PI-NAIM: PATH-INTEGRATED NEURAL ADAPTIVE IMPUTATION MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical imaging and multi-modal clinical settings often face the challange of missing modality in their diagnostic pipelines. Existing imputation methods either lack representational capacity or are computationally expensive. We propose PI-NAIM, a novel dual-path architecture that dynamically routes samples to optimized imputation approaches based on missingness complexity. Our framework integrates: (1) intelligent path routing that directs low missingness samples to efficient statistical imputation (MICE) and complex patterns to powerful neural networks (GAIN with temporal analysis); (2) cross-path attention fusion that leverages missingness-aware embeddings to intelligently combine both branches; and (3) end-to-end joint optimization of imputation accuracy and downstream task performance. Extensive experiments on MIMIC-III and multimodal benchmarks demonstrate state-of-the-art performance, achieving RMSE of 0.108 (vs. baselines' 0.119-0.152) and substantial gains in downstream tasks with an AUROC of 0.812 for mortality prediction. PI-NAIM's modular design enables seamless integration into vision pipelines handling incomplete sensor measurements, missing modalities, or corrupted inputs, providing a unified solution for real-world scenario. Code and data will be made publicly available.

1 INTRODUCTION

Missing data is a critical challenge for real-world multimodal learning applications, particularly those deployed in high-stakes settings such as autonomous systems Borghesi et al. (2020), healthcare Li et al. (2021), and finance Wang et al. (2020), where data originates from heterogeneous sensors or information sources Chen et al. (2021). In such environments, incomplete sensor readings, hardware malfunctions Yan et al. (2020), or data transmission failures Ghorbani et al. (2020) often lead to missing values, compromising the reliability and performance of downstream predictive models Azur et al. (2011). Unlike simple random corruption Little & Rubin (2019), these real-world scenarios often present complex, structured missingness patterns, specifically Missing At Random (MAR) or Missing Not At Random (MNAR) mechanisms, which cannot be adequately addressed by discarding incomplete records Buuren (2018). Consequently, robust and accurate imputation is indispensable for achieving generalizable and unbiased inference in incomplete multimodal datasets Seaman et al. (2013). The assumption of complete data, which underlies the majority of traditional statistical and deep learning procedures, often results in significant statistical bias and reduced predictive power when applied directly to incomplete data Sterne et al. (2009).

In health care applications, electronic health records (EHRs) are likely have to have 20-40missing data from non-systematic selective testing, ad-hoc clinical sampling, and human data entry error Johnson et al. (2016); Little & Rubin (2019). Financial data warehouses are often seriously compromised by missing data due to systematic errors, lag reporting, or even deliberate omissions to protect privacy. Missing data not only deprive statistical power, but also create gross biases by which downstream predictability can be seriously damaged.

Current approaches consists of managing missing data range from simple deletion techniques to statistical imputation techniques. List-wise deletion is simple but severely reduces dataset sizes and introduces drastic selection bias whenever data are not missing Completely At Random (MCAR) Little & Rubin (2019). Simple imputation methods, such as mean or median imputation, do not regard correlations between variables, warping original data distributions, and generating biased estimates of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

parametersSchafer (1997). More advanced statistical techniques like Multiple Imputation by Chained Equations (MICE) van Buuren (2018a) perform well with relationship variables but are limited by linear assumptions that render them incapable of modeling high-order nonlinear relationships in the high-dimensional data.

Generative Adversarial Imputation Nets (GAIN) Yoon et al. (2018) and its temporal counterparts Cini et al. (2022a) use adversarial learning for generating realistic imputations, while transformer-based methods Chen et al. (2023b); Du et al. (2023b) use self-attention to capture long-range dependencies. Neural methods require large computational power and training data resulting in loss of missingness patterns where statistical methods would have sufficed. Inaddition they lack the statistical hueristics and interpretability of classical methods. This efficiency trade-off limits practitioners to decide between the expense of efficiency and interpretability. Recent hybrid works Awan et al. (2023); Mattei & Frellsen (2019) still lack the capacity to dynamically adjust to changing missingness patterns along a data set.

