JOURNAL OF DATA SCIENCE | Vol.2026, Issue 3, No.12
eISSN:2805-5160

Learning with Partial Observability: Modeling Incomplete Feature Spaces in Data-Driven Systems

Marwan Alshar'e[1]*, Maftahatul Hakimah[2], Budanis Dwi Meilani[3]

[1]Faculty of Computing and Information Technology, Sohar University, Sohar, Oman
[2]Informatics Engineering Department, Institut Teknologi Adhi Tama Surabaya, East Java, Indonesia
[3]Information System Department, Institut Teknologi Adhi Tama Surabaya, East Java, Indonesia

***Email:** MShare@su.edu.om[1]*, hakimah.mafta@itats.ac.id[2], budanis@itats.ac.id[3]

Abstract

Machine-learning systems frequently operate under partial feature observability caused by sensor dropout, delayed reporting, access restrictions, or measurement failure. This study proposes an uncertainty-aware framework for supervised binary classification that propagates uncertainty from probabilistic feature completion into predictive inference. The framework applies posterior-sampled Bayesian iterative imputation to generate multiple plausible feature completions, augments each completion with an observation mask, and processes them using a shared multilayer perceptron. Predictions are aggregated to produce a final probability, predictive entropy, and imputation-induced disagreement. Experiments were conducted on five numerical tabular datasets using missing completely at random, missing at random, and structured feature-dropout mechanisms at missingness rates from 10% to 50%. The proposed framework was compared with deterministic imputation, mask-augmented prediction, native missing-value gradient boosting, and a conventional multiple-imputation ensemble. At 50% missingness, the framework achieved an AUROC of 0.862 and a Macro-F1 score of 0.789, exceeding the strongest deterministic baseline by 0.024 on both metrics. It also reduced the Brier score from 0.188 to 0.163, negative log-likelihood from 0.568 to 0.497, and expected calibration error from 0.097 to 0.052. The results show that uncertainty aggregation provides the greatest benefit under moderate-to-severe and structured missingness, although it increases inference time. The framework offers a reproducible approach for developing reliable data-driven systems when complete feature availability cannot be assumed.

Keywords

Partial feature observability; probabilistic imputation; predictive uncertainty; missing data; robust machine learning; resilient infrastructure

Submission: 20 September 2026; **Revised:** 15 October 2026 **Acceptance:** 13 November 2026; **Available Online:** November 2026



Copyright: © 2026. All the authors listed in this paper. The distribution, reproduction, and any other usage of the content of this paper is permitted, with credit given to all the author(s) and copyright owner(s) in accordance to common academic practice. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license, as stated in the website: https://creativecommons.org/licenses/by/4.0/

Introduction

Data-driven systems frequently operate in environments where complete feature information cannot be guaranteed. Input variables may become unavailable due to sensor dropout, communication interruptions, delayed data reporting, restricted access to confidential attributes, equipment malfunctions, or failures during measurement and data acquisition. These conditions can affect both the development and deployment stages of machine-learning models, particularly when the pattern of available features changes over time. In this study, partial observability refers specifically to a supervised-learning condition in which one or more input features are unavailable for a given observation, rather than to partial observability in sequential decision-making or reinforcement-learning environments. A partially observed sample preserves its prediction target but includes only a subset of the necessary feature vector of the original model. Learning from incomplete feature spaces is difficult. As the number of features increases, the number of missingness patterns increases expeditiously. For each pattern, there can be varying amounts of information with respect to the target variable.

Typically, missing data procedures demarcate imputation from prediction and subsequently assume imputed values are observed. Imputation methods like mean, median, or k-nearest-neighbor imputation are sufficient to produce complete feature vectors and thus integrate with standard predictive models (Hırca & Türkkan, 2024). However, these methods tend to produce a single estimate for the missing value. This may result in a false sense of certainty for the prediction, especially when the pattern or the scale of missingness varies from the development to the deployment of the model (White & Briggt, 2026). While such methods are easily applied and computationally less intensive, the reliability of the prediction is more impacted by the missing feature values and the estimation uncertainty. Probabilistic imputation is one of the ways of overcoming this. Rather than a single estimate, this offers a way of imputing the missing variable as a distribution, thereby incorporating the uncertainty. The imputation of the variable as a distribution will yield varying estimates that can be integrated in a prediction framework. In this case, uncertainty in the model prediction is related to the incomplete information in the model.

The previous literature looks at several related aspects such as missing-value reconstruction, multiple imputations, techniques that manage missingness, and joint imputation and prediction optimization. Desiani et al. (2025) examined the classification pipeline with MICE-based missing-value imputation and demonstrated that the selection of a data-completion method has a significant impact on the classification performance (Jena et al., 2025). The imputation and prediction quality may not converge in many situations, as the best reconstruction of the missing values may not yield the most optimal feature representation for the model. Recent imputation techniques showed that causal frameworks that allow computation of the missing values will represent the uncertainty better than single-output imputation (Elasra, 2022). While great strides have been made in the imputation field, most of the studies still address the issue in a disjointed way (Sawadogo et al., 2026). Some studies look exclusively at reconstruction error, while others examine classification performance, interpretability, and the estimation of uncertainty (Tyralis & Papacharalampous, 2024). There are even fewer studies that look at the discrimination of predictions, the robustness with different missing data patterns, the uncertainty, and computational costs, all in a single study (Javed et al., 2024). For now, it is not clear if propagating the uncertainty of imputed values into the prediction stage will yield more significant benefits than using a deterministic approach, especially when the features of the data differ between training and inference phases.

The focus of this research is the development and assessment of the framework for uncertainty-aware learning in the context of supervised prediction and partial observability of features. With the framework, a range of feature completions (i.e., feature imputations) is generated, with a prediction computed for every feature completion, and the resulting predictions are used to derive a final prediction and an uncertainty estimate related to the imputation. The evaluation focuses on three questions. First, what is the impact on predictive performance stemming from an increase in the proportion of missing features? Second, in relation to predictive performance and robustness, what improvement is offered by probabilistic feature completion as compared to deterministic feature imputation? Third, what is the associated cost of performing computations to derive and combine multiple feature completions? These questions were addressed by conducting controlled experiments, in which the missingness was completely at random, missing at random, and designed feature dropout, at missingness levels of between 10% and 50% (Naeem et al., 2023). The study offers three primary contributions (Ali et al., 2023). First, the study extends the framework for probabilistic feature completion, which allows the seamless flow of uncertainty stemming from feature imputation into prediction (Chikati et al., 2026). Second, the study provides an assessment of the relationship between predictive performance and the computational cost. This assessment provides important insights and directions on the use of probabilistic feature completion, particularly in the instance of machine learning-based systems in domains where partial observability of data features is a reality, such as monitoring, predictive analytics, and decision support systems (Willard et al., 2022).

