Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Reliable imputation of missing modalities in multimodal Alzheimer's disease data using calibrated prediction intervals

Lamyaa Al-Omairi
Melbourne Institute of Technology
Tahsien Al-Quraishi
Melbourne Institute of Technology https://orcid.org/0000-0003-3814-0431
O. S. Albahri
o.albahrey@federation.edu.au
Federation University
Raza Nowrozy
Melbourne Institute of Technology
Zenab Alrikaby
Melbourne Institute of Technology
Naseer Ali Hussien
Al-Ayen Iraqi University
Jemal Abawajy
Deakin University

Research Article

Keywords: Alzheimer's disease, Missing data imputation, Conformal prediction, Uncertainty quantification, Incomplete multimodal data, Group-conditional coverage

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11848076/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Reliable imputation of missing modalities in multimodal Alzheimer's disease data using calibrated prediction intervals

Lamyaa Al-Omairi[1], Tahsien Al-Quraishi[2], O. S. Albahri[3*], Raza Nowrozy[1], Zenab Alrikaby[1], Naseer Ali Hussien[4], Jemal Abawajy[5], for the Alzheimer's Disease Neuroimaging Initiative[†]

[1]School of IT and Engineering (SITE), Melbourne Institute of Technology, Melbourne, Victoria, Australia.
[2]School of Business, Melbourne Institute of Technology, Melbourne, Victoria, Australia.
[3]Institute of Innovation, Science and Sustainability, Federation University, Melbourne, Victoria, Australia.
[4]Department of Computer Engineering Technology, College of Technical Engineering, Al-Ayen Iraqi University (AUIQ), Nasiriyah, Iraq.
[5]Faculty of Science, Engineering and Built Environment, Deakin University, Melbourne, Victoria, Australia.

*Corresponding author(s). E-mail(s): o.albahrey@federation.edu.au;
Contributing authors: lalomairi@academic.mit.edu.au; tquraishi@mit.edu.au; rnowrozy@mit.edu.au; zalrikaby@academic.mit.edu.au; naseerali@alayen.edu.iq; jemal.abawajy@deakin.edu.au;

Abstract

Multimodal Alzheimer's disease (AD) datasets are rarely complete, and imputed biomarker values are usually analysed as if they had been measured. This study proposes reliability-aware low-rank imputation (RELiT), which imputes missing modalities and attaches to every imputed value a prediction interval calibrated separately for each modality and diagnostic group. RELiT combines nuclear-norm low-rank imputation, with diagnosis as an auxiliary covariate, and group-conditional conformal calibration on cross-fitted calibration values that

reproduce the missingness patterns present in the data; intervals of groups with too few calibration values are flagged. RELiT was evaluated on baseline clinical, amyloid positron emission tomography (PET), fluorodeoxyglucose PET, and cerebrospinal fluid data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and on the UCI Multiple Features dataset, under random, diagnosis-dependent, and real ADNI missingness patterns, against five imputation methods and standard and group-calibrated interval methods. RELiT was the most accurate imputer under random missingness. Standard conformal intervals reached nominal coverage on average but systematically under-covered participants with AD, whereas RELiT achieved nominal coverage within every diagnostic group with narrower intervals than all group-calibrated competitors. RELiT maintained reliable coverage for AD when missingness depended on diagnosis and under the real missingness patterns of the cohort. Ablations attributed most of the gain to the calibration design rather than the imputer. Calibration that mirrors how data are missing and is checked within clinically relevant groups is essential for reliable use of imputed multimodal data.

Keywords: Alzheimer's disease, Missing data imputation, Conformal prediction, Uncertainty quantification, Incomplete multimodal data, Group-conditional coverage

1 Introduction

Alzheimer's disease (AD) is a growing public-health challenge: an estimated 7.4 million Americans aged 65 and older were living with Alzheimer's dementia in 2026, and this number is projected to reach about 13 million by 2050 [1]. Characterising the disease requires clinical, imaging, and fluid biomarker data, each capturing a different aspect of its progression [2]. This reflects a broader trend in artificial intelligence (AI) towards multimodal data, which integrates complementary sources of information to build richer representations than any single modality can provide. In practice, however, multimodal datasets are rarely complete. In the Alzheimer's Disease Neuroimaging Initiative (ADNI), only 153 of 780 baseline participants (approximately 20%) had all four of magnetic resonance imaging (MRI), ^{18}F-fluorodeoxyglucose positron emission tomography (FDG-PET), cerebrospinal fluid (CSF), and proteomic measurements [3], and in a separate baseline cohort of 805 participants with MRI, approximately half lacked PET (410 participants) and half lacked CSF (401 participants) [4]. Whole modalities are typically missing for a participant, rather than isolated values, because an examination such as a lumbar puncture or a PET scan was not performed.

Missing values are therefore commonly imputed before analysis, using mean substitution, nearest-neighbour methods [5], chained equations [6], random forests [7], denoising autoencoders [8], or low-rank matrix completion [9, 10]. These methods return a single value for each missing entry, and downstream models then treat the

[†]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

imputed value as if it had been measured. Yet an imputed CSF Aβ_{42} concentration for a participant who never underwent a lumbar puncture is far less certain than a measured one, and its uncertainty varies across modalities and patient groups. If this uncertainty is ignored, imputation errors propagate silently into diagnosis, prognosis, and biomarker studies.

Three gaps limit the reliable use of imputed multimodal data. First, standard imputation methods provide no statement of how reliable each imputed value is; multiple imputation [11, 6] and Bayesian completion [12, 13] quantify uncertainty, but only under modelling assumptions whose calibration is rarely checked. Second, distribution-free conformal methods provide calibrated intervals [14, 15], including for matrix completion [16] and for prediction with missing covariates [17], but they guarantee coverage only on average over all entries, and they calibrate on values hidden at random, whereas real participants often miss several modalities at once. With modality-wise missingness and heterogeneous patient groups, average coverage can hide systematic under-coverage of exactly the participants of greatest clinical interest, such as those with AD, whose biomarker values are the most extreme. Third, many evaluations of multimodal recovery standardise data or build similarity graphs before masking, which leaks information from the hidden values, or recover quantities, such as cross-modal similarity relations, whose recoverability is assumed rather than tested.

The objective of this study is to develop and validate reliability-aware low-rank imputation (RELiT), a framework that imputes missing modalities in multimodal biomarker data and attaches to every imputed value a prediction interval whose coverage is calibrated separately for each modality and each diagnostic group. RELiT combines nuclear-norm low-rank imputation that uses diagnosis as an always-observed covariate, perturbation-derived instability of each imputed value, and group-conditional conformal calibration with feature-specific error scaling on calibration values that are cross-fitted and hidden following the missingness patterns present in the data.

The novelty of this work lies in its calibration design and in its evaluation rather than in the imputation model. Low-rank imputation and conformal prediction are established, but calibrating imputed values on cross-fitted blocks that reproduce the observed missingness patterns, separately for each modality and diagnostic group, yields per-value intervals whose coverage holds for each diagnostic group, including under missingness that depends on diagnosis and under the missingness patterns of the real cohort, while using every observed value for both imputation and calibration. The calibration is independent of the imputation model, so its contribution can be separated from that of the imputer, and the study shows, through a leakage-free protocol, where such reliability statements are and are not attainable.

The framework is evaluated on two datasets. From ADNI, baseline clinical, amyloid PET, FDG-PET, and CSF measurements of 805 participants were used [2, 18, 19]. The Multiple Features dataset of the University of California, Irvine (UCI) Machine Learning Repository, hereafter UCI Multiple Features, contains 2,000 handwritten-digit patterns, each described by six feature sets [20], and serves as a non-clinical benchmark with strong cross-modal structure.

The main contributions of this study are as follows:

1. RELiT, a framework that imputes missing modalities by nuclear-norm low-rank completion with diagnosis as an auxiliary covariate and attaches to each imputed value a prediction interval.
2. Pattern-matched, cross-fitted, group-conditional calibration, which achieves close to nominal coverage within each diagnostic group, including AD, under random, diagnosis-dependent, and real missingness patterns, whereas standard conformal intervals under-cover participants with AD and group-conditional intervals calibrated on a single random split become wide or under-cover when missingness is high; intervals of groups with too few calibration values are flagged.
3. A leakage-free evaluation against five imputation methods, including random-forest and autoencoder imputation, and against standard and group-calibrated interval methods, under simulated and real ADNI missingness patterns, together with diagnostic analyses showing that cross-modal similarity relations in ADNI are essentially unrecoverable and that common reliability signals carry little information about the error of individual imputed values.

The experiments show that RELiT is among the most accurate of the compared imputation methods, that its intervals remove the under-coverage of participants with AD that affects standard calibration, and that they are narrower than group-calibrated competitors when much of the data is missing.

The remainder of this paper is organised as follows. Section 2 reviews related work on imputation, incomplete multi-view learning, uncertainty quantification, and conformal prediction. Section 3 describes the datasets, the proposed framework, the evaluation metrics, and the experimental setup. Section 4 reports the diagnostic analysis of cross-modal relations, imputation accuracy, calibration, reliability signals, the downstream task, and the sensitivity analysis. Section 5 discusses the findings, limitations, and future work, and Section 6 concludes the paper.

