mitigates privacy risks, reduces computational overhead, and eliminates dependence on complete datasets for evaluation. By focusing on meaningful representations, these methods provide a secure, efficient, and adaptable framework for handling incomplete datasets.

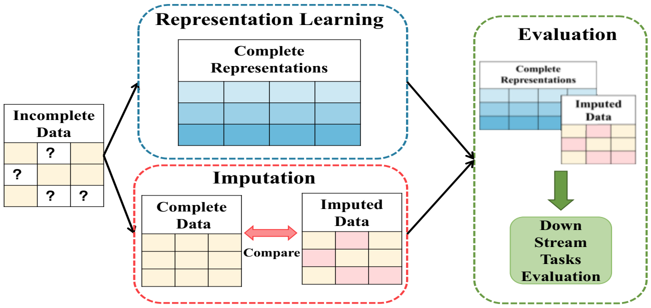


Figure 1. Comparing imputation and representation learning approaches for incomplete data handling.

Another major limitation in existing methods is their reliance on assumptions about data types and missing mechanisms. Most traditional techniques are designed for numerical data [13, 39, 9] and extend to heterogeneous datasets through preprocessing steps like one-hot encoding, which increase dimensionality and computational cost [1] . Furthermore, while many methods assume data is Missing Completely at Random (MCAR), real-world data often follows more complex patterns, such as Missing Not at Random (MNAR) or Missing at Random (MAR) [29]. Failing to address these mechanisms can result in biased outcomes [24].

Our contribution: We propose **H**eterogeneous **I**ncomplete **P**robability **M**ass **K**ernel (**HI-PMK**)—a novel, data-dependent kernel method that computes pairwise similarities directly from the observed portions of the data. HI-PMK supports both numerical and categorical features, models uncertainty without imputation, and accommodates arbitrary missing data mechanisms (MCAR, MAR, MNAR). Our approach provides a privacy-preserving, representation learning framework that is model-agnostic and readily applicable to downstream tasks such as classification and clustering. Extensive experiments on over 15 benchmark datasets with varying missing rates and mechanisms show that HI-PMK consistently outperforms both imputation-based pipelines and kernel baselines. The proposed method is simple, scalable, and robust—offering a practical alternative to conventional approaches for learning from incomplete heterogeneous data.

[*] Corresponding Author. Email: echo.zhou@deakin.edu.au

2 Related Works

Imputation Methods for Incomplete Data. Conventional techniques—such as mean substitution, MICE, k-NN, and matrix factorization [9, 39, 16, 11]—are simple but struggle with heterogeneous features and often distort downstream distributions. Generative approaches based on VAEs [21, 15, 25], GANs [40, 2], or diffusion models [37, 8, 46] model complex uncertainty, but typically assume homogeneous data, require large training samples, and are not task-aware. Recent empirical studies [45] show that such models often underperform on small tabular datasets. Surveys [22, 36, 44] confirm these limitations, particularly in handling mixed-type features and diverse missingness mechanisms. These challenges motivate imputation-free strategies like representation learning.

Representation Learning for Incomplete Data. Instead of filling in missing values, representation learning directly encodes incomplete inputs into task-specific embeddings. Popular approaches include autoencoders [7] for reconstruction and graph neural networks [43, 7, 20] for propagating partial observations. These methods excel in complex domains, e.g., multimodal [18] adb temporal where structural dependencies aid representation. However, in tabular datasets with mixed types, unstructured missingness, and limited samples, deep models can overfit or generalize poorly. In such scenarios, lightweight methods like probabilistic modeling [16] and similarity-based estimators [28] often remain competitive.

Kernel-Based Approaches for Incomplete Data. Kernel methods offer a non-parametric alternative by modeling similarities in latent space without requiring complete input. Early variants adapt Gaussian or polynomial kernels through observed-feature reweighting or regularization [34, 26, 6, 32, 33], but typically assume numerical inputs and neglect missingness structure. More recent work estimates similarity directly from partial data, including affinity learning with kernel correction [41], low-rank matrix recovery [42]. Although effective in spectral clustering or matrix completion, these methods often rely on scaling, kernel tuning, and assume real-valued inputs that limiting their application in mixed-type tabular settings. Our HI-PMK fills this gap by enabling type and uncertainty-aware similarity estimation without such assumptions.

3 Problem Formulations

Let $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ denote a dataset with m instances and n features, potentially containing missing values. We decompose $\boldsymbol{X}$ into an observed component $\boldsymbol{X}^{\mathbf{o}}$ and a missing component $\boldsymbol{X}^{\mathbf{m}}$, with a binary mask $\boldsymbol{M} \in \{0,1\}^{m\times n}$ indicating the missingness pattern: $\boldsymbol{M}_{ij} = 1$ if X_{ij} is missing, and 0 otherwise. The missingness process is governed by latent parameters Ψ, which encode factors that influence whether an entry is observed or missing. This relationship can be formalized as a conditional distribution:

$$f(\boldsymbol{M} \mid \boldsymbol{X}, \Psi),$$

where $\boldsymbol{M}$ is potentially dependent on both observed and unobserved parts of $\boldsymbol{X}$, depending on the missingness mechanism. The missing mechanism, determined by Ψ, is typically categorized into the following types:

Missing Completely at Random (MCAR). The probability of a value being missing is entirely independent of both observed ($\boldsymbol{X}^{\mathbf{o}}$) and missing ($\boldsymbol{X}^{\mathbf{m}}$) data. The missingness depends only on Ψ and can be expressed as:

$$f(\boldsymbol{M} \mid \Psi)\forall\, \boldsymbol{X}.$$

For example, in a healthcare dataset, some patients may miss follow-up medical tests due to random scheduling conflicts. Here, the observed part ($\boldsymbol{X}^{\mathbf{o}}$) includes attributes like age and previous test results, while the missing part ($\boldsymbol{X}^{\mathbf{m}}$) comprises the unrecorded follow-up test results.