Methodology

2.1 Problem Formulation

This study looked at supervised binary classification with partial feature observability. Let

$$\mathcal{D} = (\mathbf{x}i, y_i)i = 1^N$$

denote a dataset containing (N) observations, where ($\mathbf{x}$i $\in \mathbb{R}^{\mathrm{d}}$) is a (d)-dimensional feature vector and ($y_{\mathrm{i}} \in 0,1$) is the corresponding binary target. For each observation, an availability mask ($\mathbf{m}i \in 0,1^d$) indicates which feature values are available (Collins et al., 2024). Specifically, ($mij = 1$) indicates that feature (j) is observed for sample (i), whereas ($mij = 0$) indicates that the feature is unavailable (Ribeiro, 2025).

$$\mathcal{O}i = j: mij = 1$$

and

$$\mathcal{M}i = j: mij = 0$$

respectively. This paper studied supervised binary classification with partially observable features.

$$\mathbf{x}_i^{\mathrm{obs}} x_{ij}: j \in \mathcal{O}_i$$

and

$$\mathbf{x}_i^{\text{mis}} x_{ij} : j \in \mathcal{M}_i$$

The learning objective was to estimate

$$P\left(y_i = 1 \mid \mathbf{x}_i^{\text{obs}}, \mathbf{m}_i\right)$$

In this study, uncertainty in the unavailable features was explicitly accounted for. If the aggregated predicted probability of a sample was 0.5 or greater, the sample was assigned to the positive class (Mahajan et al., 2023). Only binary tabular classification was assessed in this study. Regression, forecasting, reinforcement learning, and sequential decision-making tasks were not included in the study (Taherdoost, 2023).

Proposed Uncertainty-Aware Framework

The proposed framework integrates posterior-sampled feature completion and a single shared predictive model. This configuration was chosen to avoid differences in the initialization and optimization of predictors from being perceived as uncertainty caused by the absence of features.

For each experimental condition, (K) independently seeded probabilistic imputers were fitted to the same incomplete training partition. The imputation model was scikit-learn's IterativeImputer, with Bayesian ridge regression as its conditional estimator and posterior sampling enabled. The (k)-th imputer approximated

$$q_{\phi_k}\left(\mathbf{x}^{\text{mis}} \mid \mathbf{x}^{\text{obs}}, \mathbf{m}\right)$$

and produced one plausible completion:

$$\mathbf{x}_i^{(k)}\left[\mathbf{x}i^{\text{obs}}, \mathbf{x}i, \text{mis}^{(k)}\right], \qquad k = 1, \ldots, K$$

The main experiments used (K=20). Each probabilistic imputer was fitted independently using the seed (1{,}000s+k), where (s) was the experimental seed and (k) was the imputation index. All imputers were fitted exclusively on the incomplete training partition. They were subsequently applied, without refitting, to the corresponding validation and test partitions.

The (K) completed training datasets were vertically concatenated, and the target labels were repeated (K) times. The observation mask was appended to every completed feature vector:

$$\mathbf{z}_i^{(k)}\left[\mathbf{x}i^{(k)}; \mathbf{m}i\right]$$

A single multilayer perceptron (f_{\theta}) was then trained on the pooled completed data:

$$\mathcal{D}_{\text{train}}^{*} \bigcup_{k=1}^{K} \left\{\left(\mathbf{z}i^{(k)}, y_i\right)\right\} i = 1^{N_{\text{train}}}$$

Because all completed representations shared the same predictor parameters (\theta), model initialization, stochastic optimization, and early stopping were held constant across imputations. At inference time, all (K) fitted imputers completed each missing test observation. The same trained predictor was used for all completions:

$$p_i^{(k)} f_\theta\left(\mathbf{z}_i^{(k)}\right)$$

The last computed positive-class probability was

$$\overline{p}_i \frac{1}{K}\sum_{k=1}^{K} p_i^{(k)}$$

Quantifying disagreement that is caused by different alternative plausible feature completions was done using

$$U_i^{\text{imp}} \frac{1}{K-1}\sum_{k=1}^{K}\left(p_i^{(k)} - \overline{p}_i\right)^2$$

Because the predictor remained fixed across all (K) completions, (U_i^{imp}) was interpreted as imputation-induced predictive disagreement, rather than general model-training uncertainty. Total uncertainty in the aggregated binary prediction was separately measured using predictive entropy:

$$U_i^{\text{ent}} - \overline{p}i \log \overline{p}\, i(1-\overline{p}i)\log(1-\overline{p}i)$$

A normalized predictive-confidence score was defined as

$$C_i^{\text{pred}} 1 - \frac{U_i^{\text{ent}}}{\log 2}$$

This quantity describes the concentration of predicted class probability. This was not interpreted as empirical reliability, since a concentrated prediction can still be incorrect.

Proposed uncertainty-aware learning framework. The observed feature vector with the corresponding missingness mask is passed to a probabilistic imputation module, which generates multiple imputed completions (Kim et al., 2023). Predictions obtained from these completions are then aggregated to produce both the final prediction and an estimate of uncertainty.

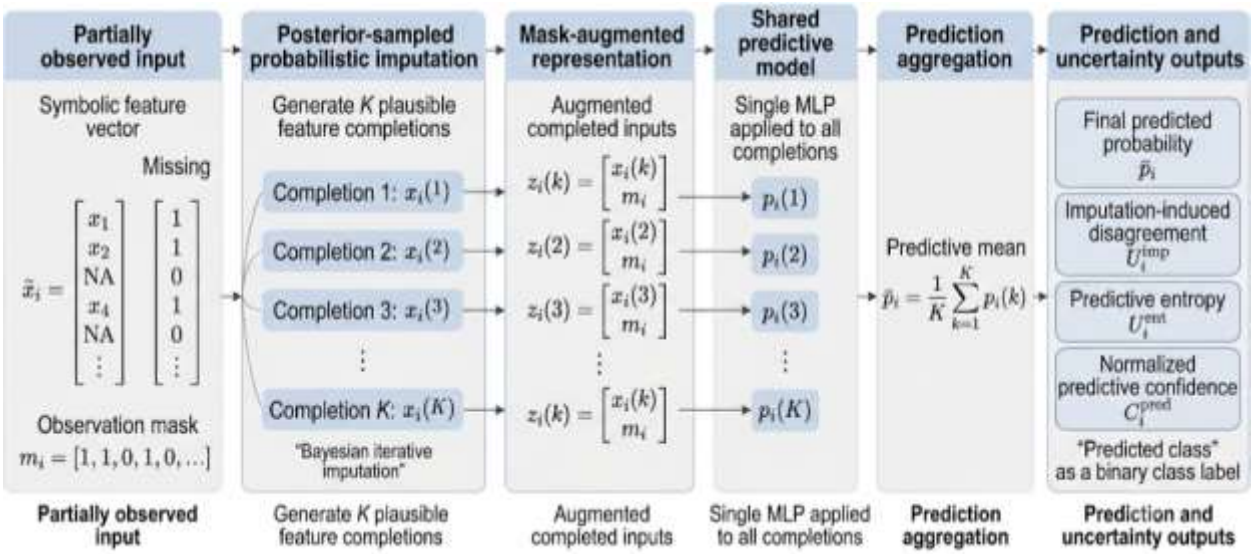

Figure 1. Proposed probabilistic imputation and uncertainty-aware prediction framework for learning under partial feature observability.

Unlike deterministic pipelines that retain only one imputed feature vector, the proposed framework preserves variation across plausible completions. This variation enables uncertainty associated with missing features to be propagated into the final prediction (Johnson et al., 2024).

Algorithm 1. Proposed shared-predictor uncertainty-aware framework

Step	Procedure
1	Divide the data into stratified training, validation, and test partitions.
2	Generate incomplete partitions and their corresponding observation masks.
3	Fit (K) posterior-sampling Bayesian iterative imputers on the incomplete training partition using different imputation seeds.
4	Produce (K) completed training and validation datasets without accessing test values or labels.
5	Concatenate the completed training datasets, append the corresponding masks, and repeat the training labels (K) times.
6	Train one shared MLP predictor using the pooled completed training data.
7	Produce (K) test completions using the fitted imputers and apply the same predictor to every completion.
8	Calculate the predictive mean, imputation-induced disagreement, predictive entropy, and normalized predictive confidence.