2 Related Work

Imputation of missing values has a long history. Simple strategies replace a missing value by the feature mean or by the average of the k nearest neighbors [5]; multivariate imputation by chained equations (MICE) iteratively regresses each incomplete feature on the others [6]; MissForest uses random forests for the same purpose [7]; and denoising autoencoders such as MIDA learn to reconstruct the data from corrupted copies [8]. Low-rank matrix completion exploits correlations across features and samples through nuclear-norm minimization [9], and soft-impute provides a scalable iterative solution by repeated soft-thresholding of singular values [10]; tensor extensions apply the same principle to multi-way data [21]. In multimodal AD research, imputation has been used to fill incomplete MRI, PET, CSF, and proteomic measurements before diagnosis [3, 22]. These methods return point estimates and are evaluated by reconstruction error or downstream accuracy; they do not state how reliable each imputed value is.

Incomplete multi-view learning addresses the same problem from the perspective of representation learning. COMPLETER recovers missing views through contrastive learning and dual cross-view prediction [23], an autoencoder-based method completes the kernel matrix of the incomplete view for AD diagnosis [22], JPLTD decomposes

a tensor of view-specific similarity graphs into low-rank and sparse components for clustering [24], parameter-free refined tensor learning adapts the treatment of singular information [25], and confidence-guided contrastive learning improves incomplete-view clustering [26]. SMIL handles severely missing modalities through Bayesian meta-learning [27]. Several of these methods recover similarity graphs or latent representations rather than measured values, and their recovery is judged only through a downstream task. Whether the cross-view relations being recovered are actually shared between modalities is usually assumed rather than tested.

Uncertainty for incomplete data has been approached in several ways. Multiple imputation quantifies imputation uncertainty through the variability between repeated imputations [11, 6], ensembles provide a simple source of predictive uncertainty [28], and Bayesian CANDECOMP/PARAFAC factorization and Bayesian robust tensor completion yield posterior uncertainty for missing entries [12, 13]. Inferential methods derive entry-wise confidence intervals for noisy matrix and tensor completion under statistical models [29, 30]. In deep multi-view learning, evidential approaches attach view-level uncertainty to classification [31] and propagate it through multiple imputations of missing views [32, 33]; latent reconstruction with uncertainty propagation has been proposed for missing modalities [34]; uncertainty-aware feature completion and three-way decisions have been used to assess or reject completed features [35, 36]; and per-modality aleatoric uncertainty has been modeled for clinical prediction with missing modalities [37]. These estimates are model-based and are rarely checked for calibration, even though predictive uncertainty can be badly miscalibrated [38].

Conformal prediction offers distribution-free calibration. Split conformal prediction with normalized scores produces intervals with finite-sample marginal coverage under exchangeability [14], conformalized quantile regression (CQR) adapts interval width through quantile models [15], Mondrian conformal prediction calibrates separately within categories to obtain category-conditional coverage [39], and weighted conformal prediction restores coverage under covariate shift [40]. Conformalized matrix completion provides distribution-free intervals for the entries of a single partially observed matrix [16], and conformal prediction with missing values targets a response variable when covariates are missing, with coverage conditional on missingness patterns [17]. Cross-conformal prediction and the jackknife+ reuse every training point for calibration through cross-fitting [41, 42]. Selective prediction uses such uncertainty to abstain on unreliable outputs [43, 44, 45]. To our knowledge, conformal calibration has not been applied to the imputed values of modality-wise missing clinical biomarkers with calibration values that reproduce the observed missingness patterns and coverage targeted separately for modalities and diagnostic groups.

Table 1 summarizes representative studies. Existing methods either impute without any reliability statement, provide model-based uncertainty without calibration guarantees, or calibrate a single matrix or a response variable without addressing modality-wise missingness and group heterogeneity in the imputed values themselves. RELiT addresses this combination, while, like the conformal methods it builds on, it assumes that calibration and test values are exchangeable within each group.

Table 1 Comparison with related studies

Study	Method	Datasets	Imp	Unc	Val	Cal	Grp	Blk	Med	Main limitation for reliable imputation
Lin et al. [23] (2021)	Contrastive learning with dual cross-view prediction	Caltech101-20, Scene-15, LandUse-21, Noisy MNIST	✓	×	×	×	×	✓	×	Point recovery of latent views; no reliability
Liu et al. [22] (2021)	Auto-encoder-based kernel completion for incomplete modalities	ADNI	✓	×	×	×	×	✓	✓	Point completion; evaluated only via diagnosis
Han et al. [31] (2021)	Evidential per-view classification with Dempster–Shafer fusion	Handwritten, CUB, PIE, Caltech101, Scene15, HMDB	×	✓	×	×	×	×	×	View-level uncertainty; no imputation
Ma et al. [27] (2021)	Bayesian meta-learning for severely missing modalities	MM-IMDb, CMU-MOSI, avMNIST	✓	×	×	×	×	✓	×	Reliability of recovered modality not quantified
Xie et al. [32] (2023)	Neighbor-based Gaussian multiple imputation with evidential fusion	YaleB, ROSMAP, Handwritten, BRCA, Scene15	✓	✓	×	×	×	✓	✓	Uncertainty serves classification, not imputed values
Gui et al. [16] (2023)	Conformalized matrix completion	Synthetic data; Rossmann sales	✓	✓	✓	✓	×	×	×	Single matrix; entry-wise missingness; marginal coverage
Zaffran et al. [17] (2023)	Conformal prediction of a response with missing covariates	Synthetic data; critical-care data	✓	✓	×	✓	✓	×	✓	Intervals for the response, not for imputed values
Lv et al. [24] (2024)	Projection learning with low-rank graph-tensor decomposition	ORL, NGs, MSRC-v1, BBCSport, RGB-D, UCI, Scene	✓	×	×	×	×	✓	×	Recovers graphs for clustering; no reliability
Chen et al. [33] (2024)	Variational imputation with conflict-aware evidential fusion	YaleB, Handwritten, ROSMAP, BRCA, Scene15	✓	✓	×	×	×	✓	✓	Uncertainty not validated against imputation error
Nguyen et al. [34] (2025)	Latent reconstruction with uncertainty propagation	CMU-MOSI, CMU-MOSEI, MM-IMDb, MMAct, UTD-MHAD	✓	✓	×	×	×	✓	×	Latent-level, uncalibrated uncertainty
Gong et al. [37] (2026)	Per-modality aleatoric uncertainty for graph-based aggregation	MIMIC-IV, eICU	×	✓	×	×	×	✓	✓	Missing modalities not imputed
This study	**Nuclear-norm imputation with pattern-matched, cross-fitted, group-conditional conformal intervals per imputed value**	**ADNI; UCI Multiple Features**	✓	✓	✓	✓	✓	✓	✓	Single clinical cohort; coverage within groups, not per participant

Imp: imputes or recovers missing values or views; Unc: estimates uncertainty; Val: uncertainty for each imputed value; Cal: distribution-free calibrated coverage; Grp: coverage conditional on groups or missingness patterns; Blk: modality- or view-wise missingness; Med: evaluated on biomedical data; ✓ yes; × no.

3 Materials and Methods

This section describes the datasets, the proposed RELiT framework, the evaluation metrics, and the experimental setup. Section 3.1 introduces the datasets, Section 3.2 presents the imputation and calibration procedure, Section 3.3 defines the evaluation metrics, and Section 3.4 specifies the missingness scenarios, parameter settings, and comparators.

3.1 Datasets

Two datasets with complementary properties were used. ADNI provides the clinical setting of interest: four modalities with few features each, unbalanced diagnostic groups, and weak structure shared across modalities. UCI Multiple Features provides a contrasting non-clinical benchmark with many features per modality, balanced groups, and strong cross-modal structure, which shows how the framework behaves when imputation is easier and groups are large.

3.1.1 ADNI

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD) [2]. Data were accessed under the ADNI Data Use Agreement.

Data were assembled from the ADNIMERGE2 data release at the baseline visit. Four modalities were defined: a clinical modality with six measures from the Mini-Mental State Examination (MMSE), the Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-Cog 11 and 13), the Functional Activities Questionnaire (FAQ), and the Clinical Dementia Rating (CDR global score and sum of boxes); an amyloid PET modality with two florbetapir (^{18}F-AV-45) summary standardized uptake value ratios (SUVRs) processed by the University of California, Berkeley, referenced to the whole cerebellum and to a composite reference region; an FDG-PET modality with the mean, maximum, and standard deviation (SD) of uptake in a meta-region of interest (meta-ROI) and in the pons–vermis reference region from the same center; and a CSF modality with Aβ_{42}, total tau, and phosphorylated tau (p-tau$_{181}$) measured by the University of Pennsylvania ADNI Biomarker Core, with tau and p-tau$_{181}$ log-transformed. All participants with the four modalities and a baseline FAQ total were included, so that the true value of every hidden measurement was known. The cohort comprised 805 participants from ADNI-GO (115) and ADNI-2 (690): 255 cognitively normal (CN), 423 with MCI, and 127 with AD (Table 2). Forty-one individual values (0.3%) were missing in the source data; they were left missing and excluded from evaluation. For the downstream analysis, participants with MCI at baseline were

labeled as converters if they received a dementia diagnosis within 36 months (94 participants) and as stable if they had a non-dementia diagnosis at or after 36 months without earlier conversion (247 participants).