Missing At Random (MAR). The probability of missingness depends only on the observed data $\boldsymbol{X}^{\mathbf{o}}$. This can be expressed as:

$$f(\boldsymbol{M} \mid \boldsymbol{X}^{\mathbf{o}}, \Psi)\forall\, \boldsymbol{X}^{\mathbf{m}}.$$

For instance, in the same healthcare dataset, older patients may be less likely to attend follow-up medical tests. Here, the missingness depends on the observed age ($\boldsymbol{X}^{\mathbf{o}}$) but is unrelated to the actual test results ($\boldsymbol{X}^{\mathbf{m}}$).

Missing Not At Random (MNAR). The probability of missingness directly depends on the values of the missing data $\boldsymbol{X}^{\mathbf{m}}$. This is expressed as:

$$f(\boldsymbol{M} \mid \boldsymbol{X}^{\mathbf{m}}, \Psi)\forall\, \boldsymbol{X}^{\mathbf{o}}.$$

For example, in the same healthcare dataset, patients with very poor test results might avoid follow-up visits due to fear or reluctance. In this case, the missingness depends directly on the test results ($\boldsymbol{X}^{\mathbf{m}}$), making it an MNAR mechanism.

4 Data-Dependent Kernel

Data-dependent kernels enhance similarity computation by incorporating local data distributions into pairwise comparisons [4, 5, 38, 49], enabling better handling of heterogeneous features and capturing nuanced relationships. However, existing methods are not well equipped to operate under incomplete data. To address this, we propose the **H**eterogeneous **I**ncomplete **P**robability **M**ass **K**ernel (HI-PMK), a novel data-dependent kernel designed to handle both heterogeneity and missingness. In HI-PMK, the **H-component** models data type diversity, while the **I-component** captures missingness-aware uncertainty, enabling robust and flexible similarity estimation.

4.1 Probability Mass Kernel (PMK)

The Probability Mass Kernel (PMK) builds upon the concept of m_0-dissimilarity [4, 27], extending traditional distance metrics by incorporating the local data distribution surrounding the objects being compared. Unlike data-independent kernels, such as Gaussian or Laplacian kernels—where the similarity between two points depends solely on their spatial distance—PMK adjusts similarity based on the density of data in the surrounding region. For example, in sparse regions, two points at the same distance may be considered more similar than in densely populated regions. This adaptability allows PMK to capture complex structures and patterns in data, resulting in more accurate similarity measures.

Let $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ be a dataset with m instances and n features. The i-th instance can be denoted as a vector $\mathbf{x}^{(i)} = \langle x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}\rangle$, where $i \in [1, m]$. For a given feature k, let $R_k(x_k^{(i)}, x_k^{(j)})$ represent the region covering instances $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ for feature k. Specifically, the size of this region, $|R_k(x_k^{(i)}, x_k^{(j)})|$, represents the number of data points whose k-th feature values lie within this range. This count quantifies local data density, enhancing the similarity measure by incorporating contextual information.

The m_0-dissimilarity between $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ is calculated as:

$$m_0(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \left(\frac{1}{n}\sum_{k=1}^{n} \log\frac{|R_k(x_k^{(i)}, x_k^{(j)})|}{m}\right) \quad (1)$$

Intuitively, if many data points fall within the region, the two instances are considered more dissimilar, as a dense data distribution surrounds them. Conversely, sparse regions yield small dissimilarity scores.

To efficiently compute $|R_k(x_k^{(i)}, x_k^{(j)})|$, each feature k is discretised into b bins, and a pre-computed bin data mass is used. A matrix stores the data masses between all bin pairs for each feature, enabling fast lookups to determine the region size and significantly speeding up the computation process. Finally, the m_0 score is normalized to obtain the PMK:

$$\text{PMK}(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \frac{2 \cdot m_0(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})}{m_0(\mathbf{x}^{(i)}, \mathbf{x}^{(i)}) + m_0(\mathbf{x}^{(j)}, \mathbf{x}^{(j)})}. \tag{2}$$

This normalization ensures symmetry and self-similarity, aligning PMK with Mercer kernel requirements and enabling its integration with kernel-based learning frameworks.

4.2 PMK for Heterogeneous Data

The original PMK was limited to numerical data. To handle mixed data types, it was extended with the **H-Component**, enabling similarities across numerical and categorical features. This enhanced framework is called H-PMK.

In Equation 1, $\frac{|R_k(x_k^{(i)}, x_k^{(j)})|}{m}$ is the probability mass in the region, denoted as $P(R_k(x_k^{(i)}, x_k^{(j)}))$, which is calculated by discretizing the values of the numerical features. For categorical features, $P(R_k(x_k^{(i)}, x_k^{(j)}))$ can be computed using probabilities of categorical labels. The computation depends on whether the categorical feature k is **ordinal** (where values follow a natural order, such as size = S, M, L, XL, XXL}) or **nominal** (where values have no inherent order, such as color = Red, Green, Blue, Yellow, White}), as outlined below:

$$P(R_k(x_k^{(i)}, x_k^{(j)})) = \begin{cases} \displaystyle\sum_{z_k=\min(x_k^{(i)}, x_k^{(j)})}^{\max(x_k^{(i)}, x_k^{(j)})} P(z_k), & \text{ordinal } k \\ P(x_k^{(i)} \vee x_k^{(j)}), & \text{nominal } k \end{cases} \tag{3}$$

where $P(z_k)$ represents the frequency of the feature value z_k divided by the total number of instances m, and $P(x_k^{(i)} \vee x_k^{(j)})$ denotes the probability of a feature k having the label of $x_k^{(i)}$ or $x_k^{(j)}$.

4.3 PMK for Incomplete Data

The **I-Component** extends PMK to handle incomplete data by addressing missing values under various mechanisms. This enhancement, called I-PMK, ensures robust performance on incomplete datasets.

Separated Bucket $\mathcal{B}_k$ for Missingness Frequency. Since PMK relies on probability mass computed from the data distribution, we treat missing values as a distinct category. For each feature k, we introduce a special bucket $\mathcal{B}_k$ that collects all missing entries in that feature. The size of this bucket, denoted $|\mathcal{B}_k|$, reflects the empirical frequency of missingness and is later used to adjust pairwise similarity computations involving missing values.