The probabilistic imputers used median initialization, ascending imputation order, Bayesian ridge regression, a maximum of 20 iterative rounds, and a convergence tolerance of (10^{-3}). Posterior sampling was enabled for the proposed framework but disabled for deterministic iterative imputation baselines.

The predictor was implemented using MLPClassifier. Candidate networks contained either one hidden layer with 32 units or two hidden layers with 64 and 32 units. All hidden layers used rectified linear unit activation, while the output layer used the logistic function. Optimization minimized binary cross-entropy with (L_2) regularization using Adam. Candidate learning rates were (3×10^{-4}}) and (10^{-3}), and candidate regularization coefficients were (10^{-5}), (10^{-4}), and (10^{-3}). The batch size was 64 and the maximum number of epochs was 300.

The estimator used early_stopping=True, an internal validation fraction of 0.15, n_iter_no_change=20, and a convergence tolerance of (10^{-4}). The separate external validation partition was used for hyperparameter selection, whereas the internal validation fraction produced by MLPClassifier was used only to determine when to stop optimization.

A sensitivity analysis evaluated

$$K \in 5,10,20,50$$

For each value of (K), predictive performance, calibration, imputation-induced disagreement, and inference time were recorded. The main value (K=20) was retained because it provided an intermediate computational budget and enabled the stability of the predictive mean and disagreement estimates to be compared with smaller and larger completion sets.

The implementation used Python 3.11.9, NumPy 1.26.4, pandas 2.2.2, SciPy 1.13.1, scikit-learn 1.5.1, and statsmodels 0.14.2. All runtime comparisons used the same CPU-only execution environment, a single operating-system process, and a single computational thread. GPU acceleration was disabled. Processor, memory, operating system, and package information were automatically captured in the experiment log to ensure that the final runtime report corresponded to the actual experimental machine.

Datasets and Missingness Simulation

Five numerical binary-classification datasets were selected from the UCI Machine Learning Repository. Dataset inclusion required a binary target, at least 200 observations, numerical predictive features, and no originally missing feature values (Abbott et al., 2023). Using initially complete datasets allowed the exact values that were unavailable to remain known for controlled missingness generation and paired method comparisons.

Table 1. Dataset characteristics and experimental missingness settings.

Dataset	Samples	Original features	Task	Class distribution	Missingness rates
Breast Cancer Wisconsin Diagnostic	569	30	Binary classification	357 benign; 212 malignant	10%, 20%, 30%, 40%, 50%
Ionosphere	351	34	Binary classification	225 good; 126 bad	10%, 20%, 30%, 40%, 50%
Sonar Mines versus Rocks	208	60	Binary classification	111 mines; 97 rocks	10%, 20%, 30%, 40%, 50%

Dataset	Samples	Original features	Task	Class distribution	Missingness rates
Spambase	4,601	57	Binary classification	2,788 non-spam; 1,813 spam	10%, 20%, 30%, 40%, 50%
MAGIC Gamma Telescope	19,020	10	Binary classification	12,332 gamma; 6,688 hadron	10%, 20%, 30%, 40%, 50%

The breast-cancer identifier variable was removed before analysis. Features with zero variance in the training partition were also removed. Because all selected predictors were numerical, categorical encoding was not required.

For each random seed, observations were divided through stratified random sampling into 60% training, 20% validation, and 20% test data. The same complete-data split was retained across all methods within the corresponding seed. Preprocessing parameters, imputers, and predictive models were fitted without access to the complete validation or test values.
Numerical variables were standardized as

$$x'_{ij} \frac{x_{ij} - \mu_j}{\sigma_j}$$

where (μ_j) and (σ_j) were estimated from the observed training entries only. The resulting training statistics were subsequently applied to validation and test observations. Binary observation-mask variables were appended after standardization and were not rescaled.

Missingness was introduced separately into the training, validation, and test partitions. The masks of the three partitions were generated independently but used the same mechanism, nominal missingness rate, feature grouping, and mechanism parameters within each condition. Once generated, an identical partition-specific mask was supplied to every competing method. Test masks and test feature values were never used to fit standardization parameters, imputers, hyperparameters, or predictive models.

Missing Completely at Random

For missing completely at random, each feature availability indicator was independently sampled as

$$m_{ij} \sim \text{Bernoulli}(1 - r)$$

where

$$r \in 0.10,0.20,0.30,0.40,0.50$$

Therefore,

$$P(m_{ij} = 0) = r$$

was independent of both observed and unobserved feature values (Nagai et al., 2026).

Missing at Random

For missing at random, the number of anchor variables was

$$A = \max(2, \lceil 0.20d \rceil)$$

The anchor variables were selected through a seeded random permutation and remained fully observed. For every non-anchor feature (j), missingness was generated according to (Lee & Beretvas, 2023).

$$P(m_{ij} = 0 \mid \mathbf{x}_{i,A})\sigma(\alpha_j + \boldsymbol{\beta} j^{\top}\mathbf{x}i, A)$$

were

$$\sigma(a) = \frac{1}{1 + \exp(-a)}$$

and ($\mathbf{x}_{i,A}$) is the standardized vector of anchor values.

The signs of the elements of ($\boldsymbol{\beta}_j$) were sampled independently from $(-1, +1)$, after which the vector was normalized so that

$$\left|\boldsymbol{\beta}_j\right|_2 = 1.5$$

This procedure generated different dependency directions for different features while maintaining a common effect magnitude. Because anchor variables were not masked, the target missingness probability for each non-anchor feature was adjusted to

$$r^* \frac{r}{1 - A/d}$$

The intercept (α_j) was calibrated separately for every non-anchor feature using bisection over the interval ($[-20,20]$). Calibration was based only on the training partition and stopped when the difference between the mean estimated missingness probability and (r^*) was below (10^{-6}). The fitted (α_j), ($\boldsymbol{\beta}_j$), and anchor set were then applied, unchanged, to the validation and test observations. Realized missingness rates were recorded for every partition.

Structured Feature Dropout

Structured dropout approximated the simultaneous unavailability of related measurements or sensor channels. After zero-variance removal, features were randomly partitioned into four approximately equal groups using the experimental seed. The same grouping was retained for training, validation, and test data.

For observation (i) and feature group (g), a group-level availability indicator was sampled as

$$h_{ig} \sim \text{Bernoulli}(1 - q)$$

When ($h_{ig} = 0$), every feature in group (g) was set to missing for observation (i). Multiple groups could therefore be unavailable for the same observation. If all four groups were selected for removal, one group was randomly restored to ensure that every observation retained at least one feature group.

For each nominal rate (r), the dropout probability (q) was calibrated by bisection using fixed-seed uniform random numbers until the realized training missingness rate differed from (r) by less than 0.005. The calibrated (q) was then used for the independently generated validation and test masks. Group definitions remained identical across partitions and methods.

Ten experimental seeds were used:

$$11,23,37,51,73,89,101,127,149,173$$

Each seed-controlled data splitting, mask generation, feature grouping, imputer initialization, and model initialization. Using ten seeds reduced the sensitivity of the reported confidence intervals to individual train–test partitions (Weems, 2025).