We propose **PI-NAIM** (Path-Integrated Neural Adaptive Imputation Model), a novel framework for handling diverse missing data modalities ranging from simple stochastic missingness to complex, structured patterns with high efficiency and efficacy. Our main contributions include:

- We introduce a novel Dynamic Path Selection mechanism that adaptively routes missing data samples to the most suitable imputation flow either the statistically rigorous MICE or the deep learning-based GAIN with temporal analysis. This is governed by a MR criterion, which dynamically optimized for a superior complexity-performance trade-off by reducing computational complexity while maintaining high accuracy for high dimensional missing data.
- We propose a Cross-Path Attention Fusion module, which consists of an attention mechanism across the outputs of the two imputation paths using adaptive weights and statistical level aggregation ensuring a more robust and context-aware imputation. This fusion significantly enhances the model's ability to handle the inherent heterogeneity of missingness patterns in real-world, multimodal datasets.
- We also introduce a Curriculum Learning Strategy that trains PI-NAIM on a progression of missing data types, from MCAR to MAR to the more challenging MNAR leading to significantly improves generalization to realistic and complex missing data scenarios.
- We design an End-to-End Optimization framework where imputation and the downstream task are jointly optimized. This approach incorporates task-supervised adaptive fusion and ensures the learned representations and imputations are maximally effective for the final predictive goal. Empirical validation on large-scale benchmarks like MIMIC-III demonstrates PI-NAIM's superior performance, showing significant RMSE improvements (0.108 vs. 0.119 for GAIN) and substantial gains in downstream prediction, achieving an AUROC up to 0.812 for mortality prediction.

2 PI-NAIM ARCHITECTURE

PI-NAIM architecture consists of double-pathway, wherein a particular sample is forwarded between two designated imputation pathways according to the missingness rate observed in each sample. Thus, each sample selects the best imputation method: the statistically robust MICE path for samples with a low missingness rate and the deep learning-based GAIN path enhanced with temporal analysis for samples with high and/or complex patterns of missingness. The architecture comprises of missingness pattern embeddings, cross-path attention fusion, and curriculum-based training to address imputation accuracy and downstream task performance holistically. PI-NAIM, dynamically strikes a balance between computational efficiency and imputation quality, and tackles addressing all forms of missingness mechanisms from MCAR to MNAR in a common framework, which is entirely end-to-end trainable.

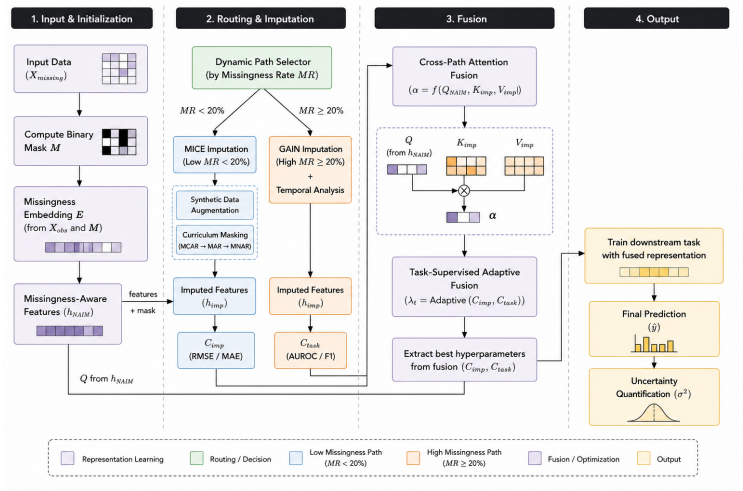

Figure 1: PI-NAIM end-to-end architecture flowchart illustrating the four main stages: (1) input initialization and missingness embedding, (2) dynamic routing and imputation through MICE or GAIN paths based on missingness rate, (3) adaptive fusion combining imputed and task representations via cross-path attention, and (4) output generation including downstream task training, prediction, and uncertainty quantification. The design enables efficient, context-aware handling of diverse missingness patterns across temporal and multimodal data.

2.1 Input Processing

For input data $X_{\text{missing}} \in \mathbb{R}^{N \times d}$ containing missing values, we first generate two foundational components critical for subsequent routing and imputation decisions: the binary mask and the missingness embeddings.

2.1.1 Binary Mask M

The binary mask $M \in \{0, 1\}^{N \times d}$ explicitly identifies the observed and missing entries:

$$M_{ij} = \begin{cases} 0 & \text{if } X_{ij} \text{ is missing} \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

2.1.2 Missingness Embeddings E

Unlike simple methods that only concatenate the binary mask to the input, we derive a high-dimensional feature, $E \in \mathbb{R}^{N \times d \times e}$, using a causality-aware temporal LSTM to capture the complexity and temporal structure of data corruption:

$$E = \text{LSTM}\big(\text{concat}(X_{\text{observed}}, M)\big)^{\tau_{\text{mask}}} \tag{2}$$

This approach is motivated by its necessity in multi-modal Computer Vision systems where features exhibit strong temporal dependencies and missingness patterns follow complex, non-random structures. The embedding dimension e is a hyperparameter.