Adjusting Probability Mass Estimation for Incomplete Entries (Maximizing Uncertainty). To compute $P(R_k(x_k^{(i)}, x_k^{(j)}))$ when one or both values are missing, we adopt a conservative strategy that reflects the maximum plausible dissimilarity under uncertainty. Specifically, we approximate the probability mass using the largest possible region consistent with the observed data distribution. This design aligns with the principle of *Maximizing Uncertainty (MaxU)*, which treats missing values as potentially taking any value in the feature space.

Only one of them is missing: For any numeric or ordinal feature k, we treat them similarly since numerical data is converted into ordinal through discretization. In this context, x_k represents the observed value in the pair we are analyzing, while '*?*' denotes the missing value. The region size $|R_k(x_k, ?)|$ can be estimated by examining the data masses between the bin containing x_k ($Bin(x_k)$) and the bins representing the first (minimum value) and last (maximum value) in the feature range. We let $\mathcal{M}_L(x_k)$ and $\mathcal{M}_R(x_k)$ represent the data masses to the left and right of $Bin(x_k)$, respectively, including the mass of $Bin(x_k)$ itself (as shown in Figure 2). Since '*?*' denotes a missing value and can correspond to any possible value, we take a conservative approach named Maximize Uncertainty (MU) and assign the maximal possible dissimilarity as follows:

$$|R_k(x_k, ?)| = \max(\mathcal{M}_L(x_k), \mathcal{M}_R(x_k)) + |\mathcal{B}_k| \tag{4}$$

If k is a nominal feature, $|R_k(x_k, ?)|$ is computed based on the frequencies of nominal labels:

$$|R_k(x_k, ?)| = \mathcal{M}(x_k) + \max_{a \in \mathcal{S}_k} \mathcal{M}(a) + |\mathcal{B}_k| \tag{5}$$

Here, $\mathcal{M}(x_k)$ is the frequency of the observed label x_k, and $\mathcal{S}_k$ represents the set of all possible nominal labels for feature k. The term $\max_{a \in \mathcal{S}_k} \mathcal{M}(a)$ accounts for the most frequent label, assuming '*?*' could correspond to it, resulting in maximal dissimilarity. In both cases, we include the frequency of missing values ($|\mathcal{B}_k|$) to ensure that the contribution of missing data is properly accounted for, reflecting their potential to take on any value.

Both of them are missing: When both entries are missing, we adopt a conservative strategy to approximate their maximum possible dissimilarity. For numerical or ordinal features, the worst-case assumption is that the two missing values differ maximally, leading to:

$$|R_k(?, ?)| = m, \tag{6}$$

where m denotes the total number of instances in the dataset. For nominal features, since the missing entries may belong to two different categories, we estimate the maximal disagreement by summing the frequencies of the two most common categories:

$$|R_k(?, ?)| = \max_{a \in \mathcal{S}_k} \mathcal{M}(a) + \max_{b \in \mathcal{S}_k \setminus \{a\}} \mathcal{M}(b) + |\mathcal{B}_k|, \tag{7}$$

where $\mathcal{S}_k$ is the set of possible categories for feature k, $\mathcal{M}(a)$ denotes the frequency of label a, and $|\mathcal{B}_k|$ is the number of missing entries in feature k. This formulation ensures that completely missing pairs are penalized with maximal dissimilarity, while preserving consistency across feature types.

4.4 Methodology Analysis

Separate Bucket $\mathcal{B}_k$. The bucket $\mathcal{B}_k$ stores all missing entries for feature k and plays a key role in capturing implicit patterns under structured missingness. Under MCAR, where missingness occurs

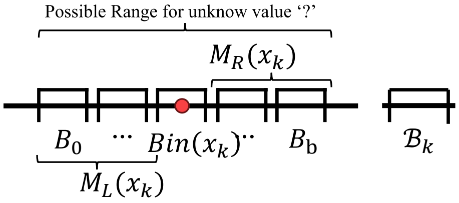
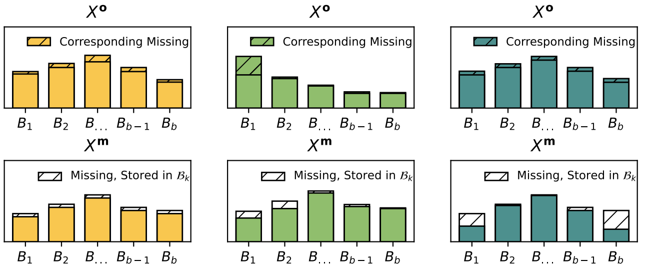


Figure 2. Illustration of probability mass adjustment for missing values in numeric or ordinal feature k. The red dot marks the observed value x_k; B_0 and B_b represent the first and last bins, respectively; and $\mathcal{B}_k$ denotes the designated bin for missing values.



(a) MCAR (b) MAR (c) MNAR

Figure 3. Visualization of mass bucket distributions under different missingness mechanisms. In $\boldsymbol{X}^{\mathbf{o}}$, the term *Corresponding Missing* refers to observed values associated with missing entries in $\boldsymbol{X}^{\mathbf{m}}$. In $\boldsymbol{X}^{\mathbf{m}}$, values are grouped and stored into buckets $\mathcal{B}_k$, shows the distribution of missingness.

uniformly at random, $\mathcal{B}_k$ has limited effect due to the lack of informative structure. However, under MAR and MNAR, $\mathcal{B}_k$ becomes highly informative. In MAR settings, missingness depends on observed values ($\boldsymbol{X}^{\mathbf{o}}$), which are often correlated with unobserved values ($\boldsymbol{X}^{\mathbf{m}}$). As shown in Figure 3 (b), $\mathcal{B}_k$ implicitly captures this correlation structure by accumulating contextually similar missing entries. In MNAR cases (Figure 3 (c)), missingness depends directly on the missing values themselves. Since these unobserved entries tend to concentrate around specific value ranges or categories, $\mathcal{B}_k$ acts as a proxy for these latent patterns. Thus, the inclusion of $\mathcal{B}_k$ significantly enhances robustness under non-random missing mechanisms.