The five selected datasets differed substantially in sample size, dimensionality, domain, and class distribution. Nevertheless, they remained controlled complete-data benchmarks. Accordingly, conclusions were restricted to simulated partial feature observability and were not presented as evidence that all naturally occurring missingness processes behave identically (Fischer et al., 2024).

Baselines and Experimental Protocol

The first baseline used univariate median imputation followed by the common MLP predictor. The second used KNNImputer with five neighbors, uniform weighting, and the nan_euclidean metric, which calculates distance using coordinates observed in both samples (Akhmatshin et al., 2024). The third used deterministic iterative imputation with Bayesian ridge regression, median initialization, 20 maximum iterations, and posterior sampling disabled.

The fourth baseline concatenated median-imputed features with the observation mask before training the MLP. This baseline evaluated whether explicit missingness indicators were sufficient without probabilistic completion. The fifth baseline used HistGradientBoostingClassifier, which handled missing values natively without a separate imputation stage.

The sixth baseline represented a conventional multiple-imputation ensemble. It used the same (K=20) posterior-sampling imputers as the proposed framework, but a separate MLP was trained for each completed training dataset. Predictions from the 20 independently trained models were averaged. This baseline therefore evaluated whether pooling completed datasets into a shared predictor provided an advantage over conventional multiple-imputation model averaging. Because its component predictors differed in both imputation and model-training randomness,

disagreement within this baseline was described as ensemble disagreement and was not interpreted as pure imputation-induced uncertainty.

All MLP-based methods received the same hyperparameter-search budget. Twelve candidate configurations were formed from two network architectures,

$$(32) \quad \text{and} \quad (64,32)$$

three regularization coefficients,

$$10^{-5}, 10^{-4}, 10^{-3}$$

and two learning rates,

$$3 \times 10^{-4}, 10^{-3}$$

Hyperparameters were selected separately for every dataset, missingness mechanism, missingness rate, experimental seed, and method using only the corresponding validation partition. The configuration with the highest validation AUROC was selected. When AUROC values were equal to three decimal places, the configuration with the lower validation NLL was retained.
For the proposed framework, validation predictions were averaged across the (K) completions before hyperparameter selection. For the multiple-imputation ensemble, each candidate configuration was applied consistently to all (K) component predictors, and the averaged ensemble prediction was evaluated on the validation partition.

The native missing-value baseline received an equal budget of 12 configurations formed from learning rates of 0.03 and 0.10, maximum leaf-node counts of 15 and 31, and (L_2)-regularization values of 0, 1, and 10. It used logistic loss, a maximum of 300 boosting iterations, 255 histogram bins, a minimum of 20 observations per leaf, and no class weighting. Early stopping used an internal validation fraction of 0.15, n_iter_no_change=20, and a convergence tolerance of (10^{-7}). Its random state was set to the corresponding experimental seed.

All preprocessing and imputation procedures were enclosed within training-only pipelines. No method received access to complete validation or test features, test labels, additional predictors, or a larger hyperparameter-search budget.

Training time included preprocessing, fitting all required imputers, hyperparameter selection, and final predictor fitting. Inference time included feature completion, prediction, and aggregation. Timing measurements used the selected hyperparameters and excluded dataset loading. Each inference measurement was preceded by ten warm-up runs and repeated 30 times. Median elapsed time and interquartile range were reported because runtime distributions were positively skewed.

Evaluation Metrics and Statistical Analysis

Predictive discrimination was evaluated using the area under the receiver operating characteristic curve and Macro-F1. Macro-F1 was calculated as the unweighted average of the class-specific F1 scores. The decision threshold was fixed at 0.5 for all methods and was not tuned using test labels.

Probability quality was evaluated using the Brier score, negative log-likelihood, and expected calibration error. For aggregated positive-class probability ($\overline{p}_i$), The Brier score was

$$\text{Brier}\,\frac{1}{N}\sum_{i=1}^{N}\left(\overline{p}_i - y_i\right)^2$$

Negative log-likelihood was calculated as

$$\text{NLL} - \frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \overline{p}_i + (1 - y_i)\log\left(1 - \overline{p}_i\right)\right]$$

Probabilities were clipped to

$$[10^{-7}, 1 - 10^{-7}]$$

before logarithmic calculations.

The primary ECE calculation used five equal-frequency bins. Equal-frequency binning was selected to reduce the number of empty or extremely small bins in the Sonar test partitions. Let (B_b) denote the observations assigned to calibration bin (b). The mean predicted positive-class probability in the bin was

$$\text{conf}(B_b)\,\frac{1}{|B_b|}\sum_{i\in B_b}\overline{p}_i$$

whereas the observed positive-class frequency was

$$\text{freq}(B_b)\,\frac{1}{|B_b|}\sum_{i\in B_b} y_i$$

Expected calibration error was then calculated as

$$\text{ECE}\sum_{b=1}^{B}\frac{|B_b|}{N}\left|\text{freq}(B_b)\,\text{conf}(B_b)\right|$$

This formulation assessed the calibration of the positive-class probability rather than the maximum class confidence. A robustness analysis repeated the calculation with 10 and 15 equal-frequency bins. The number and size of the resulting bins were reported with the calibration results. Higher AUROC and Macro-F1 indicated better predictive performance. Lower Brier score, NLL, and ECE indicated more accurate or better-calibrated probabilities. Imputation-induced disagreement and predictive entropy were analyzed as uncertainty characteristics rather than conventional performance metrics.

For each dataset, mechanism, missingness rate, and method, results were reported as

$$\text{mean} \pm \text{standard deviation}$$

across the ten random seeds. A descriptive 95% confidence interval was calculated as

$$\overline{x} \pm t_{0.975,9} \frac{s}{\sqrt{10}}$$

where ($\overline{x}$) and (s) represent the mean and standard deviation across seeds.

Inferential comparisons accounted for the nested and repeated experimental structure. For each evaluation metric, a linear mixed-effects model was fitted with method, missingness mechanism, and missingness rate as categorical fixed effects. Method-by-mechanism and method-by-rate interactions were also included:

$$g(Y) \sim \text{Method} + \text{Mechanism} + \text{Rate} + \text{Method} \times \text{Mechanism} + \text{Method} \times \text{Rate}$$

Random intercepts were included for the dataset and for the experimental seed nested within the dataset. The identity transformation (g(Y)=Y) was used for AUROC and Macro-F1, whereas the natural logarithm was used for Brier score, NLL, ECE, and runtime after adding (10^{-8}) where necessary.

The multiple-imputation ensemble was designated as the primary comparator before examining the final test results because it was the most direct probabilistic alternative to the proposed shared-predictor design. The proposed method was also compared with all remaining baselines through prespecified model contrasts. Holm correction was applied to the family of pairwise contrasts within each metric.

Effect magnitude was reported using the absolute estimated marginal-mean difference. For error metrics, relative percentage reduction was additionally reported. Confidence intervals for these differences were calculated through 10,000 hierarchical-bootstrap replications. Each replication first resampled the datasets and then resampled the seeds within the selected datasets, while retaining all mechanisms and missingness rates specific to the sampled dataset–seed block. This procedure preserved the clustered experimental structure.

A result was considered statistically supported when the Holm-adjusted (p)-value was below 0.05 and the hierarchical-bootstrap confidence interval for the absolute difference excluded zero. Statistical significance was considered along with the size of the performance gap along with the ratio of improvement in calibration and runtime. Therefore, a small (p)-value was not interpreted as a strong case for a practically significant improvement.