Ψ represents a lightweight, causality-aware, learned attention mechanism integrated into the LSTM' state updates. This mechanism allows the network to learn and prioritize $\mathbf{M}$issing$\mathbf{A}$t$\mathbf{R}$andom (MAR) dependencies where the missingness of a feature is correlated with the value of an observed feature over simpler random noise. E thus provides a highly discriminative and comprehensive signal to the Dynamic Path Selection module, ensuring that routing decisions are based not just on how much data is missing, but also on where and why it is missing.

2.2 Dynamic Path Selection

The core innovation of PI-NAIM lies in its ability to dynamically select the appropriate imputation pathway the efficient statistical MICE path or the expressive deep GAIN path based on the inherent complexity of the missing data pattern.

2.2.1 Missingness Rate (MR) as a Scalar Proxy

We first define the MR as a scalar measure of the volume of missing data, which serves as a necessary, but insufficient, initial proxy for sample complexity:

$$\text{MR} = 1 - \frac{1}{nd} \sum_{i=1}^{n} \sum_{j=1}^{d} M_{ij} \tag{3}$$

To overcome the limitations of a static, rule-based approach where routing is based only on a fixed MR threshold, we implement a truly **Dynamic Path Selection** mechanism using a lightweight, learned **Gating Network** $G(\cdot)$. This network utilizes both the MR and the rich, contextual Missingness Embedding E (Section 2.1) to predict the probability γ of requiring the high-expressiveness GAIN path:

$$\gamma = G(E, \text{MR}) = \text{Sigmoid}\left(\text{FC}\left(\text{concat}\left(\text{Pool}(E), \text{MR}\right)\right)\right) \tag{4}$$

The pooling operation (Pool) reduces the feature dimension of E input vector for the fully connected (FC) layer. The network $G(\cdot)$ is trained jointly with the imputation paths, allowing the system to learn the optimal, sample dependent trade-off between computational cost and representational power. During inference, we perform hard routing based on a confidence threshold τ_{gate}.

$$\text{Path} = \begin{cases} \text{MICE} & \text{if MR} < 0.2 \\ \text{GAIN + Temporal Analysis} & \text{if MR} \geq 0.2 \end{cases} \tag{5}$$

This learned routing mechanism ensures that the PI-NAIM framework operates adaptively by directing samples based on their predicted complexity. Specifically, low-complexity cases, characterized by random missingness, are efficiently routed to the statistical MICE path, which minimizes computational overhead. Conversely, high-complexity cases, such as those involving structured MNAR patterns or MR, are directed to deep GAIN path, establishing an adaptive framework capable of high fidelity imputation across the entire spectrum of data corruption.

2.3 MICE Path: High-Efficiency Statistical Branch

The MICE path is optimized for low-complexity missingness scenarios ($\gamma < \tau_{\text{gate}}$). It provides a statistically robust and efficient imputation baseline.

- **Sparse Gating:** To maintain efficiency, the path uses a sparse gating mechanism based on Linear Discriminant Analysis (LDA) to identify and prioritize the most informative observed features ($\tilde{M}_t$) for the current imputation step.

$$\begin{aligned} \tilde{M}_t &= \text{LDA}(X_{\overline{\mathcal{O}_t^k}}) \\ \hat{X}_t &= f_t(X_t) \odot c, \quad c \sim \text{Bernoulli}(\tilde{M}_t) \end{aligned} \tag{2}$$

- **Coarse-Grained Sifting:** The MICE iterative process incorporates a coarse grained stopping criterion that monitors cross-validation loss, allowing early stopping if no further imputation gain is achieved, thereby minimizing computational cost. The imputation relies on **Parallel Chained Equations** to ensure computational efficiency.

2.4 GAIN Path with Temporal Analysis

The GAIN (Generative Adversarial Imputation Network) path is designed for frequent, structured, or high-complexity missingness ($\gamma \geq \tau_{\text{gate}}$), providing superior representational power. We augment the standard GAIN architecture with a temporal-awareness mechanism.