Maximizing Uncertainty (MaxU). The uncertainty-aware term is grounded in the *principle of maximum entropy* [17], which assigns the largest plausible dissimilarity when a value is missing. MaxU conservatively models the missing entry as potentially any value in the feature domain $\mathcal{X}_k$, leading to the worst-case dissimilarity: $|R_k(x_k, ?)| = \sup_{x' \in \mathcal{X}_k} |R_k(x_k, x')|$. This strategy implicitly assumes a uniform distribution over all plausible values, avoiding arbitrary assumptions and improving robustness under adversarial or structured missingness. Compared to alternative schemes like *Average Uncertainty* (AvgU) and *Minimum Uncertainty* (MinU), MaxU aligns with the minimax principle [14], preserving discriminative power by maximizing entropy. While AvgU assumes a mean-case estimate and MinU favors optimistic matches, both are prone to failure under high sparsity or biased missing patterns. MaxU avoids such pitfalls, reducing imputation bias and maintaining kernel expressiveness across diverse missing mechanisms.

4.5 *Limitation and Complexity Analysis*

The original M_0 similarity [4] computes dissimilarity via bin-based histograms, resulting in a preprocessing time complexity of $O(mnb + nb^2)$, where m is the number of instances, n the number of features, and b the number of bins. This scaling becomes prohibitive for large or high-dimensional datasets. By contrast, our proposed HI-PMK avoids binning and instead uses precomputed probability masses based on the observed data distribution. The resulting similarity matrix can be computed in $O(m^2 \cdot n)$ time and stored in $O(m^2)$ space—matching the complexity of standard similarity measures such as Euclidean or cosine distance. This efficiency enables HI-PMK to scale to moderate-sized datasets while supporting mixed feature types and structured missingness. A detailed runtime analysis is provided in Section 5.3.

5 Experimental Results

We evaluate HI-PMK on both clustering and classification tasks. While PMK was initially designed for clustering, our results show that HI-PMK generalizes well to classification, highlighting its robustness across downstream applications. Additional experiments are available in the Supplementary Material [47]. **Code:** github.com/echoid/Incomplete-Heter-Kernel

5.1 *Experimental Setup*

5.1.1 *Datasets.*

We evaluate two types of datasets from the UCI Machine Learning Repository, as summarized in Table 1. **(i) Real-World Datasets with Synthetic Missingness:** We introduce MCAR, MAR, and MNAR patterns at varying missing rates into 10 fully observed datasets to systematically evaluate robustness under controlled conditions. **(ii) Real-World Incomplete Datasets:** Six naturally incomplete datasets with unknown mechanisms (possibly a mixture of mechanisms), reflecting practical challenges in real-world settings.

Dataset	**N**	**Ord**	**Nom**	**Num**	**C**	**M.R.**
			Complete Datasets			
Adult	48,842	1	7	6	4	-
Australian	69,014	0	8	6	2	-
Banknote	1,372	0	0	5	2	-
Breast	2,869	4	5	0	2	-
Car	1,728	6	0	0	4	-
Heart	303	0	8	5	5	-
Sonar	208	0	0	60	2	-
Spam	4,601	0	0	57	2	-
Student	649	11	16	2	5	-
Wine	4,898	0	0	12	2	-
			Incomplete Datasets			
Hepat	155	0	13	6	2	5.67%
Horse	368	13	1	8	2	23.80%
Kidney	400	2	10	12	2	10.54%
Mammo	961	4	0	1	2	3.37%
Pima	768	0	0	8	2	12.24%
Wiscon	699	9	0	0	2	0.25%

Table 1. Summary of datasets. **N** = #Instances, **Ord/Nom/Num** = #Ordinal/Nominal/Numerical Features, **C** = #Classes, **M.R.** = Missing Rate.

Synthetic Incomplete Data Generation. We introduced missingness into complete datasets under three standard mechanisms. **MCAR** removes values uniformly at random, independent of features or values. **MAR** introduces missingness based on observed features, following the strategy in [23]. **MNAR** depends on unobserved values: for numerical features, extreme values (e.g., high/low percentiles) are more likely to be missing; for ordinal features, boundary categories have higher missing rates; for nominal features, certain categories are assigned greater missing probabilities.

5.1.2 *Benchmark Methods for Comparison*

Imputation-Based Methods. We include representative imputation techniques: `Mean/Mode Imputer` as a simple baseline,

Model	Hepat		Horse		Kidney		Mammo		Pima		Wiscon	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
Mean	0.8129	0.0015	0.8398	0.0078	0.9775	0.0067	0.8148	0.0959	**0.7735**	0.0111	0.9628	0.7295
MICE	0.8129	0.0021	**0.8506**	0.0024	0.9700	0.0074	**0.8241**	0.0959	0.7722	0.0910	0.9614	0.7427
EM	0.8000	0.0025	0.8398	0.0079	0.9500	0.0045	0.8200	0.0911	0.7696	0.0211	0.9614	0.7387
MisF	0.8065	0.0014	0.8262	0.0078	0.9650	0.0023	0.8200	0.0959	0.7722	0.0169	0.9628	0.7335
GAIN	**0.8274**	0.0012	0.8501	0.0071	0.8525	0.0160	0.8106	0.0932	0.7462	0.0596	0.9642	0.7387
genRBF	0.7935	0.0772	0.6304	0.0166	0.6250	0.0132	0.5140	0.0002	0.6510	0.0052	0.5564	0.4505
KPCA	0.7935	0.0159	0.6850	0.0583	0.6250	0.0056	0.8127	0.0117	0.6510	0.0092	0.9500	0.5186
PPCA	0.8000	0.0015	**0.8506**	0.0540	0.9625	0.0081	**0.8241**	0.0959	0.7722	0.0909	0.9628	0.7427
Simp	-	0.0019	-	0.0001	-	0.0424	-	0.0951	-	0.0556	-	0.0000
Gow	-	0.1968	-	**0.1280**	-	0.3899	-	0.2970	-	0.1069	-	0.6939
HI-PMK	0.8065	**0.2001**	**0.8506**	0.1066	**0.9875**	**0.6663**	**0.8241**	**0.3271**	**0.7735**	**0.1374**	**0.9700**	**0.7585**

Table 2. NMI scores for clustering tasks on incomplete datasets and Classification accuracy for incomplete datasets. The bold values indicate the best-performing models, while the underlined values represent the second-best performance for each dataset.