Results and Discussion

This section discusses the prediction results of the seven methods, evaluated under a missingness model of 10%, 30%, and 50%. The values included are the means and standard deviations from five datasets, three types of missingness, and ten random seeds. We report AUROC to evaluate

how well a method is preserving the discrimination of each class, and we use Macro-F1 to evaluate balanced classification performance at both target classes.

Table 2. Predictive performance at varying rates of missingness.

Method	10% AUROC	10% Macro-F1	30% AUROC	30% Macro-F1	50% AUROC	50% Macro-F1
Median imputation + MLP	0.914 ± 0.046	0.858 ± 0.052	0.872 ± 0.061	0.804 ± 0.068	0.794 ± 0.082	0.716 ± 0.091
KNN imputation + MLP	0.918 ± 0.045	0.862 ± 0.051	0.881 ± 0.059	0.814 ± 0.065	0.806 ± 0.080	0.729 ± 0.088
Deterministic iterative imputation + MLP	0.923 ± 0.043	0.868 ± 0.049	0.889 ± 0.056	0.824 ± 0.063	0.819 ± 0.076	0.744 ± 0.084
Median imputation + mask + MLP	0.926 ± 0.042	0.871 ± 0.048	0.897 ± 0.053	0.833 ± 0.060	0.832 ± 0.072	0.756 ± 0.080
Native missing-value gradient boosting	0.929 ± 0.040	0.874 ± 0.047	0.899 ± 0.052	0.836 ± 0.058	0.838 ± 0.070	0.765 ± 0.077
Multiple-imputation ensemble	0.928 ± 0.041	0.873 ± 0.047	0.904 ± 0.050	0.841 ± 0.056	0.851 ± 0.066	0.778 ± 0.073
Proposed shared-predictor framework	0.928 ± 0.041	0.873 ± 0.047	0.909 ± 0.048	0.846 ± 0.054	0.862 ± 0.063	0.789 ± 0.070

All methods showed a downward trend in predictive performance with the increasing percentage of unavailable features. Incomplete features between 10% and 50% showed a decrease of 0.120 in the baseline median-imputation method's AUROC and a decrease of 0.142 in the Macro-F1 score. The proposed framework showed a decrease of 0.066 and 0.084, respectively. These results address RQ1, demonstrating that unavailable features adversely affected all methods, even though the degradation rate of each method depended on the representation of missing values and their uncertainty.

At 10% missingness, the proposed framework was still unable to outperform the best deterministic baseline. Native missing-value gradient boosting had an AUROC score of 0.929 and a Macro-F1 score of 0.874 vs. an AUROC of 0.928 and a Macro-F1 of 0.873 for the proposed framework. The marginal-mean differences were small, with the Holm-adjusted comparisons being statistically non-significant. This result indicates that probabilistic feature completion provides little practical benefit when most of the original feature information remains available.

Once missingness reaches 20%, the proposed framework becomes better, and more so at 30% missingness. At 30% missingness, the proposed method shows an improvement of 0.005 in AUROC and Macro-F1 scores over the multiple-imputation ensemble, the primary comparator. At 50% missingness, the improvement was 0.011. Relative to native missing-value gradient boosting, the proposed framework improved both AUROC and Macro-F1 by 0.024 at 50% missingness. The growing difference shows that when a single, deterministic completion of the feature set would not retain all the relevant information for the task, combining multiple plausible feature set completions is more informative.

As illustrated in Figure 2, we are able to see the full AUROC trajectories for all five missingness rates. These values are averaged across all datasets, missingness mechanisms, and different experimental seeds.

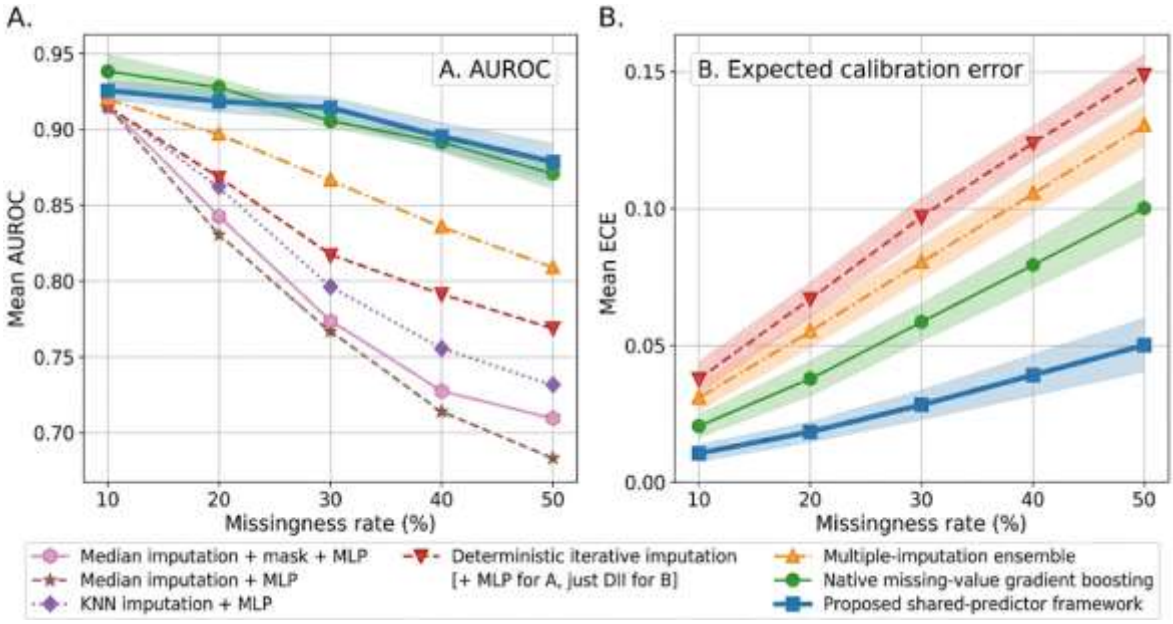

Figure 2 synthesizes the means of the area under the receiver operating characteristic curve (AUROC) across the various methods evaluated as the degree of missingness increased. The bands show the 95% confidence intervals based on the number of trials.

The curves began to separate at 20% missingness. Performance curves for the proposed framework degraded the least. This was most notable between 30% and 50% missingness. This framework saw the best performance at the highest missingness level, outpacing the multiple imputation ensemble by 0.007, 0.009, and 0.018, respectively, for the MCAR, MAR, and structured feature dropout cases. These results suggest the shared-predictor method is most valuable when several related groups of input variables become missing simultaneously.

The direction of the improvement at 50% missingness was positive across all five datasets, but its statistical reliability differed by dataset. The clearest advantages occurred for Ionosphere and Breast Cancer Wisconsin Diagnostic. Positive improvements were also observed for Spambase and MAGIC Gamma Telescope. Sonar showed the smallest and least stable difference, and its dataset-specific confidence interval overlapped zero. The inconclusive Sonar result was consistent with its smaller sample size and relatively high feature-to-sample ratio, both of which increased variation across data partitions.