Table 2 Baseline characteristics of the ADNI cohort

Characteristic	All	CN	MCI	AD	p
Age, years	72.6 (7.2)	73.1 (6.0)	71.7 (7.4)	74.4 (8.4)	<0.001
Female, n (%)	376 (46.7)	137 (53.7)	188 (44.4)	51 (40.2)	0.017
Education, years	16.3 (2.6)	16.7 (2.4)	16.2 (2.7)	15.7 (2.6)	0.002
APOE $\varepsilon 4$, n (%)	365 (45.3)	75 (29.4)	205 (48.5)	85 (66.9)	<0.001
MMSE	27.6 (2.6)	29.0 (1.2)	28.1 (1.7)	23.1 (2.1)	<0.001
ADAS-Cog 13	15.5 (9.6)	8.9 (4.3)	14.9 (6.8)	30.9 (8.4)	<0.001
CDR sum of boxes	1.50 (1.74)	0.05 (0.16)	1.46 (0.89)	4.55 (1.73)	<0.001
FAQ total	3.6 (5.9)	0.3 (1.2)	2.7 (3.7)	13.3 (7.1)	<0.001
Amyloid SUVR[a]	1.22 (0.26)	1.11 (0.20)	1.22 (0.25)	1.43 (0.24)	<0.001
FDG meta-ROI[b]	1.26 (0.09)	1.30 (0.05)	1.27 (0.07)	1.15 (0.12)	<0.001
CSF $A\beta_{42}$, pg/mL	176 (53)	198 (50)	175 (52)	136 (37)	<0.001
CSF total tau, pg/mL	87 (54)	66 (32)	86 (53)	132 (64)	<0.001
CSF p-tau$_{181}$, pg/mL	42.0 (25.5)	34.8 (17.8)	41.2 (24.4)	59.0 (33.6)	<0.001

Values are mean (SD) unless stated otherwise; p-values are from one-way analysis of variance (ANOVA; continuous variables) or the χ^2 test (categorical variables), using available values. All: n=805; CN: n=255; MCI: n=423; AD: n=127. APOE: apolipoprotein E. [a]Florbetapir SUVR, whole-cerebellum reference. [b]Mean FDG uptake in the meta-ROI.

3.1.2 UCI Multiple Features

The UCI Multiple Features dataset contains 2,000 handwritten-digit patterns, 200 for each digit from 0 to 9, each represented by six feature sets treated as modalities: 76 Fourier coefficients, 216 profile correlations, 64 Karhunen–Loève coefficients, 240 pixel averages, 47 Zernike moments, and 6 morphological features [20]. The dataset is publicly available from the UCI Machine Learning Repository at https://archive.ics.uci.edu/dataset/72/multiple+features. The complete dataset was used.. The digit label defined the calibration groups but, unlike diagnosis in ADNI, was not used as an imputation covariate, because it is the target of the downstream classification task. Table 3 summarizes both datasets.

Table 3 Summary of the datasets

	ADNI	UCI Multiple Features
Samples	805 participants	2,000 digit patterns
Groups	CN 255, MCI 423, AD 127	10 digits × 200
Modalities	4	6
Features per modality	6, 2, 6, 3	76, 216, 64, 240, 47, 6
Cross-modal neighbor overlap (chance)	2.4% (1.6%)	23.3% (0.7%)

Cross-modal neighbor overlap: proportion of k-nearest-neighbor (kNN, $k = 10$) edges in one modality that are also edges in another, averaged over modality pairs; chance level in parentheses.

3.2 The RELiT framework

RELiT takes a data matrix in which whole modalities are missing for some participants and returns, for each missing value, an imputed value and a calibrated prediction interval. It has three stages (Fig. 1): a calibration design that hides observed blocks following the missingness patterns present in the data, in cross-fitted folds, with leakage-free standardization (Section 3.2.1); low-rank imputation with an auxiliary group covariate and perturbation-derived instability (Section 3.2.2); and group-conditional conformal intervals (Section 3.2.3).

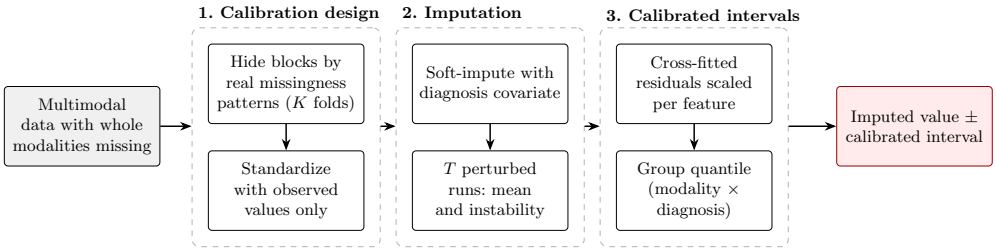


Fig. 1 Overview of RELiT

3.2.1 Calibration design and standardization

Let $X \in \mathbb{R}^{n \times p}$ contain p features of n participants, partitioned into V modalities. The missing values in the (participant, modality) blocks form the test set $\mathcal{T}$, which is imputed and assigned intervals.

Calibration requires imputed values whose true values are known, obtained by hiding observed blocks. If blocks are hidden at random, calibration participants lose one modality while many test participants miss several, so calibration errors are systematically smaller than test errors. RELiT therefore hides blocks following the missingness patterns that actually occur: for every participant i with at least two observed modalities, the pattern of observed modalities of a randomly chosen incomplete participant of the same diagnostic group is drawn, and the modalities of i that this pattern lacks are hidden, provided that at least one modality of i remains observed. Participants are randomly split into K folds; in fold k, the selected blocks of the participants in that fold are hidden and imputed from all remaining data. Every eligible participant thus contributes calibration values exactly once, as in cross-conformal prediction and the jackknife+ [41, 42], so the calibration set $\mathcal{C}$ is several times larger than a single held-out split, and the final imputation of $\mathcal{T}$ uses all observed values.

In every imputation, each feature is centered and scaled using only the values observed under the mask of that imputation, so that no hidden value influences the imputation or its standardization. All errors and interval widths below are expressed in units of the SD of the observed values of each feature unless stated otherwise.

3.2.2 Low-rank imputation and instability

Let Z be the standardized matrix and Ω the set of observed entries. The diagnostic group of each participant, which is recorded for every ADNI participant, is appended as always-observed indicator columns G, giving $\tilde{Z} = [Z, G]$. Missing entries, both test values and the values hidden in each calibration fold, are imputed by soft-impute [10]: starting from zero, the iteration

$$\tilde{Z}^{(t)} = Q \operatorname{diag}(\boldsymbol{s}) R^\top \quad \text{(singular value decomposition)}, \tag{1}$$

$$L^{(t)} = Q \operatorname{diag}\big(\max\{\boldsymbol{s} - \lambda, 0\}\big) R^\top, \tag{2}$$

$$\tilde{Z}_{ij}^{(t+1)} = \begin{cases} \tilde{Z}_{ij}, & (i,j) \in \Omega, \\ L_{ij}^{(t)}, & (i,j) \notin \Omega, \end{cases} \tag{3}$$

is repeated until the relative change in the imputed entries falls below 10^{-4}. The threshold λ is a fixed fraction λ_{rel} of the largest singular value of the zero-filled matrix, and at most r singular components are computed. Soft-thresholding corresponds to nuclear-norm regularization and, unlike hard rank truncation, avoids overfitting when entire modalities are missing.

To measure how stable each imputed value is, the final imputation is repeated T times after adding Gaussian noise with SD ϵ to the observed standardized values. The mean of the T imputations is the final imputed value $\hat{Z}_{ij}$, and their SD σ_{ij} is its instability; the imputations within calibration folds are deterministic.

3.2.3 Group-conditional intervals and reliability signals

For a calibration value $q \in \mathcal{C}$ with feature $j(q)$, the residual is $e_q = \hat{Z}_q - Z_q$, where $\hat{Z}_q$ is the imputation from its fold. Features within a modality differ in how predictable they are, so residuals are scaled by a feature-specific error scale s_j, the mean absolute residual of feature j; for a calibration value, s_j is estimated from the other folds only, and for test values from all folds. Each value is assigned to a group g defined by its modality and the participant's diagnostic group. For group g, with m_g scores $|e_q|/s_{j(q)}$ and miscoverage level α, let $\hat{\tau}_g$ be the $\lceil (m_g + 1)(1 - \alpha) \rceil$-th smallest score. If m_g is too small for this quantile to exist, the quantile of the diagnostic group pooled over modalities is used, and failing that the largest quantile among all groups, which is conservative; intervals obtained in this way are flagged as not calibrated within their own group. Each test value receives the interval

$$\mathcal{I}_q = \Big[\hat{Z}_q - \hat{\tau}_{g(q)}\, s_{j(q)},\ \ \hat{Z}_q + \hat{\tau}_{g(q)}\, s_{j(q)} \Big], \qquad q \in \mathcal{T}, \tag{4}$$

which is transformed back to original units by the feature's mean and SD. If calibration and test values within a group are exchangeable, Mondrian conformal theory guarantees coverage of at least $1 - \alpha$ within each group for a single held-out split [39, 14]; for cross-fitted residuals, the guarantee weakens to $1 - 2\alpha$ in the worst case, although coverage close to $1 - \alpha$ is typical in practice [41, 42]. Pattern matching makes calibration

and test values comparable in which and how many modalities are missing, and grouping by diagnosis keeps them comparable when missingness depends on diagnosis; the coverage guarantee therefore applies to the extent that, within each group, a participant whose blocks are hidden by a drawn pattern is exchangeable with a participant who has that pattern, which the experiments assess empirically.