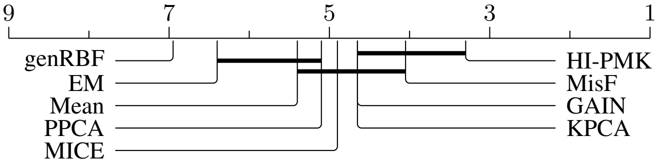


Figure 4. Critical Difference (CD) Diagram showing the ranking of models based on their F1 scores across all datasets. Smaller ranking value represent better performance.

MICE [39] for its iterative inference, EM [9, 19] for likelihood-based estimation, and MisF [35], a tree-based approach tailored for mixed-type data. Although recent studies [10, 31, 45] question the efficacy of deep models on tabular data, we also evaluate GAIN [40], a generative method representative of deep learning-based imputers.

Kernel Representation Methods. We assess genRBF [33], which directly models similarity in the presence of missing data, and two classical kernel learning methods—KPCA [30, 48] and PPCA [12]—which rely on prior imputation (via MICE) and highlight the limitations of conventional kernels when applied to heterogeneous or incomplete datasets.

Clustering Methods. We include Simp and Gow [27], two similarity-based clustering methods designed for heterogeneous data. To enable clustering on incomplete datasets, we first apply MICE imputation to ensure complete inputs.

5.1.3 Evaluation Metrics

For clustering, we applied K-means with the number of clusters K set to the ground-truth number of classes. Performance was evaluated using Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI), where ARI accounts for chance agreement. Each experiment was repeated five times, and results were averaged for stability.

For classification, we used Support Vector Machines (SVM) with RBF kernels. On complete datasets, we used standard RBF kernels; for incomplete binary-class datasets, RBF kernels were computed over similarity matrices. Methods like genRBF and HI-PMK, which rely on pairwise similarities, were evaluated using SVMs with precomputed kernels. We performed 5-fold cross-validation, tuning hyperparameters (e.g., C, kernel width) via nested inner 5-fold CV. Final models were trained on the full training set and evaluated using F1 score and accuracy.

5.2 Analysis of Results

Incomplete Data with Unknown Missing Mechanism. Table 2 presents classification accuracy and clustering NMI on real-world incomplete datasets. As expected, Simp and Gow, being clustering-only methods, are not applicable to classification. HI-PMK consistently achieves the best or near-best performance, demonstrating its robustness under unknown and mixed missing mechanisms. Although these datasets contain relatively low missing rates, HI-PMK maintains strong results without relying on imputation or prior knowledge of the missingness pattern.

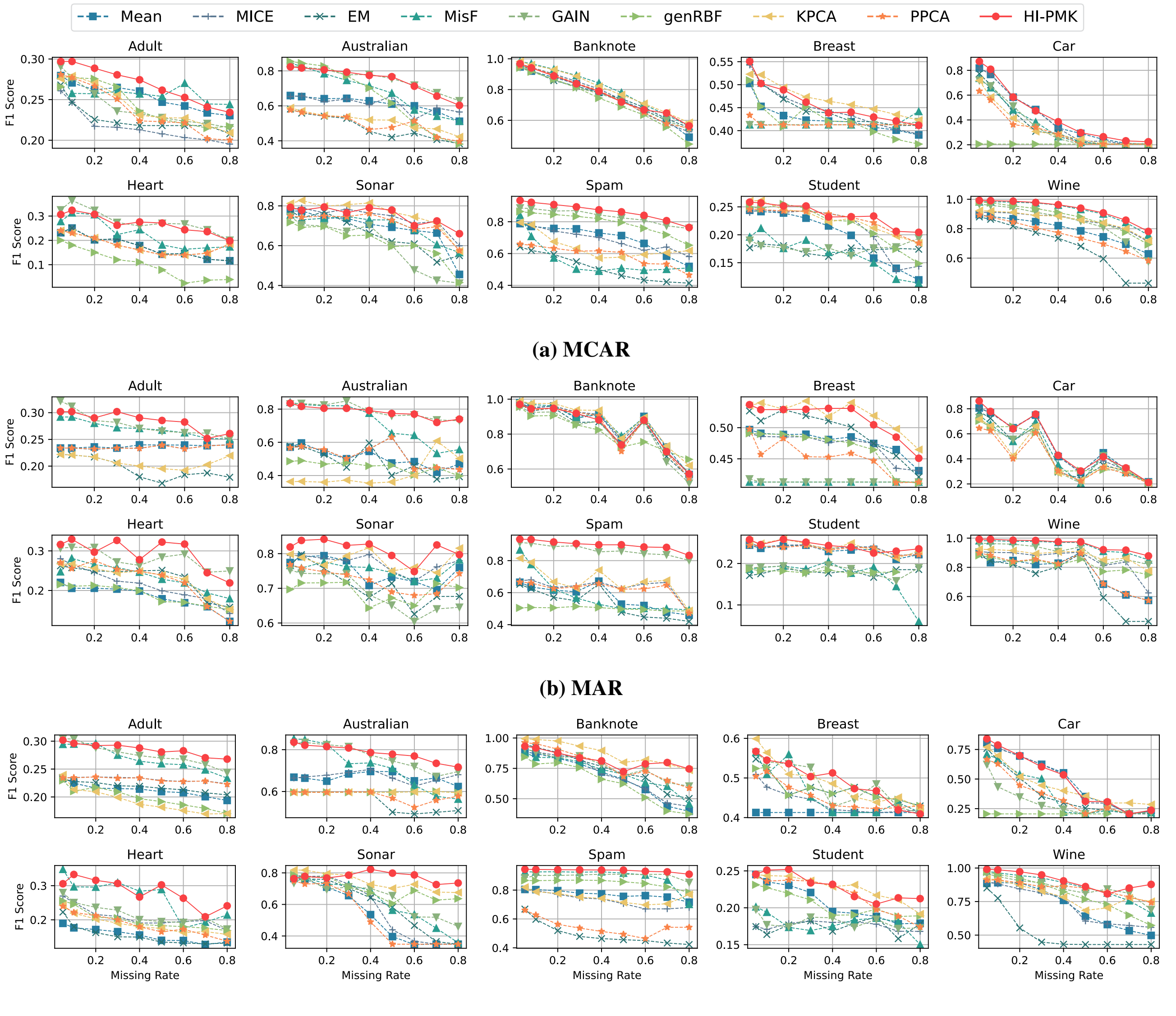
Complete Data with Synthetic Missing Mechanism. Figure 5 reports F1 scores across varying missing rates and mechanisms, with Table 3 showing detailed results at 20% missingness. HI-PMK consistently ranks among the top performers under all mechanisms. It is particularly effective on both numerical datasets and mixed-type datasets, confirming its adaptability. Unlike most baselines, its performance remains stable even as missingness increases. HI-PMK also generalizes well to high-dimensional datasets like *Sonar* and *Spam*. The Critical Difference (CD) diagram in Figure 4 ranks methods based on F1 scores, with HI-PMK achieving the top rank overall, followed by MisF and GAIN.