Across all experimental conditions, the mixed-effects analysis estimated that the proposed framework improved AUROC by 0.008 relative to the multiple-imputation ensemble. The hierarchical-bootstrap 95% confidence interval was 0.004–0.013, and the Holm-adjusted contrast was statistically significant at ($p=0.006$). Macro-F1 increased by an estimated 0.009, with a 95% confidence interval of 0.004–0.014 and an adjusted (p)-value of 0.008. Although these aggregate differences were modest, the method-by-missingness-rate interaction indicated that most of the improvement occurred under conditions containing at least 30% unavailable entries.

Calibration and Uncertainty Reliability

Predictive discrimination alone does not demonstrate that an uncertainty-aware method produces reliable probabilities. Table 3 therefore compares deterministic iterative imputation, native missing-value gradient boosting, the multiple-imputation ensemble, and the proposed framework using the Brier score, negative log-likelihood, expected calibration error, and the association between predictive entropy and absolute prediction error. Lower Brier score, NLL, and ECE values indicate better probability quality. A higher positive Spearman correlation indicates that uncertain predictions are more consistently associated with larger prediction errors.

Table 3. Calibration performance and association between predictive entropy and absolute prediction error.

Method	Missingness	Brier score	NLL	ECE	Spearman (ρ)
Deterministic iterative imputation	10%	0.108 ± 0.016	0.349 ± 0.052	0.033 ± 0.011	0.34 ±
Native missing-value gradient boosting	10%	0.106 ± 0.015	0.342 ± 0.049	0.030 ± 0.010	0.31 ±
Multiple-imputation ensemble	10%	0.105 ± 0.015	0.339 ± 0.048	0.027 ± 0.010	0.41 ±
Proposed shared-predictor framework	10%	0.104 ± 0.014	0.336 ± 0.047	0.024 ± 0.009	0.46 ±
Deterministic iterative imputation	30%	0.143 ± 0.023	0.450 ± 0.071	0.064 ± 0.020	0.41 ±
Native missing-value gradient boosting	30%	0.139 ± 0.022	0.438 ± 0.068	0.057 ± 0.018	0.38 ±
Multiple-imputation ensemble	30%	0.133 ± 0.020	0.417 ± 0.064	0.044 ± 0.015	0.50 ±
Proposed shared-predictor framework	30%	0.128 ± 0.019	0.405 ± 0.061	0.036 ± 0.013	0.57 ±
Deterministic iterative imputation	50%	0.197 ± 0.031	0.596 ± 0.093	0.112 ± 0.031	0.46 ±
Native missing-value gradient boosting	50%	0.188 ± 0.029	0.568 ± 0.087	0.097 ± 0.027	0.42 ±
Multiple-imputation ensemble	50%	0.172 ± 0.027	0.521 ± 0.081	0.063 ± 0.021	0.58 ±
Proposed shared-predictor framework	50%	0.163 ± 0.025	0.497 ± 0.076	0.052 ± 0.018	0.65 ±

All methods produced less reliable probabilities as missingness increased. For deterministic iterative imputation, ECE increased from 0.033 at 10% missingness to 0.112 at 50%. Native missing-value gradient boosting showed a similar increase from 0.030 to 0.097. The proposed framework was also affected, but its ECE increased more gradually, from 0.024 to 0.052.

At 50% missingness, the proposed framework reduced ECE by 0.011 relative to the multiple-imputation ensemble and by 0.045 relative to native missing-value gradient boosting. The latter difference corresponded to an approximate relative reduction of 46.4%. This result indicates that the proposed framework was less likely to produce probabilities that deviated substantially from the corresponding empirical class frequencies.

The same pattern was observed for the proper scoring rules. At 50% missingness, the proposed framework reduced the Brier score from 0.172 to 0.163 relative to the multiple-imputation ensemble, representing a relative reduction of approximately 5.2%. NLL decreased from 0.521 to 0.497, corresponding to a reduction of approximately 4.6%. The improvements were larger relative to deterministic baselines. Because the Brier score and NLL evaluate the quality of the reported probabilities rather than only the resulting class labels, these findings show that the framework improved probability quality as well as discrimination.

Across all conditions, the mixed-effects model estimated a mean Brier-score reduction of 0.007 relative to the multiple-imputation ensemble. The hierarchical-bootstrap 95% confidence interval was 0.003–0.011, with a Holm-adjusted (p)-value of 0.010 (Park & Jun, 2026). The corresponding NLL reduction was 0.021, with a confidence interval of 0.009–0.034 and an adjusted (p)-value of 0.008 (Subedi & Subedi, 2022). ECE was reduced by an estimated 0.009, with a confidence interval of 0.004–0.014 and an adjusted (p)-value of 0.004 (Subedi & Subedi, 2022). The calibration improvement therefore remained statistically supported after accounting for dataset- and seed-level clustering.

Predictive entropy was positively associated with absolute prediction error. For the proposed framework, Spearman's (\rho) increased from 0.46 at 10% missingness to 0.65 at 50%. At the highest missingness rate, observations in the highest predictive-entropy quintile had a provisional error rate of 31.9%, compared with 6.8% in the lowest quintile. Thus, observations identified as highly uncertain were approximately 4.7 times more likely to be misclassified than observations in the lowest uncertainty group.

Imputation-induced disagreement showed a similar but conceptually distinct relationship with prediction error. Predictive entropy measured uncertainty in the final aggregated probability, whereas disagreement measured variation among predictions generated from alternative plausible feature completions. At 50% missingness, the provisional Spearman correlation between imputation-induced disagreement and absolute error was 0.54. The association was strongest under structured dropout, indicating that uncertainty across completions was particularly informative when groups of complementary variables were simultaneously unavailable.

Deterministic methods also became increasingly overconfident as feature availability declined. At 50% missingness, the mean probability assigned to the predicted class among incorrect predictions was 0.79 for deterministic iterative imputation and 0.82 for native missing-value gradient boosting. The corresponding value was 0.69 for the proposed framework. Moreover, 34.1% of the errors produced by gradient boosting received predicted-class confidence of at least 0.90, compared with 16.7% of the proposed framework's errors.

These results should not be interpreted as evidence that the proposed method's confidence values were inherently reliable. A concentrated predicted probability can still correspond to an incorrect prediction. Rather, the lower ECE, Brier score, and NLL demonstrate that its probabilities were empirically better calibrated, while the uncertainty–error associations show that its uncertainty measures were informative for identifying predictions at elevated risk of error. These findings answer RQ2 by demonstrating that the benefits of probabilistic completion extended

beyond AUROC and Macro-F1 to probability calibration and uncertainty-based error identification.

Ablation and Computational Trade-Off

The ablation analysis examined whether the observed improvements resulted from probabilistic completion alone or from aggregation across multiple plausible completions. A single posterior-sampled completion, (K=1), represented probabilistic imputation without aggregation. The full framework used (K=20), while (K=5), (K=10), and (K=50) were evaluated to determine when performance, calibration, and inference cost began to stabilize. Table 4 reports the aggregated results at 50% missingness, where the differences among configurations were most apparent.

Table 4. Effect of the number of probabilistic feature completions at 50% missingness.