To examine whether the reliability of individual values can be predicted beyond the group level, a single held-out split was used, in which 10% of observed blocks were hidden at random and divided into two halves $\mathcal{C}_1$ and $\mathcal{C}_2$. Four signals were computed for each imputed value: the instability σ_q, the absolute deviation of $\hat{Z}_q$ from the feature mean, the participant's root-mean-square fit residual on its observed values, and the number of observed modalities. A gradient-boosted regression model fitted on $\mathcal{C}_1$ combined them, with the modality as a categorical input, into a learned reliability score. For the downstream analysis, each participant's reliability was summarized by the mean width of the intervals of their imputed values, and the participants with the widest intervals could be set aside (selective prediction).

3.3 Evaluation metrics

Imputation accuracy on the test values $\mathcal{T}$ was measured by the root mean square error (RMSE) and mean absolute error (MAE) in SD units. Intervals were evaluated by empirical coverage, overall and within each modality, diagnostic group, and feature, and by mean width. Reliability signals were evaluated by the Spearman correlation between the signal and the absolute error, by the excess area under the risk–coverage curve (E-AURC), which is the area under the curve of the MAE of the retained values against the fraction retained minus that of an oracle ranking by the true error [45], so that lower values indicate better ranking, and by the relative reduction in MAE when the 50% most reliable values are retained. Downstream performance was measured by the area under the receiver operating characteristic curve (AUC) for MCI conversion in ADNI and by accuracy for digit classification in UCI Multiple Features. Differences between methods across settings were tested with the two-sided Wilcoxon signed-rank test; p-values are reported without adjustment for multiple comparisons and should be interpreted as descriptive.

3.4 Experimental setup

The experiments were designed to assess imputation accuracy and interval calibration under controlled conditions in which the true value of every missing measurement is known. Missingness was therefore imposed on complete data under three mechanisms of increasing realism, and RELiT was compared with established imputation methods and with standard and group-calibrated interval methods under identical masks.

3.4.1 Missingness scenarios

Modality-wise missingness was simulated at rates of 10%, 30%, 50%, and 70%. Under missingness completely at random (MCAR), each (participant, modality) block was removed independently with the given probability. Under diagnosis-dependent missingness, which is missing at random given diagnosis, the probability was multiplied

by 0.5 for CN, 1.0 for MCI, and 1.5 for AD participants and rescaled to the same overall rate, so that AD participants, whose values are most extreme, were the most likely to have missing modalities. Any participant left with no observed modality had one modality restored. To reproduce real missingness, a third scenario assigned to each participant a pattern of available modalities drawn from the empirical baseline patterns of ADNI-GO and ADNI-2 participants with the same diagnosis (1,181 participants with at least one modality; Table 4). In this scenario, about 23% of participants had only clinical data, and CSF was the modality most often missing. ADNI experiments used ten seeds per rate and UCI Multiple Features five, giving 40 and 20 settings under MCAR; the diagnosis-dependent scenario was run on ADNI at rates of 30%, 50%, and 70% and the real-pattern scenario with ten seeds each. The calibration design was developed in pilot runs on ADNI; to avoid selecting it on the evaluation data, all reported ADNI results use ten seeds, and therefore missingness masks, that were not used during development.

Table 4 Real ADNI missingness patterns

C	A	F	S	CN	MCI	AD	All (%)
•	•	•	•	255	423	130	808 (68.3)
•	○	○	○	119	117	37	273 (23.1)
•	•	•	○	32	34	15	81 (6.8)
Other				6	9	4	19 (1.6)

Baseline patterns of ADNI-GO and ADNI-2 participants. C: clinical; A: amyloid PET; F: FDG-PET; S: CSF; • observed; ○ missing. Four participants with no modality are excluded.

3.4.2 Parameter settings and comparators

RELiT used $\lambda_{\text{rel}} = 0.10$, $\epsilon = 0.04$, $K = 5$ calibration folds, and miscoverage $\alpha = 0.10$ unless stated otherwise, with at most $r = 10$ components and $T = 20$ perturbed imputations for ADNI, and $r = 50$ and $T = 10$ for the larger UCI matrix. These values were fixed before the experiments; their influence was examined in a one-at-a-time sensitivity analysis on ADNI at 50% missingness ($\lambda_{\text{rel}} \in \{0.05, 0.20\}$, $T \in \{5, 10, 40\}$, $\epsilon \in \{0.02, 0.08\}$, $K \in \{3, 10\}$). The learned reliability score used a histogram-based gradient-boosting regressor (scikit-learn [46], 200 iterations, learning rate 0.05).

Imputation was compared with mean imputation; kNN imputation ($k = 10$) [5]; MICE with Bayesian ridge regression (10 iterations; for UCI, the 30 most correlated features per variable) [6]; MissForest-style imputation with chained random forests (100 trees and 5 iterations for ADNI; 10 trees, 2 iterations, and the 30 most correlated features for the larger UCI matrix) [7]; and denoising-autoencoder imputation in the style of MIDA [8], a multilayer perceptron (hidden layers 64–16–64 for ADNI and 128–32–128 for UCI, with at most 200 and 60 training epochs, respectively) trained to reconstruct the data from copies with 25% of modality blocks dropped, refitted

three times. All methods used all observed values and, in ADNI, the same diagnosis covariate.

Interval methods were compared in two families. Standard methods used a single held-out split (10% of observed blocks hidden at random) with one quantile per modality: split conformal intervals around the RELiT, kNN, and MICE imputations; intervals normalised by the instability σ or by the learned score; and CQR with gradient-boosted quantile models [15]. Group-calibrated methods applied the same modality-by-diagnosis groups, per-feature scaling, and fallback as RELiT to MICE and to CQR with the single held-out split; these are the most direct competitors, because they add RELiT's grouping to established methods. Ablations of RELiT used groups by modality only, a single held-out split instead of pattern-matched cross-fitting, MICE in place of low-rank imputation within RELiT's calibration (to separate the contributions of the imputer and the calibration), and no diagnosis covariate or diagnosis groups.

3.4.3 Downstream tasks and graph diagnostic

For ADNI, the 341 labelled MCI participants were used to predict 36-month conversion with ell_2-regularisedd logistic regression on all 17 features, evaluated by five-fold stratified cross-validation, after imputing the originally missing modalities with each imputation method; the complete data served as an upper bound. Selective prediction retained the 80% of participants with the narrowest RELiT intervals and was compared with retaining a random 80%. For UCI Multiple Features, digit classification accuracy was evaluated in the same way.

To test whether cross-modal similarity relations can be recovered, which several incomplete multi-view methods assume, a kNN graph ($k = 10$) was built for each modality, and the proportion of edges in one modality that were also edges in another was computed and compared with chance (Table 3). Recovery of the similarity graph of a missing modality from the other modalities was also attempted directly, with a graph-based low-rank recovery approach, as reported in Section 4.1.

All experiments were implemented in Python with NumPy, SciPy, and scikit-learn and run on two CPU cores.

4 Results

This section first examines whether cross-modal similarity relations can be recovered in ADNI (Section 4.1), which motivates imputing measured values directly. It then reports imputation accuracy (Section 4.2), the calibration of prediction intervals under random missingness (Section 4.3) and under diagnosis-dependent and real missingness (Section 4.4), the reliability of individual imputed values (Section 4.5), downstream prediction (Section 4.6), and the sensitivity analysis, ablation, and runtime (Section 4.7).

4.1 Recoverability of cross-modal similarity relations

Cross-modal neighbourhoods were almost unrelated in ADNI. On average, 2.4% of the kNN edges of one modality were also edges of another, against a chance level

of 1.6%, and the Spearman correlation between dense similarity matrices of different modalities averaged 0.21 (range 0.11–0.34). In UCI Multiple Features, edge overlap was 23.3% against a chance level of 0.7% (Table 3). Consistent with this, recovering the similarity graph of a missing ADNI modality from the other modalities, with a graph-based low-rank recovery approach, did not outperform a simple average of the observed modalities' graphs: across two seeds and four missingness rates, the MAE on true edges was 0.343 for the graph-based recovery, 0.337 for the average, and 0.346 for a consensus initialisation. Intervals for graph entries reached 91% overall coverage, but covered only 9–12% of the true edges, which form 1.7% of the entries; the overall coverage was produced by the near-zero non-edge entries. Cross-modal similarity relations are therefore not a recoverable target in ADNI, and the remaining analyses impute the measured values themselves.