5.3 Scalability

We assess the runtime scalability of HI-PMK under varying sample sizes, feature dimensions, and missing rates using synthetic datasets. The evaluation settings are as follows:

- **Sample Size:** n=30 features, missing rate 30%;
- **Feature Dimension:** n=1000 features, missing rate 30%;
- **Missing Rate:** d=30 samples, n=1000 features.

Sample Size. HI-PMK demonstrates efficient scaling with increasing sample size and remains faster than deep models like GAIN for $d \leq$ 2000, making it well-suited for small to medium-scale datasets.
Feature Dimension. As dimensionality increases, HI-PMK's runtime grows moderately due to kernel computations. It consistently outperforms kernel methods such as KPCA and PPCA, which exhibit steeper growth under high dimensions.
Missing Rate. HI-PMK maintains stable runtime across a wide range of missing rates. In contrast, deep generative models incur additional cost due to prolonged training. This stability highlights HI-PMK's robustness under high sparsity.



(a) MCAR

(b) MAR

(c) MNAR

Figure 5. F1 scores for classification tasks under different missingness mechanisms (MCAR, MAR, MNAR), evaluated across varying missing rates. Each plot illustrates how classifier performance responds to increased data degradation, highlighting sensitivity to the underlying missing data pattern.

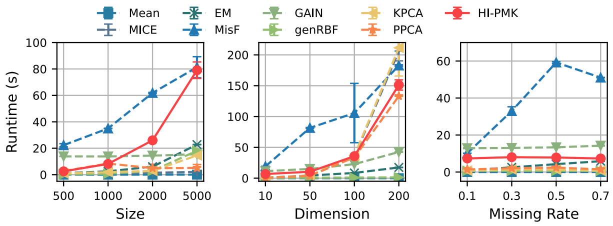


Figure 6. Scalability across sample size, dimension, and missing rate.

5.4 Ablation Study

We conduct two ablation studies to assess the contribution of HI-PMK components: (1) a module-level decomposition and (2) a comparison of uncertainty modeling strategies. Table 4 reports average F1 scores over complete datasets under MCAR, MAR, and MNAR conditions, and average classification and clustering performance on real incomplete datasets.

Component-Level Analysis. We evaluate: **PMK** (baseline using complete data only), **H-PMK** (PMK with `MICE` Imputer), **I-PMK-w/o MaxU** (removes uncertainty-aware term), **I-PMK-w/o** $\mathcal{B}_k$ (excludes frequency-based missing bucket adjustment), **I-PMK** (includes both MaxU and $\mathcal{B}_k$), and **HI-PMK** (our model).

Uncertainty Strategy Comparison. We further compare **MaxU** with **AvgU** and **MinU**, which estimate average-case and optimistic dissimilarities, respectively. MaxU assumes worst-case uncertainty by assigning the largest plausible dissimilarity when values are missing. Results show that MaxU consistently yields superior performance, especially under MNAR, validating the benefit of conservative uncertainty modeling [17]. **MaxU leads to significant performance gains**, particularly in incomplete data with structured missingness. Its conservative treatment of uncertainty helps avoid biased

Model	Adult	Australian	Banknote	Breast	Car	Heart	Sonar	Spam	Student	Wine
MCAR										
Mean	0.2546	0.6127	0.7619	0.4278	0.4377	0.1774	0.6912	0.6978	0.1985	0.7962
MICE	0.2178	0.6123	0.7725	0.4508	0.4285	0.1774	0.7406	0.6933	0.2126	0.8563
EM	0.2269	0.4800	0.7709	0.4515	0.3736	0.1774	0.6628	0.5148	0.1742	0.6900
MisF	0.2584	0.6909	**0.8109**	0.4160	0.3741	0.2281	0.6977	0.5694	0.1660	0.9234
GAIN	0.2423	0.7447	0.7613	0.4135	0.3688	0.2608	0.6046	0.8351	0.1781	0.8463
genRBF	0.2443	0.6575	0.7345	0.4236	0.2059	0.1045	0.6202	0.7939	0.2195	0.8827
KPCA	0.2438	0.5161	0.8102	0.4307	0.3517	0.1817	0.7519	0.6500	0.2285	0.8610
PPCA	0.2384	0.4987	0.7702	0.4148	0.3339	0.1817	0.7142	0.5909	0.2281	0.7617
HI- PMK	**0.2697**	**0.7500**	0.8107	**0.4605**	**0.4602**	**0.2694**	**0.7538**	**0.8673**	**0.2363**	**0.9329**
MAR										
Mean	0.2371	0.5115	0.8116	0.4781	0.5181	0.1848	0.7507	0.5755	0.2329	0.7780
MICE	0.2351	0.5245	0.8161	0.4715	0.5096	0.2159	0.7739	0.6347	0.2374	0.8515
EM	0.1960	0.4840	0.8255	0.4943	0.4560	0.2310	0.7117	0.5374	0.1792	0.7179
MisF	0.2710	0.7191	0.8550	0.4127	0.4626	0.2355	0.7556	0.5948	0.1706	0.9470
GAIN	0.2791	**0.7905**	0.8106	0.4133	0.4458	0.2814	0.6892	0.8665	0.1840	0.9077
genRBF	0.2017	0.4543	0.8089	0.4667	0.4513	0.1889	0.6919	0.5002	0.1808	0.8235
KPCA	0.2084	0.4098	0.8107	0.5108	0.4264	0.2313	0.7819	0.6770	0.2402	0.8843
PPCA	0.2351	0.5236	0.8360	0.4526	0.4027	0.2270	0.7261	0.6229	0.2352	0.7857
HI- PMK	**0.2852**	0.7841	**0.8390**	**0.5141**	**0.5250**	**0.2945**	**0.8129**	**0.8984**	**0.2429**	**0.9577**
MNAR										
Model	Adult	Australian	Banknote	Breast	Car	Heart	Sonar	Spam	Student	Wine
Mean	0.2129	0.6663	0.6855	0.4127	0.5025	0.1555	0.5402	0.7712	0.2072	0.7237
MICE	0.2316	0.6704	0.7034	0.4413	0.4950	0.2098	0.5809	0.7310	0.1757	0.7268
EM	0.2184	0.5538	0.7506	0.4746	0.3871	0.1547	0.5987	0.4988	0.1743	0.5309
MisF	0.2694	0.7200	0.7222	0.4638	0.3901	0.2672	0.6216	0.8913	0.1793	0.8658
GAIN	0.2769	0.7678	0.7710	0.4849	0.3117	0.2178	0.6414	0.8979	0.1806	0.8543
genRBF	0.1990	0.5967	0.6390	0.4719	0.2059	0.2008	0.6973	0.8493	0.2050	0.8073
KPCA	0.1940	0.5967	0.8011	0.4919	0.4527	0.1882	0.7471	0.7516	0.2275	0.8034
PPCA	0.2311	0.5791	0.7915	0.4541	0.3885	0.1870	0.5289	0.5493	0.2181	0.8483
HI- PMK	**0.2858**	**0.7851**	**0.8253**	**0.4925**	**0.5030**	**0.2825**	**0.7737**	**0.9347**	**0.2290**	**0.9111**