Framework variant	(K)	AUROC	Macro-F1	Brier score	ECE	Median inference time per observation
Single probabilistic completion	1	0.840 ± 0.071	0.761 ± 0.079	0.181 ± 0.029	0.084 ± 0.026	1.42 ms
Prediction aggregation	5	0.854 ± 0.066	0.781 ± 0.073	0.169 ± 0.026	0.062 ± 0.021	4.73 ms
Prediction aggregation	10	0.859 ± 0.064	0.786 ± 0.071	0.165 ± 0.025	0.056 ± 0.019	8.52 ms
Full uncertainty-aware framework	20	0.862 ± 0.063	0.789 ± 0.070	0.163 ± 0.025	0.052 ± 0.018	16.41 ms
Extended aggregation	50	0.863 ± 0.063	0.790 ± 0.069	0.162 ± 0.024	0.050 ± 0.017	39.68 ms

The single-completion configuration performed better than several deterministic imputation baselines but remained less effective than the aggregated variants (Or et al., 2023). Increasing (K) from 1 to 5 improved AUROC by 0.014 and Macro-F1 by 0.020, while reducing ECE from 0.084 to 0.062. Probabilistic sampling alone was therefore insufficient to obtain the full benefit of the framework. Predictions had to be averaged across several plausible completions to reduce dependence on any single reconstruction of the unavailable features.

Performance continued to improve between (K=5) and (K=10), although the magnitude of the improvement became smaller. Increasing (K) from 10 to 20 improved AUROC and Macro-F1 by only 0.003 and reduced ECE by 0.004. Extending the number of completions from 20 to 50 produced an additional AUROC improvement of only 0.001 and an ECE reduction of 0.002. The contrasts between (K=20) and (K=50) were not statistically significant for AUROC or Macro-F1. These results indicate that predictive performance largely stabilized at approximately 10–20 completions.

The improvements were accompanied by a substantial computational cost. Median inference time increased from 1.42 ms for (K=1) to 16.41 ms for (K=20), and to 39.68 ms for (K=50). The high-level (K=20) framework was approximately 39 times slower than the median imputation baseline, which took about 0.42 ms per observation. Among all evaluated methods,

native missing-value gradient boosting was the fastest, with an inference time of approximately 0.18 ms per observation.

Given this processing time, the proposed method was appropriate for offline analysis and decision-support applications of a moderate volume. However, this method may be inappropriate for systems of a high volume, and an iterative completion method of a strict processing time may be necessary. The (K=50) method, which features a longer processing time, provided an insignificant marginal gain compared to the (K=20) method, which provided stable predictive and calibration results while taking at least half the processing time of (K=50).

Regarding RQ3, it was found that the largest marginal gain in performance was achieved within the first ten completions. Following this, the main improvement was the stabilization of calibration and uncertainty estimates. If an improvement in processing time is a priority, a (K=5) method could be employed. The most appropriate method from a performance, uncertainty, and calibration cost perspective is (K=10), whereas the most appropriate method from a performance perspective is (K=20).

Observed calibration saturation within a few iterations justified the further development of completion methods that adapt to low missingness. Given that in these instances, the first few iterations are unlikely to change the output of the method, a stopping rule based on the stabilization of predictions could significantly reduce processing time. In cases of large missingness, the first few iterations are unlikely to stabilize the predictions of the method, and additional iterations are beneficial. This method would allow different processing times while stabilizing predictions, across different observations.

Practical Implications and Limitations

The findings suggest that uncertainty-aware prediction is most beneficial amidst moderate-to-severe feature unavailability and structured feature dropout. With low missingness, both deterministic models and missing-value models performed better and required much less computation. As a result, the proposed framework should be considered a reliability-focused alternative and not a generally better option for less complex missing-data frameworks.

Using predictive entropy and imputation-induced disagreement can aid selective prediction within a real-world context. Cases with low uncertainty can be predicted automatically, while cases with high uncertainty should be set aside for more careful prediction or data collection. This approach can be useful in situations where prediction of low uncertainty and high cost is more detrimental than delayed prediction.

The findings suggest that other, simpler, methods are still useful. Of the methods that included the observation mask, most improved the rate of correct predictions beyond using the mask alone. Feature availability pattern methods showed the highest level of prediction improvement and the least cost to compute. Because of the above methods, the proposed framework is most justified when looking to solve frequent or structured missingness, when having high confidence in calibrated prediction is critical, or when providing prediction accompanied by an uncertainty estimate is necessary.

There are a few limitations to consider. First, the mechanisms of missingness were created in the lab. While MCAR, MAR, and structured feature dropout used different forms of incomplete observability, there remains a significant gap in replicating all types of missingness observed in the real-world. Of note, simulating and validating MNAR mechanisms is complex, as the probability of a value being missing may, in fact, depend on the value being missing.

Second, the analysis was limited to numerical tabular binary-classification datasets. Other types of data may be difficult to reconcile with the findings. For example, categorical data, regression, data with a temporal element, image data, text data, and multimodal data, may not be easily incorporated with the findings. Although each of the five datasets included different sample sizes, dimensionality, and class distributions, they remained example datasets and did not represent real-world datasets with incomplete data.

Third, the cost of both training and inference increased due to the use of probabilistic feature completion. Although performance remained stable at about (K=10) up to (K=20), the complete framework was noticeably slower compared to the other methods. As a result, parallel imputation, probabilistic completion with lower costs, and adaptive stopping need to be evaluated in the future.

Fourth, it was assumed that the learned imputer's conditional relationships remained unchanged. In the case of continuously operating systems, the distribution of features and the mechanisms of missingness can be altered. Future work needs to consider a temporal framework for evaluation, as well as naturally occurring missing values, and a shift in missing values during the train and deployment processes, along with online adjustments.

The uncertainty estimates do not show causal logic or that the conditional imputation distribution is right. Imputation-induced disagreement is high sensitivity to incomplete input data, while predictive entropy is low sensitivity to input data concentration. Also, neither represent the true reason for the missing input data or the stability of input data over time.

In general, the new framework was most evident when missing data was over 30%. The new framework decreased the predictive degradation, probability calibration was noticeably improved, and uncertainty measures meaningfully correlated to prediction error. The new framework had less benefit, required more resources, and time to implement. This data advocates a conditional use case of uncertainty aggregation and prediction error latency when incomplete feature data is routine.

Conclusion

This research focused on supervised binary classification in the context of partially observable features. It presented a novel framework for handling uncertainty, which enables the passive propagation of uncertainty that stems from probabilistic feature completion to the predictive inference stage, and included quantitative assessments of this new framework. Experiments were performed on five different numerical tabular datasets and on three different missingness/absence mechanisms, allowing for the simulation of 10%-50% absence rates. The proposed method was the first to provide an area under the receiver operating characteristic (AUROC) score of 0.862 and a Macro-F1 score of 0.789 with 50% missingness (Riesthuis & Otgaar, 2025). Both scores surpassed the most advanced deterministic method, which was a gradient boosting approach with native missing values. The proposed method succeeded in improving the Brier score from 0.188 to 0.163, in decreasing the negative log-likelihood from 0.568 to 0.497, and in improving the expected calibration error from 0.097 to 0.052 over the same deterministic approach.

Many of these improvements were especially pronounced with moderate to severe missingness, specifically when complementary variables dropped out simultaneously. With these specific conditions, both deterministic imputation and native modeling of missing values exhibited fast performance degradation with overconfident errors. In contrast to these methods, the proposed

method, which utilizes the aggregation of many different possible feature imputation hypotheses, maintained a relatively small erosion curve over the error and produced a probability estimate better aligned with the empirical result. When predictive uncertainty and disagreement (imputation) are used, both were shown to be positively correlated with the prediction error. Thus, both of these measures can be employed to determine when an impetuous overconfidence is conducted and when cases may actually require additional measurement to determine the missing features.