4.2 Imputation accuracy

Under MCAR, RELiT imputed missing ADNI modalities with an RMSE of 0.819 SD units, against 0.845 for MICE ($p = 0.01$), 0.854 for kNN imputation ($p < 0.001$), 0.861 for autoencoder imputation ($p < 0.001$), 0.885 for MissForest ($p < 0.001$), and 0.994 for mean imputation (Table 5). MICE was more accurate at 10% missingness (0.757 against 0.787), but its error increased more steeply with the missingness rate, reaching 0.949 at 70% against 0.853 for RELiT. Gains were largest for the amyloid PET, FDG-PET, and CSF modalities, whereas kNN imputation was the most accurate for the clinical modality (RMSE 0.633 against 0.690). Under diagnosis-dependent missingness, all methods were less accurate; the RMSE of RELiT was 0.959, against 0.946 for kNN imputation ($p = 0.11$), 0.955 for MICE ($p = 0.53$), 1.014 for MissForest ($p < 0.001$), and 1.052 for autoencoder imputation ($p < 0.001$). Under the real ADNI missingness patterns, the RMSE was 0.896 for MICE, 0.915 for RELiT ($p = 0.002$), 0.931 for kNN imputation, 0.954 for autoencoder imputation, and 0.975 for MissForest. In UCI Multiple Features, RELiT had the lowest RMSE overall (0.706 against 0.747 for the best comparator, autoencoder imputation; $p = 0.02$), although autoencoder imputation was more accurate at 10% missingness (0.543 against 0.625), and MICE became unstable at 50% and 70% missingness. RELiT is therefore among the most accurate imputers in every scenario, but its accuracy advantage is not uniform.

Table 5 Imputation error

	ADNI				UCI
Method	**MCAR**	**Dx-dep.**	**Real**	**No dx**	**MCAR**
Mean	0.994 (0.035)	1.138 (0.048)	1.006 (0.021)	0.994 (0.035)	1.003 (0.002)
kNN	0.854 (0.040)	**0.946** (0.029)	0.931 (0.024)	0.902 (0.033)	0.750 (0.093)
MICE	0.845 (0.082)	0.955 (0.090)	**0.896** (0.023)	0.915 (0.079)	1.181 (1.174)
MissForest	0.885 (0.051)	1.014 (0.062)	0.975 (0.027)	0.949 (0.053)	0.879 (0.103)
Autoencoder	0.861 (0.069)	1.052 (0.089)	0.954 (0.022)	0.915 (0.059)	0.747 (0.155)
RELiT	**0.819** (0.040)	0.959 (0.062)	0.915 (0.020)	**0.878** (0.041)	**0.706** (0.068)

RMSE in SD units, mean (SD) across settings. Dx-dep.: diagnosis-dependent missingness; Real: real ADNI patterns; No dx: MCAR without the diagnosis covariate. The large SD of MICE on UCI reflects divergence in some settings. Best value per column in bold.

4.3 Calibration under random missingness

All interval methods reached close to the nominal 90% coverage when averaged over all imputed ADNI values under MCAR (0.891–0.916; Table 6). Averages, however, concealed a systematic difference between diagnostic groups (Fig. 2). Methods with one quantile per modality over-covered CN participants (0.920–0.948) and under-covered participants with AD (0.778–0.819). All methods calibrated by modality and diagnosis reduced this gap, but at different costs in width. RELiT covered 0.912 of CN, 0.908 of MCI, and 0.921 of AD values, with AD coverage between 0.916 and 0.925 at every rate, and its intervals were the narrowest of all methods that reached nominal AD coverage: 15.3% narrower than MICE with group-conditional conformal intervals (2.639 against 3.155 SD units; $p < 0.001$), 18.9% narrower than group-conditional CQR ($p < 0.001$), and 7.7% narrower than MICE with per-modality intervals ($p < 0.001$), which under-covered AD values. The advantage over group-conditional MICE grew with the missingness rate, from 7.0% at 10% to 22.0% at 70% missingness (Fig. 4), because a single held-out split leaves few calibration values for small groups, whereas cross-fitting used every eligible participant (1,435–4,010 calibration values, against 247–615 in the calibration half of a 10% split). Among methods that reached nominal AD coverage, RELiT also had the best worst-feature coverage (0.850 against 0.794 for group-conditional MICE).

The ablations separate the sources of these gains. With groups by modality only, AD coverage fell to 0.777. With a single held-out split instead of pattern-matched cross-fitting, AD coverage was 0.917 and intervals were 10.1% wider ($p < 0.001$). With MICE in place of low-rank imputation inside RELiT's calibration, AD coverage was 0.913 and intervals were 1.5% wider ($p = 0.23$), 9.1% wider at 70% missingness; most of the improvement in interval width therefore comes from the calibration design, and a smaller part from the imputation model. At other miscoverage levels, RELiT covered 0.957 of values at $\alpha = 0.05$ and 0.815 at $\alpha = 0.20$. Coverage of individual features ranged from 0.877 (ADAS-Cog 13) to 0.952 (CDR global). In original units, the mean half-width of the 90% intervals was 2.9 points for MMSE, 10.7 points for ADAS-Cog 13, 1.7 points for CDR sum of boxes, 0.34 for amyloid SUVR (whole cerebellum), 0.11 for FDG meta-ROI uptake, and 76 pg/mL for CSF Aβ_{42}.

In UCI Multiple Features, where the ten-digit groups are large and balanced, RELiT intervals were conservative, covering 0.921 of values, because each calibration fold is imputed from less data than the final imputation, and the strong low-rank structure of these data makes imputation error depend more on the amount of data. They were 9.1% narrower than group-conditional MICE intervals and 14.6% narrower than group-conditional CQR (both $p < 0.001$), but 3.6% wider than RELiT with a single held-out split, which covered 0.894 of values. Pattern-matched cross-fitting is therefore most useful when groups are small or unbalanced, as in ADNI, and a single split suffices when every group is large.

4.4 Diagnosis-dependent and real missingness

When participants with AD were three times as likely as CN participants to have missing modalities, coverage of AD values fell to 0.669–0.761 for per-modality methods

Table 6 Coverage and width of 90% intervals in ADNI

Method	Coverage	Width	CN	MCI	AD	Worst feat.
MCAR						
Per-modality, RELiT imputation	0.905	2.726	0.942	0.920	0.778	0.815
Per-modality, σ-normalized	0.899	2.819	0.933	0.909	0.799	0.787
Per-modality, learned score	0.898	2.710	0.932	0.908	0.800	0.809
Per-modality CQR	0.891	3.066	0.920	0.903	0.797	–
Per-modality, kNN	0.906	2.887	0.948	0.906	0.819	0.818
Per-modality, MICE	0.907	2.883	0.944	0.919	0.792	0.815
Group-cond. CQR	0.902	3.262	0.907	0.902	0.891	–
Group-cond., MICE	0.916	3.155	0.910	0.919	0.920	0.794
RELiT	0.911	**2.639**	0.912	0.908	0.921	0.850
Diagnosis-dependent						
Per-modality, RELiT imputation	0.857	2.716	0.935	0.897	0.687	0.723
Per-modality CQR	0.845	3.056	0.920	0.886	0.669	–
Per-modality, MICE	0.874	2.973	0.931	0.904	0.748	0.758
Group-cond. CQR	0.873	3.471	0.903	0.913	0.731	–
Group-cond., MICE	0.896	3.616	0.914	0.919	0.813	0.749
RELiT	0.899	3.023	0.901	0.908	0.867	0.829
Real patterns						
Per-modality, RELiT imputation	0.894	2.931	0.915	0.890	0.855	0.820
Per-modality CQR	0.909	3.473	0.930	0.902	0.877	–
Per-modality, MICE	0.893	2.877	0.910	0.892	0.856	0.823
Group-cond. CQR	0.918	3.667	0.927	0.913	0.911	–
Group-cond., MICE	0.902	**3.130**	0.909	0.894	0.907	0.798
RELiT	0.901	2.917	0.901	0.902	0.899	0.863

Mean across settings; width in SD units. Per-modality: one conformal quantile per modality with a single held-out split; group-cond.: one quantile per modality and diagnostic group. Bold: narrowest width among methods with AD coverage of at least 0.90.

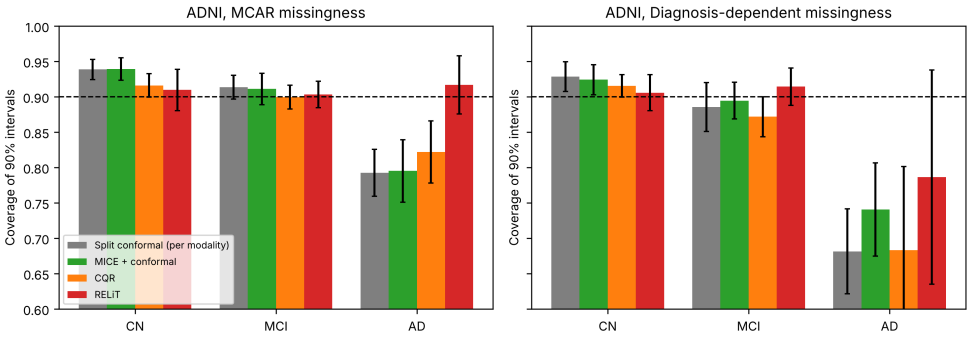


Fig. 2 Coverage by diagnostic group

(Table 6), because calibration values were drawn mostly from CN and MCI participants. Group-conditional calibration with a single held-out split restored AD coverage at 30% missingness (0.917 for MICE) but not at higher rates (0.864 and 0.658), because few observed AD modalities remained to be held out (Table 7). RELiT, which reuses every eligible participant through cross-fitting, covered 0.929 and 0.928 of AD values at 30% and 50% missingness and 0.744 at 70%, against 0.658 for group-conditional MICE and 0.537 for group-conditional CQR. At 70% missingness, most participants with AD retained only one modality and could not contribute calibration values; 19.1% of

test values then received fallback quantiles and were flagged, and 73.9% of the flagged values were covered. Intervals were 14.9% narrower than those of group-conditional MICE ($p < 0.001$). Coverage of CN and MCI values was between 0.901 and 0.908.