Table 3. Classification results showing the F1 scores of HI-PMK and other baseline methods at a missing rate of 20% across multiple datasets and mechanisms (MCAR, MAR, and MNAR).The highest score for each dataset has been highlighted in bold, and the second-highest score is underlined.

similarity estimates. $\mathcal{B}_k$ **enhances robustness for categorical features**, especially under MAR and MNAR, where missingness correlates with specific values or categories. While these components offer marginal gains on complete data, **they are critical for performance under real-world incomplete conditions**. **HI-PMK achieves the best overall results**, benefiting from the combined effects of hybrid modeling and uncertainty-aware similarity estimation.

These findings confirm the necessity of jointly modeling uncertainty and missing value frequency for achieving generalization across diverse missingness patterns. The effectiveness of HI-PMK stems from its principled design that integrates both data semantics and statistical uncertainty.

Variant	MCAR	MAR	MNAR	Cls.	Clus.
PMK	0.509	0.507	0.502	0.751	0.253
H-PMK	0.549	0.521	0.530	0.844	0.340
I-PMK-w/o MU	0.572	0.581	0.572	0.816	0.328
I-PMK-w/o $\mathcal{B}_k$	0.551	0.608	0.610	0.818	0.316
I-PMK	0.592	0.637	0.613	0.827	0.339
HI-PMK	**0.630**	**0.651**	**0.647**	**0.869**	**0.362**
HI-PMK-AvgU	0.622	0.616	0.603	0.860	0.359
HI-PMK-MinU	0.592	0.587	0.593	0.842	0.312
HI-PMK-MaxU	**0.630**	**0.651**	**0.647**	**0.869**	**0.362**

Table 4. Ablation study comparing components and uncertainty strategies.

6 Conclusion

We introduced HI-PMK, a data-dependent kernel framework that directly models similarity under incomplete and heterogeneous settings without requiring imputation. By incorporating uncertainty-aware modeling and frequency-based correction, HI-PMK handles diverse missingness mechanisms and supports mixed-type features. Extensive experiments demonstrate consistent gains in classification and clustering tasks across 15+ benchmarks. Future work will focus on reducing quadratic complexity and extending the method to temporal or graph-structured domains, where missingness patterns exhibit richer dependencies.

Acknowledgment

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

References

[1] A. Agresti. *Categorical data analysis*, volume 792. John Wiley & Sons, 2012.

[2] M. A. Al-taezi, Y. Wang, P. Zhu, Q. Hu, and A. Al-Badwi. Improved generative adversarial network with deep metric learning for missing data imputation. *Neurocomputing*, 570:127062, 2024.

[3] P. D. Allison. Missing data. *The SAGE handbook of quantitative methods in psychology*, pages 72–89, 2009.

[4] S. Aryal, K. M. Ting, T. Washio, and G. Haffari. Data-dependent dissimilarity measure: an effective alternative to geometric distance measures. *Knowledge and information systems*, 53(2):479–506, 2017.

[5] S. Aryal, K. M. Ting, T. Washio, and G. Haffari. A comparative study of data-dependent approaches without learning in measuring similarities of data objects. *Data mining and knowledge discovery*, 34(1):124–162, 2020.

[6] L. A. Belanche, V. Kobayashi, and T. Aluja. Handling missing values in kernel methods with application to microbiology data. *Neurocomputing*, 141:110–116, 2014.

[7] F. M. Bianchi, L. Livi, K. Ø. Mikalsen, M. Kampffmeyer, and R. Jenssen. Learning representations of multivariate time series with missing data. *Pattern Recognition*, 96:106973, 2019.

[8] Z. Chen, H. Li, F. Wang, O. Zhang, H. Xu, X. Jiang, Z. Song, and H. Wang. Rethinking the diffusion models for missing data imputation: A gradient flow perspective. *Advances in Neural Information Processing Systems*, 37:112050–112103, 2024.

[9] Z. Ghahramani and M. Jordan. Supervised learning from incomplete data via an em approach. *Advances in neural information processing systems*, 6, 1993.

[10] L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 507–520. Curran Associates, Inc., 2022.

[11] Y. Gui, R. Barber, and C. Ma. Conformalized matrix completion. *Advances in Neural Information Processing Systems*, 36:4820–4844, 2023.

[12] H. Hegde, N. Shimpi, A. Panny, I. Glurich, P. Christie, and A. Acharya. Mice vs ppca: Missing data imputation in healthcare. *Informatics in Medicine Unlocked*, 17:100275, 2019.