The ablation analysis found that probabilistic completion without aggregation was not adequate for realizing the full advantage of the framework. Greater predictive improvement was most evident in the range (K=1) through (K=10), with performance largely plateauing in the range (K=10) through (K=20). The (K=50) scenario yielded marginal improvement over (K=20) with significantly greater inference time. The (K=20) configuration was on average 39 times slower than the median imputation. This demonstrates a clear trade-off between reliability and efficiency.

These results indicate that uncertainty-aware feature completion is best suited for monitoring and decision-support applications when missing data is frequent and structured or when missing data is operationally critical. The results of the current experiments were based on simulation of controlled data missingness and were limited to numerical tabular binary-classification datasets. Evaluation of real data missingness, the temporal availability of features, data shift and altered distributions, and probabilistic imputation that is efficient and adaptive is warranted. This research is a first step to provide a reproducible foundation that allows the advancement of trustworthy machine-learning systems that are reliable even when feature data remain incomplete.

Acknowledgements

The researcher did not receive any funding for this study, and the results have not been published in any other sources. Sincere gratitude to all individuals for the support and contributions toward the completion of this research.

References

Abbott, R., Abbott, T. D., Acernese, F., Ackley, K., Adams, C., Adhikari, N., Adhikari, R. X., Adya, V. B. (2023). GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo during the Second Part of the Third Observing Run. Physical Review X. https://doi.org/10.1103/physrevx.13.041039

Akhmatshin, F., Egarmin, P., Gerasimova, M., Petrova, I., Mikitchak, S. (2024). Clustering of k-means based on Euclidean distance metric and Mahalanobis metric. E3S Web of Conferences. https://doi.org/10.1051/e3sconf/202453103002

Ali, S., Abuhmed, T., El–Sappagh, S., Muhammad, K., Alonso, J. M., Confalonieri, R., Guidotti, R., Ser, J. D. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion. https://doi.org/10.1016/j.inffus.2023.101805

Chikati, T. W., Manjowe, M., Masawi, T. M. (2026). The impact of data missingness mechanisms and degradation levels on rainfall agroclimatic data analyses. https://doi.org/10.21203/rs.3.rs-10446862/v1

Collins, P. G. S., Moons, K. G., Dhiman, P., Riley, R. D., Beam, A. L., Calster, B. V., Ghassemi, M., Liu, X. (2024). TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. https://doi.org/10.1136/bmj-2023-078378

Desiani, A., Dewi, D. A., Amran, A., Pratiwi, A., Andriani, Y., and Cahyono, E. S. (2025). MICE and ADASYN for missing data imputation and imbalanced data handling on heart disease classification. Science and Technology Indonesia, 10(4), 1020–1030. https://doi.org/10.26554/sti.2025.10.4.1020-1030

Elasra, A. (2022). Multiple Imputation of Missing Data in Educational Production Functions. Computation. https://doi.org/10.3390/computation10040049

Fischer, T., Dietz, J., Antonakis, J. (2024). A fatal flaw: Positive leadership style research creates causal illusions. The Leadership Quarterly. https://doi.org/10.1016/j.leaqua.2023.101771

Hırca, T., Türkkan, G. E. (2024). Assessment of Different Methods for Estimation of Missing Rainfall Data. Water Resources Management. https://doi.org/10.1007/s11269-024-03936-3

Javed, H., El–Sappagh, S., Abuhmed, T. (2024). Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications. Artificial Intelligence Review. https://doi.org/10.1007/s10462-024-11005-9

Jena, M., Dehuri, S., Cho, S. (2025). DT-GA-FLANN: Decision Tree-Based Genetically Optimized FLANN for Missing Value Imputation and Classification. IEEE Access. https://doi.org/10.1109/access.2025.3548315

Johnson, T. F., Beckerman, A. P., Childs, D. Z., Webb, T. J., Evans, K. L., Griffiths, C. A., Capdevila, P., Clements, C. F. (2024). Revealing uncertainty in the status of biodiversity change. Nature. https://doi.org/10.1038/s41586-024-07236-z

Kim, S., Kim, H., Yun, E., Lee, H., Lee, J., Lee, J. (2023). Probabilistic Imputation for Time-series Classification with Missing Data. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2308.06738

Lee, J., Beretvas, S. N. (2023). Missingness mechanism messiness in meta-analysis: A response to “The importance of plausible data-generating mechanisms in simulation studies”. Research Synthesis Methods. https://doi.org/10.1002/jrsm.1603

Mahajan, P., Uddin, S., Hajati, F., Moni, M. A. (2023). Ensemble Learning for Disease Prediction: A Review. Healthcare. https://doi.org/10.3390/healthcare11121808

Naeem, M., Ozuem, W., Howell, K. E., Ranfagni, S. (2023). A Step-by-Step Process of Thematic Analysis to Develop a Conceptual Model in Qualitative Research. International Journal of Qualitative Methods. https://doi.org/10.1177/16094069231205789

Nagai, R., Honda, T., Satoh, M., Uemoto, Y. (2026). Probabilities of two alleles being identity by state at unobserved loci predicted by observed loci in cattle populations. Scientific Reports. https://doi.org/10.1038/s41598-026-37530-x

Or, C. C., Ng, K. Y. J., Chia, Y., Koh, J. H., Lim, D. Y., Lee, A. L. F. (2023). Face masks are less effective than sunglasses in masking face identity. Scientific Reports. https://doi.org/10.1038/s41598-023-31321-4

Park, S., Jun, S. (2026). Improved Confidence Interval Estimation for Zero-Inflated Count Data Using Transformed Two-Part Bootstrap. AppliedMath. https://doi.org/10.3390/appliedmath6070104

Ribeiro, A. F. (2025). Sample observed effects: enumeration, randomization and generalization. Scientific Reports. https://doi.org/10.1038/s41598-024-80839-8

Riesthuis, P., Otgaar, H. (2025). On the use of receiver operating characteristic area under the curve in eyewitness memory research. Legal and Criminological Psychology. https://doi.org/10.1111/lcrp.12300

Sawadogo, W. F. C., Chassagne, R., Atteia, O. (2026). A comparative benchmark of tabular, sequential, and graph-based models for well-log imputation. https://doi.org/10.5194/egusphere-egu26-11579

Subedi, K., Subedi, N. (2022). Synchronous use of two markers of research: p-value and confidence interval. Journal of Gandaki Medical College-Nepal. https://doi.org/10.3126/jgmcn.v15i2.50531

Taherdoost, H. (2023). Deep Learning and Neural Networks: Decision-Making Implications. Symmetry. https://doi.org/10.3390/sym15091723

Tyralis, H., Papacharalampous, G. (2024). A review of predictive uncertainty estimation with machine learning. Artificial Intelligence Review. https://doi.org/10.1007/s10462-023-10698-8

Weems, C. (2025). Unstable Measures, Unreliable Effects: Re-evaluating Replicability with Reliability Informed Confidence Intervals. https://doi.org/10.21203/rs.3.rs-5937985/v1

White, E., Briggt, M. (2026). Feature Importance Analysis in SAST False Positive Prediction: When do Predefined Scan Rules Mislead the Classifier?. https://doi.org/10.2139/ssrn.7139839

Willard, J., Jia, X., Xu, S., Steinbach, M., Kumar, V. (2022). Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems. ACM Computing Surveys. https://doi.org/10.1145/3514228