Under the real ADNI missingness patterns, in which about 23% of participants had only clinical data, per-modality methods covered 0.855–0.877 of AD values. RELiT covered 0.901 of all values and 0.899 of AD values, with intervals 6.6% narrower than group-conditional MICE ($p = 0.002$) and 20.1% narrower than group-conditional CQR ($p = 0.002$); with MICE inside RELiT's calibration, intervals were 1.7% narrower than with low-rank imputation ($p = 0.004$), consistent with the higher accuracy of MICE in this scenario. No intervals required a fallback.

Table 7 AD coverage under diagnosis-dependent missingness

	30%	**50%**	**70%**
Per-modality conformal (RELiT imputation)	0.754	0.666	0.641
Per-modality, MICE imputation	0.808	0.753	0.683
Group-conditional CQR	0.908	0.749	0.537
Group-conditional, MICE imputation	0.917	0.864	0.658
RELiT with a single held-out split	**0.930**	0.810	0.628
RELiT	0.929	**0.928**	**0.744**
RELiT intervals flagged as fallback	0.0%	2.1%	19.1%

Coverage of AD values by 90% intervals, ten settings per rate; last row: share of RELiT intervals flagged as fallback.

4.5 Reliability of individual imputed values

Beyond the group level, the reliability signals carried little information about the error of individual values (Table 8). In ADNI under MCAR, the perturbation instability σ was weakly negatively correlated with the absolute error ($\rho = -0.14$), so that retaining the 50% most stable values increased the MAE by 8.8%. The learned score was the most informative signal ($\rho = 0.15$, 10.5% MAE reduction), but intervals normalised by it were only 0.6% narrower than unnormalized ones and covered fewer values (0.898; Table 6). Under the real missingness patterns, no signal had $|\rho|$ above 0.05. In UCI Multiple Features, the learned score was more informative ($\rho = 0.35$, 22.2% reduction). Risk–coverage curves are shown in Fig. 3. Per-value reliability in ADNI is therefore expressed adequately by the group-level interval width, and finer reliability scores add little.

4.6 Downstream prediction

Prediction of 36-month MCI conversion was similar after imputation by any method (Table 9): AUC was 0.886 with RELiT, 0.884 with MICE ($p = 0.19$), 0.885 with mean imputation, 0.885 with autoencoder imputation, 0.876 with MissForest, and 0.871 with kNN imputation, compared with 0.923 with complete data. Retaining the 80%

Table 8 Reliability signals

	ADNI			UCI Multiple Features		
Signal	ρ	**E-AURC**	**Reduction**	ρ	**E-AURC**	**Reduction**
Instability σ	−0.14	0.427	−8.8%	0.03	0.331	3.7%
Deviation from mean	−0.05	0.393	−4.1%	−0.16	0.380	−10.7%
Fit residual	0.04	0.348	4.6%	0.12	0.296	7.5%
Observed modalities	−0.04	0.385	−1.1%	0.15	0.259	9.3%
Learned score	0.15	0.293	10.5%	0.35	0.191	22.2%

Spearman ρ with the absolute error; E-AURC (lower is better); MAE reduction when the 50% most reliable values are retained. MCAR, single held-out split.

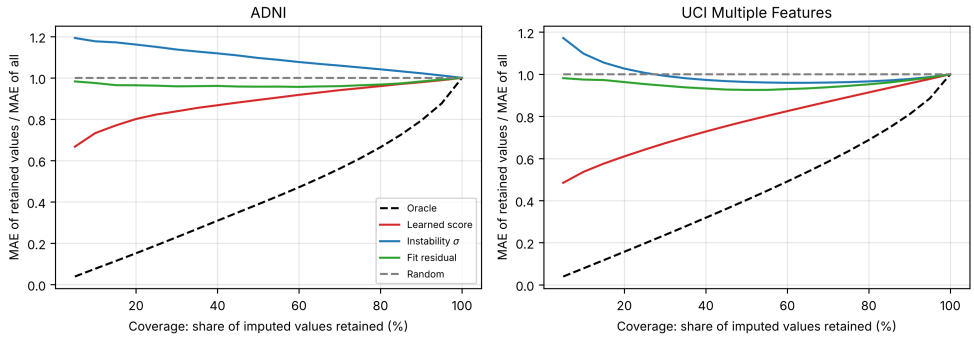


Fig. 3 Risk–coverage curves

of participants with the narrowest intervals did not improve the AUC (0.882 against 0.886 for a random 80%). In UCI Multiple Features, digit classification accuracy was 0.948 with RELiT and 0.943 with MICE ($p < 0.001$), against 0.984 with complete data. Better imputation thus yielded at most small downstream gains, and interval width did not identify participants whose predictions were less accurate.

4.7 Sensitivity analysis, ablation, and runtime

At 50% MCAR missingness in ADNI, RELiT was insensitive to its parameters (Table 10): across the regularization fraction, the number of perturbed imputations, the perturbation size, and the number of folds, RMSE ranged from 0.822 to 0.826, overall coverage from 0.915 to 0.920, AD coverage from 0.916 to 0.925, and mean width from 2.680 to 2.711 SD units. Removing the diagnosis covariate and diagnosis groups increased the RMSE of RELiT from 0.819 to 0.878 and reduced AD coverage from 0.921 to 0.755, similar to the per-modality comparators; even without diagnosis, RELiT remained the most accurate imputation method (Table 5). A complete RELiT run, including the five calibration folds, took 5.9 s for ADNI and 142 s for UCI Multiple Features on two CPU cores, against 38.6 s and 82 s for MissForest.

Table 9 Downstream performance

Imputation	ADNI (AUC)	UCI Multiple Features (accuracy)
Mean imputation	0.885 (0.030)	0.939 (0.040)
kNN imputation	0.871 (0.041)	0.928 (0.054)
MICE	0.884 (0.030)	0.943 (0.041)
MissForest	0.876 (0.038)	0.935 (0.050)
Autoencoder (MIDA-style)	0.885 (0.030)	0.944 (0.043)
RELiT	0.886 (0.029)	0.948 (0.036)
RELiT, narrowest 80%	0.882 (0.032)	0.948 (0.035)
RELiT, random 80%	0.886 (0.029)	0.948 (0.036)
Complete data	0.923 (0.007)	0.984 (0.001)

Mean (SD) under MCAR. ADNI: AUC for 36-month MCI-to-AD conversion; UCI: digit classification accuracy.

Table 10 Sensitivity analysis

Configuration	RMSE	Coverage	Width	AD coverage
Default	0.822	0.918	2.693	0.921
$\lambda_{\text{rel}} = 0.05$	0.822	0.917	2.691	0.921
$\lambda_{\text{rel}} = 0.20$	0.826	0.915	2.705	0.916
$T = 5$	0.822	0.918	2.693	0.921
$T = 10$	0.822	0.918	2.693	0.921
$T = 40$	0.822	0.917	2.693	0.921
$\epsilon = 0.02$	0.822	0.918	2.693	0.921
$\epsilon = 0.08$	0.823	0.917	2.693	0.920
$K = 3$	0.822	0.920	2.711	0.925
$K = 10$	0.822	0.915	2.680	0.918

ADNI, 50% MCAR missingness, five settings per configuration. Default: $\lambda_{\text{rel}} = 0.10$, $T = 20$, $\epsilon = 0.04$, $K = 5$.

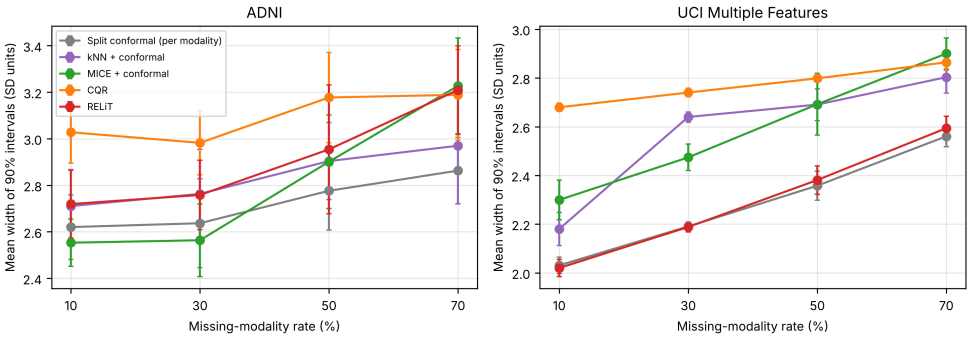


Fig. 4 Interval width by missingness rate

5 Discussion

This study set out to attach calibrated reliability statements to imputed multimodal biomarker values. Four findings stand out. First, standard conformal intervals, whether built around low-rank, kNN, or MICE imputations or obtained by quantile regression, reach close to nominal coverage on average under random missingness while under-covering participants with AD by 8–12 percentage points under random missingness, by up to 23 points when missingness depends on diagnosis, and by 2–5 points under the real ADNI missingness patterns. Because AD participants are those whose biomarker values matter most for diagnosis and trial enrichment, average coverage is a misleading summary in this setting.