[13] R. Houari, A. Bounceur, A. K. Tari, and M. T. Kecha. Handling missing data problems with sampling methods. In *2014 International conference on advanced networking distributed systems and applications*, pages 99–104. IEEE, 2014.

[14] P. J. Huber and E. M. Ronchetti. *Robust statistics*. John Wiley & Sons, 2011.

[15] N. B. Ipsen, P.-A. Mattei, and J. Frellsen. not-miwae: Deep generative modelling with missing not at random data. In *ICLR 2021-International Conference on Learning Representations*, 2021.

[16] J. K. Kim and J. Shao. *Statistical methods for handling incomplete data*. Chapman and Hall/CRC, 2021.

[17] G. J. Klir. Principles of uncertainty: What are they? why do we need them? *Fuzzy sets and systems*, 74(1):15–31, 1995.

[18] Z. Lian, L. Chen, L. Sun, B. Liu, and J. Tao. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on pattern analysis and machine intelligence*, 45(7):8419–8432, 2023.

[19] R. Little and D. B. Rubin. *Incomplete data*. John Wiley & Sons, Inc, 2014.

[20] I. Marisca, A. Cini, and C. Alippi. Learning to reconstruct missing data from spatiotemporal graphs with sparse observations. *Advances in Neural Information Processing Systems*, 35:32069–32082, 2022.

[21] P.-A. Mattei and J. Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*, pages 4413–4423. PMLR, 2019.

[22] X. Miao, Y. Wu, L. Chen, Y. Gao, and J. Yin. An experimental survey of missing data imputation algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):6630–6650, 2022.

[23] B. Muzellec, J. Josse, C. Boyer, and M. Cuturi. Missing data imputation using optimal transport. In *International Conference on Machine Learning*, pages 7130–7140. PMLR, 2020.

[24] S. Nakagawa. Missing data: mechanisms, methods and messages. *Ecological statistics: Contemporary theory and application*, pages 81–105, 2015.

[25] A. Nazabal, P. M. Olmos, Z. Ghahramani, and I. Valera. Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, 107: 107501, 2020.

[26] K. Pelckmans, J. De Brabanter, J. A. Suykens, and B. De Moor. Handling missing values in support vector machine classifiers. *Neural Networks*, 18(5-6):684–692, 2005.

[27] Z. Rasool, S. Aryal, M. R. Bouadjenek, and R. Dazeley. Overcoming weaknesses of density peak clustering using a data-dependent similarity measure. *Pattern Recognition*, 137:109287, 2023. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2022.109287. URL https://www.sciencedirect.com/science/article/pii/S003132032200766X.

[28] R. Razavi-Far, B. Cheng, M. Saif, and M. Ahmadi. Similarity-learning information-fusion schemes for missing data imputation. *Knowledge-Based Systems*, 187:104805, 2020.

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

[30] G. Sanguinetti and N. D. Lawrence. Missing data in kernel pca. In *European Conference on Machine Learning*, pages 751–758. Springer, 2006.

[31] R. Shwartz-Ziv and A. Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

[32] M. Śmieja, Ł. Struski, J. Tabor, B. Zieliński, and P. Spurek. Processing of missing data by neural networks. *Advances in neural information processing systems*, 31, 2018.

[33] M. Śmieja, Ł. Struski, J. Tabor, and M. Marzec. Generalized rbf kernel for incomplete data. *Knowledge-Based Systems*, 173:150–162, 2019.

[34] A. J. Smola, S. V. N. Vishwanathan, and T. Hofmann. Kernel methods for missing variables. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 325–332, 2005.

[35] D. J. Stekhoven and P. Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

[36] Y. Sun, J. Li, Y. Xu, T. Zhang, and X. Wang. Deep learning versus conventional methods for missing data imputation: A review and comparative study. *Expert Systems with Applications*, 227:120201, 2023.

[37] Y. Tashiro, J. Song, Y. Song, and S. Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems*, 34:24804–24816, 2021.

[38] K. M. Ting, J. R. Wells, and T. Washio. Isolation kernel: the x factor in efficient and effective large scale online kernel learning. *Data Mining and Knowledge Discovery*, 35(6):2282–2312, 2021.

[39] S. Van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45: 1–67, 2011.

[40] J. Yoon, J. Jordon, and M. Schaar. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pages 5689–5698. PMLR, 2018.

[41] F. Yu, R. Zhao, Z. Shi, Y. Lu, J. Fan, Y. Zeng, J. Mao, and W. Li. Boosting spectral clustering on incomplete data via kernel correction and affinity learning. *Advances in Neural Information Processing Systems*, 36:72583–72603, 2023.

[42] F. Yu, Y. Zeng, J. Mao, and W. Li. A theory-driven approach to inner product matrix estimation for incomplete data: An eigenvalue perspective. In *Proceedings of the ACM on Web Conference 2025*, pages 4077–4088, 2025.

[43] W. Zheng, E. W. Huang, N. Rao, S. Katariya, Z. Wang, and K. Subbian. Cold brew: Distilling graph node representations with incomplete or missing neighborhoods. In *International Conference on Learning Representations*, 2022.

[44] Y. Zhou, S. Aryal, and M. R. Bouadjenek. A comprehensive review of handling missing data: Exploring special missing mechanisms. *arXiv preprint arXiv:2404.04905*, 2024.

[45] Y. Zhou, M. R. Bouadjenek, and S. Aryal. Missing data imputation: Do advanced ml/dl techniques outperform traditional approaches? In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 100–115. Springer, 2024.

[46] Y. Zhou, M. R. Bouadjenek, and S. Aryal. Missddim: Deterministic and efficient conditional diffusion for tabular data imputation, 2025. URL https://arxiv.org/abs/2508.03083.

[47] Y. Zhou, M. R. Bouadjenek, J. Wells, and S. Aryal. HI-PMK: A data-dependent kernel for incomplete heterogeneous data representation, 2025. URL https://arxiv.org/abs/2501.04300.

[48] Z. Zhou, J. Mo, and Y. Shi. Data imputation and dimensionality reduction using deep learning in industrial data. In *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, pages 2329–2333. IEEE, 2017.

[49] Y. Zhu and K. M. Ting. Kernel-based clustering via isolation distributional kernel. *Information Systems*, 117:102212, 2023.