Second, how the calibration values are produced matters as much as how they are grouped. Calibrating by modality and diagnosis removes the AD under-coverage, but with a single random held-out split it needs wide intervals, and it fails when missingness is high, because small groups retain few calibration values. Hiding blocks according to the missingness patterns that occur in the data and cross-fitting across participants addressed both problems: RELiT reached nominal AD coverage with intervals 15% narrower than group-conditional MICE and 19% narrower than group-conditional CQR under random missingness, up to 22% narrower at 70% missingness, and it maintained AD coverage at 50% diagnosis-dependent missingness, where single-split group calibration under-covered. The ablations showed that this calibration design, rather than the imputation model, accounts for most of the gain: MICE inside RELiT's calibration gave intervals within 2% of RELiT on average, although 9% wider at 70% missingness. The framework is therefore best viewed as a calibration layer that can be combined with any imputer.

Third, nuclear-norm low-rank imputation with diagnosis as an auxiliary covariate was the most accurate of six imputation methods under random missingness, particularly at high missingness rates and for the imaging and CSF modalities, but not uniformly: MICE was more accurate at 10% missingness and under the real missingness patterns, kNN imputation was better for the clinical modality, and random-forest and autoencoder imputation were less accurate throughout, plausibly because 805 participants and 17 features are too few for flexible models. Downstream conversion prediction was nearly identical across imputation methods, which agrees with the general observation that downstream accuracy is an insensitive measure of imputation quality and supports evaluating imputed values directly.

Fourth, two widely made assumptions did not hold in ADNI. Cross-modal similarity relations were close to chance, so methods that recover a missing modality's similarity graph from other modalities have little shared structure to exploit in such data; an apparently high overall coverage of graph entries was produced by trivially predictable zero entries. Perturbation instability, a common reliability signal, was weakly negatively related to the error of individual values, and no signal, including a learned combination, ranked individual values well enough to make selective use worthwhile. These results caution against treating uncertainty scores as reliability without validating them against the error, and they suggest that in tabular biomarker data the most dependable statement of reliability is a calibrated interval at the level of groups.

Compared with conformalized matrix completion [16], which provides marginal coverage for entries missing individually from a single matrix, RELiT calibrates on hidden modality blocks that reproduce the observed missingness patterns and conditions on modality and diagnosis. Compared with conformal prediction with missing covariates [17], which conditions on missingness patterns to obtain valid intervals for a response, RELiT targets the imputed values themselves. Compared with evidential and variational multi-view methods [32, 33, 34], RELiT trades representational flexibility for coverage that can be checked within clinically relevant groups.

In practice, RELiT intervals can be reported alongside imputed values in cohort studies, used to propagate imputation uncertainty into biomarker analyses, or used to decide whether an imputed CSF or PET value is precise enough for a given purpose; for example, a half-width of about 76 pg/mL for CSF Aβ_{42} indicates that an imputed value is rarely sufficient to classify a participant's amyloid status near a cut-off. Flagged intervals identify groups for which the data do not support a calibrated statement.

5.1 Limitations and future work

This study has several limitations. True values were known only because missingness was imposed on participants with complete data; the real-pattern scenario reproduces the observed patterns, but not necessarily the reasons for missingness, which also depend on study phase, site, and participant burden. Coverage assumes that calibration and test values are exchangeable within each group, and cross-fitting guarantees only $1 - 2\alpha$ in the worst case, although observed coverage was close to $1 - \alpha$ in ADNI; in UCI Multiple Features, where fold imputations are noticeably less accurate than the final imputation, cross-fitted intervals were conservative (coverage 0.921), and a single held-out split gave narrower intervals. At 70% diagnosis-dependent missingness, most participants with AD retained a single modality and could not provide calibration values, so AD coverage reached only 0.744; RELiT flags these intervals but does not repair them, and calibration on external complete cohorts or weighted conformal prediction [40] could address this. Coverage holds for groups, not for individual features or participants: the least-covered feature in a setting averaged 0.850. The reliability of individual values could not be ranked beyond the group level. The diagnostic grouping requires a diagnosis for every participant, which may not be available when imputation is needed; other always-observed variables, such as age or APOE status, could define groups instead. Finally, the analysis used a single, highly educated research cohort with 17 summary features, and the second dataset was non-clinical.

Future work should validate RELiT on independent cohorts, such as the Australian Imaging, Biomarker and Lifestyle study or the National Alzheimer's Coordinating Center, including calibration on one cohort and application to another, extend it to regional imaging features and longitudinal visits, and propagate the calibrated intervals into downstream models through multiple imputation or interval-aware learning.

6 Conclusion

This study introduced RELiT, a reliability-aware framework that imputes missing modalities in multimodal Alzheimer's disease data and attaches to every imputed value a prediction interval calibrated within modalities and diagnostic groups. Its central contribution is a calibration design in which calibration values reproduce the missingness patterns observed in the data and are cross-fitted across participants, so that every observed value contributes to both imputation and calibration.

In ADNI, RELiT was among the most accurate of the six imputation methods evaluated and removed the systematic under-coverage of participants with AD that affects standard conformal intervals, under random, diagnosis-dependent, and real missingness patterns. It achieved this with narrower intervals than group-calibrated alternatives, and the ablation analyses showed that most of this gain arises from the calibration design rather than from the imputation model, so the approach can be combined with other imputers. Where a diagnostic group lacked sufficient calibration data, RELiT flagged the affected intervals rather than reporting them as calibrated, making the limits of its reliability statements explicit.

The study also showed that cross-modal similarity relations in ADNI are close to chance and that commonly used reliability signals carry little information about individual imputation errors. Together, these findings indicate that imputed biomarker values should be accompanied by calibrated uncertainty, and that this calibration should mirror how data are actually missing and be verified within the patient groups of greatest clinical interest, not only on average. Validation in independent cohorts and integration of the calibrated intervals into downstream clinical models are the natural next steps.

Acknowledgements

Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19 AG024904). The grantee organisation is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

Abbreviations

Abbreviation	Definition
$A\beta_{42}$	Amyloid-β 1–42
AD	Alzheimer's disease
ADAS-Cog	Alzheimer's Disease Assessment Scale–Cognitive Subscale
ADNI	Alzheimer's Disease Neuroimaging Initiative
AI	Artificial intelligence
ANOVA	Analysis of variance
APOE	Apolipoprotein E
AUC	Area under the receiver operating characteristic curve
CDR	Clinical Dementia Rating
CN	Cognitively normal
CQR	Conformalized quantile regression
CSF	Cerebrospinal fluid
E-AURC	Excess area under the risk–coverage curve
FAQ	Functional Activities Questionnaire
FDG	^{18}F-fluorodeoxyglucose
FNIH	Foundation for the National Institutes of Health
kNN	k-nearest neighbor
MAE	Mean absolute error
MCAR	Missing completely at random
MCI	Mild cognitive impairment
MICE	Multivariate imputation by chained equations
MMSE	Mini-Mental State Examination
MRI	Magnetic resonance imaging
PET	Positron emission tomography
p-tau$_{181}$	Tau phosphorylated at threonine 181
RELiT	Reliability-aware low-rank imputation
RMSE	Root mean square error
ROI	Region of interest
SD	Standard deviation
SUVR	Standardised uptake value ratio
UCI	University of California, Irvine (Machine Learning Repository)

Declarations

Ethics approval and consent to participate. Not applicable. This study is a secondary analysis of existing, de-identified datasets and did not involve direct recruitment of human participants. The Alzheimer's Disease Neuroimaging Initiative (ADNI) was approved by the institutional review boards of participating institutions, and written informed consent was obtained from participants or their authorised representatives at enrolment. The UCI Multiple Features dataset contains handwritten-digit patterns and does not contain identifiable participant information.

Consent for publication. Not applicable.

Clinical trial number. Not applicable.

Data availability. The UCI Multiple Features dataset is publicly available from the UCI Machine Learning Repository at https://archive.ics.uci.edu/dataset/72/multiple+features. The ADNI data are available to qualified researchers on application at https://adni.loni.usc.edu, subject to the ADNI Data Use Agreement. ADNI participant-level data cannot be redistributed by the authors.

Code availability. The analysis code used in this study is available from the corresponding author on reasonable request.

Competing interests. The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this manuscript.

Funding. This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions (CRediT). **L.A.** and **T.A.**: Conceptualisation, Methodology, Software, Formal analysis, Investigation, Data curation, Visualisation, Writing – original draft, Writing – review & editing. **O.S.A.**, **R.N.**, **Z.A.**, **N.H.**, and **J.A.**: Supervision, Writing – review & editing. All authors read and approved the final manuscript.

Declaration of generative AI and AI-assisted technologies in the writing process. During the preparation of this manuscript, the authors used Claude (Anthropic), a large language model assistant, to assist with code development, experimental workflows, and manuscript drafting and editing. The authors reviewed and edited the AI-assisted content and take full responsibility for the accuracy, integrity, and conclusions of the published work.

References

[1] Alzheimer's Association.: 2026 Alzheimer's Disease Facts and Figures. Available: https://www.alz.org/alzheimers-dementia/facts-figures. Alzheimer's Association, Chicago, IL.

[2] Weiner MW, et al. The Alzheimer's Disease Neuroimaging Initiative: Progress report and future plans. Alzheimers Dement. 2010 May;6(3):202–211.e7. https://doi.org/10.1016/j.jalz.2010.03.007.

[3] Yuan L, Wang Y, Thompson PM, Narayan VA, Ye J. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. NeuroImage. 2012;61(3):622–632. https://doi.org/10.1016/j.neuroimage.2012.03.059.

[4] Dong A, Li Z, Wang M, Shen D, Liu M. High-order Laplacian regularized low-rank representation for multimodal dementia diagnosis. Front Neurosci. 2021;15:634124. https://doi.org/10.3389/fnins.2021.634124.

[5] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520–525. https://doi.org/10.1093/bioinformatics/17.6.520.

[6] van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. J Stat Softw. 2011;45(3):1–67. https://doi.org/10.18637/jss.v045.i03.

[7] Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112–118. https://doi.org/10.1093/bioinformatics/btr597.

[8] Gondara L, Wang K. MIDA: Multiple Imputation Using Denoising Autoencoders. In: Advances in Knowledge Discovery and Data Mining (PAKDD 2018). vol. 10939 of Lecture Notes in Computer Science. Cham: Springer; 2018. p. 260–272.

[9] Candès EJ, Recht B. Exact matrix completion via convex optimization. Found Comput Math. 2009 Dec;9(6):717–772. https://doi.org/10.1007/s10208-009-9045-5.

[10] Mazumder R, Hastie T, Tibshirani R. Spectral regularization algorithms for learning large incomplete matrices. J Mach Learn Res. 2010;11(80):2287–2322.

[11] Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons; 1987.

[12] Zhao Q, Zhang L, Cichocki A. Bayesian CP factorization of incomplete tensors with automatic rank determination. IEEE Trans Pattern Anal Mach Intell. 2015 Sep;37(9):1751–1763. https://doi.org/10.1109/TPAMI.2015.2392756.

[13] Wang X, Yu PLH, Yang W, Su J. Bayesian robust tensor completion via CP decomposition. Pattern Recognit Lett. 2022 Nov;163:121–128. https://doi.org/10.1016/j.patrec.2022.10.005.

[14] Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L. Distribution-free predictive inference for regression. J Am Stat Assoc. 2018;113(523):1094–1111. https://doi.org/10.1080/01621459.2017.1307116.

[15] Romano Y, Patterson E, Candès EJ. Conformalized Quantile Regression. In: Advances in Neural Information Processing Systems. vol. 32; 2019. .

[16] Gui Y, Barber RF, Ma C. Conformalized Matrix Completion. In: Advances in Neural Information Processing Systems. vol. 36; 2023. .

[17] Zaffran M, Dieuleveut A, Josse J, Romano Y. Conformal Prediction with Missing Values. In: Proceedings of the 40th International Conference on Machine Learning. vol. 202 of Proceedings of Machine Learning Research; 2023. p. 40578–40604.

[18] Estarellas M, Oxtoby NP, Schott JM, Alexander DC, Young AL. Multimodal subtypes identified in Alzheimer's Disease Neuroimaging Initiative participants by missing-data-enabled subtype and stage inference. Brain Commun. 2024 Jul;6(4):fcae219. https://doi.org/10.1093/braincomms/fcae219.

[19] Ritter K, Schumacher J, Weygandt M, Buchert R, Allefeld C, Haynes JD. Multimodal prediction of conversion to Alzheimer's disease based on incomplete biomarkers. Alzheimers Dement Diagn Assess Dis Monit. 2015 Jun;1(2):206–215. https://doi.org/10.1016/j.dadm.2015.01.006.

[20] Duin R.: Multiple Features. UCI Machine Learning Repository.

[21] Zhang Z, Aeron S. Exact tensor completion using t-SVD. IEEE Trans Signal Process. 2017 Mar;65(6):1511–1526. https://doi.org/10.1109/TSP.2016.2639466.

[22] Liu Y, Fan L, Zhang C, Zhou T, Xiao Z, Geng L, et al. Incomplete multi-modal representation learning for Alzheimer's disease diagnosis. Med Image Anal. 2021 Apr;69:101953. https://doi.org/10.1016/j.media.2020.101953.

[23] Lin Y, Gou Y, Liu Z, Li B, Lv J, Peng X. COMPLETER: Incomplete Multi-View Clustering via Contrastive Prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 11174–11183.

[24] Lv W, Zhang C, Li H, Jia X, Chen C. Joint Projection Learning and Tensor Decomposition-Based Incomplete Multiview Clustering. IEEE Trans Neural Netw Learn Syst. 2024 Dec;35(12):17559–17570. https://doi.org/10.1109/TNNLS.2023.3306006.

[25] Yang J, et al. Parameter-free multi-view clustering via refined tensor learning. Neurocomputing. 2025 Dec;656:131497. https://doi.org/10.1016/j.neucom.2025.131497.

[26] Chao G, Jiang Y, Chu D. Incomplete Contrastive Multi-View Clustering with High-Confidence Guiding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38; 2024. Available: https://ojs.aaai.org/index.php/AAAI/article/view/29000.

[27] Ma M, Ren J, Zhao L, Tulyakov S, Wu C, Peng X. SMIL: Multimodal Learning with Severely Missing Modality. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35; 2021. .

[28] Lakshminarayanan B, Pritzel A, Blundell C. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In: Advances in Neural Information Processing Systems. vol. 30; 2017. .

[29] Chen Y, Fan J, Ma C, Yan Y. Inference and uncertainty quantification for noisy matrix completion. Proc Natl Acad Sci USA. 2019 Nov;116(46):22931–22937.

https://doi.org/10.1073/pnas.1910053116.

[30] Cai C, Poor HV, Chen Y. Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality. In: Proceedings of the 37th International Conference on Machine Learning. vol. 119 of Proceedings of Machine Learning Research; 2020. p. 1271–1282. Available: https://proceedings.mlr.press/v119/cai20c.html.

[31] Han Z, Zhang C, Fu H, Zhou JT. Trusted Multi-View Classification. In: International Conference on Learning Representations; 2021. .

[32] Xie M, Han Z, Zhang C, Bai Y, Hu Q. Exploring and Exploiting Uncertainty for Incomplete Multi-View Classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. .

[33] Chen M, Huang H, Li Q.: Towards Robust Uncertainty-Aware Incomplete Multi-View Classification. arXiv:2409.06270.

[34] Nguyen DA, Do QH, Doan KD, Do MN.: Are you SURE? Enhancing Multimodal Pretraining with Missing Modalities through Uncertainty Estimation. arXiv:2504.13465.

[35] Wang W, Fu S, Wang J, Liu B, Tang C, Ou W. Uncertainty-aware adaptive feature completion networks for incomplete multi-view learning. Eng Appl Artif Intell. 2026 Jul;176:114676. https://doi.org/10.1016/j.engappai.2026.114676.

[36] Wang H, Fang Y, Wang X, Min F. Three-Way Multimodal Learning with Severely Missing Modalities. Information. 2026 Apr;17(4):384. https://doi.org/10.3390/info17040384.

[37] Gong L, Liu Y, Sun L, Bi Y, Liu J, Zhu X.: Embracing Aleatoric Uncertainty in Medical Multimodal Learning with Missing Modalities. arXiv:2601.21950.

[38] Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, et al. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. In: Advances in Neural Information Processing Systems. vol. 32; 2019. .

[39] Vovk V, Gammerman A, Shafer G. Algorithmic Learning in a Random World. New York: Springer; 2005.

[40] Tibshirani RJ, Barber RF, Candès EJ, Ramdas A. Conformal Prediction Under Covariate Shift. In: Advances in Neural Information Processing Systems. vol. 32; 2019. .

[41] Vovk V. Cross-Conformal Predictors. Annals of Mathematics and Artificial Intelligence. 2015;74(1–2):9–28.

[42] Barber RF, Candès EJ, Ramdas A, Tibshirani RJ. Predictive Inference with the Jackknife+. The Annals of Statistics. 2021;49(1):486–507.

[43] Geifman Y, El-Yaniv R. Selective Classification for Deep Neural Networks. In: Advances in Neural Information Processing Systems. vol. 30; 2017. .

[44] Geifman Y, El-Yaniv R. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In: Proceedings of the 36th International Conference on Machine Learning. vol. 97 of Proceedings of Machine Learning Research; 2019. p. 2151–2159. Available: https://proceedings.mlr.press/v97/geifman19a.html.

[45] Geifman Y, Uziel G, El-Yaniv R. Bias-Reduced Uncertainty Estimation for Deep Neural Classifiers. In: International Conference on Learning Representations; 2019. .

[46] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–2830.