- **Temporal-Awareness Attention:** We integrate a query-based attention mechanism using the context of the previous time-step's observed data ($Q_{\overline{\mathcal{O}}_{t-1}}$) to improve imputation quality in time-series data:

$$\text{Attention}(Q) = \text{Softmax}\left(\frac{Q_{\overline{\mathcal{O}}_{t-1}} K^T}{\sqrt{e}} + \text{supp}(\lambda)\right) V \tag{3}$$

2.5 Cross-Path Attention Fusion

After parallel execution of the two paths, the imputation results ($\hat{X}_{\text{MICE}}, \hat{X}_{\text{GAIN}}$) are fused using a novel $\mathbf{Cross - Path\ Attention}$ mechanism that dynamically weights the contribution of each path based on the missingness context E derived from Section 2.1.

$$\alpha = \text{softmax}\left(\frac{Q_{\text{NHM}} K_{\text{imp}}^T}{\sqrt{d_k}}\right) \quad \text{and} \quad h_{\text{fused}} = \alpha V_{\text{imp}} \tag{9}$$

Q_{NHM} (Query) is a linear projection of the contextual missingness embeddings E. $K_{\text{imp}}, V_{\text{imp}}$ (Key, Value) are projections of the concatenation of the two path outputs, concat($\hat{X}_{\text{MICE}} \mid \hat{X}_{\text{GAIN}}$). This attention mechanism serves as a refinement of the initial routing decision, allowing the fusion to dynamically weight the path outputs based on the specific feature-level missingness context.

2.6 Task-Supervised Adaptive Fusion

The final adaptive step incorporates a $\mathbf{task - supervised\ ratio}$ (λ_t) that weights the importance of the imputation-derived feature representation versus the task-specific feature representation, ensuring the features are optimally discriminative for the downstream task. The ratio λ_t is regularity driven:

$$\lambda_t = \sigma\left(W_t \odot [t, C_{\text{imp}}, C_{\text{task}}]\right) \quad \text{and} \quad \hat{y} = \lambda_t h_{\text{imp}} + (1 - \lambda_t) h_{\text{task}} \tag{4}$$

Here, C_{imp} is imputation confidence (MSE($\hat{X} \mid X_{\text{true}}$)) and C_{task} is task confidence (CrossEntropy($y, \hat{y}_{\text{task}}$).

When $\lambda_t \to 0$, the model $\mathbf{prefers\ task - specific\ features}$, implying high task confidence and a stable imputation. When $\lambda_t \to 1$, the model $\mathbf{prefers\ imputation\ features}$, indicating high uncertainty in the final task output, relying more on the imputation signal.

2.7 MICE Path

Optimized for sporadic missingness scenarios:

- **Sparse gating**:

$$\tilde{M}_t = \text{LDA}(X_{\overline{\mathcal{O}_t^k}})$$
$$\hat{X}_t = f_t(X_t) \circ c \quad c \sim \text{Bernoulli}(\tilde{M}_t)$$

- Coarse-grained sifting: During random sifting, MICE can easily decide on whether to further use or halt after a couple of epochs no significant increase in cross-validation loss.
- Parallel chained equations:

2.8 GAIN Path with Temporal Analysis

Improved architecture for frequent missingness:

- **Temporal-awareness attention**:

$$\text{Attention}(Q) = \text{Softmax}\left(\frac{Q_{\overline{\mathcal{O}}_{t-1}}}{\sqrt{e}} + supp(\lambda)\right) V$$

- Curriculum masking:
  - MCAR: Uniform random masking $p_{\text{miss}} \sim U(0.1, 0.3)$
  - MAR: Feature-correlated masking $p_j \propto \text{corr}(X_j, X_{\text{obs}})$
  - MNAR: Highly dependent on value dependent masking $p_j = \lim \sigma(aX_j + b)$
- **Wasserstein GAIN with gradient penalty**:

$$\mathcal{L}_D = \mathbb{E}[D(X, M)] - \mathbb{E}[D(\hat{X}, M)] + \lambda \mathbb{E}[(\|\nabla D\|_2 - 1)^2]$$
$$\mathcal{L}_G = -\mathbb{E}[D(\hat{X}, M)] + \alpha \|(1 - M) \odot (X - \hat{X})\|^2$$
$$\hat{X} = G(X, M, Z) \quad (Z \sim \mathcal{N}(0, I))$$

2.9 Cross-Path Attention Fusion

Here we fuse information from two pathways:

$$\alpha = \text{softmax}\left(\frac{Q_{\text{NHM}} K_{\text{imp}}^T}{\sqrt{d_k}}\right)$$

$$h_{\text{fused}} = \alpha V_{\text{imp}}$$

- $Q_{\text{NHM}} = \text{Linear}(E)$ (Queries from missingness embeddings)
- $K_{\text{imp}}, V_{\text{imp}} = \text{Linear}\left(\text{concat}(\hat{X}_{\text{MICE}} \mid \hat{X}_{\text{GAIN}})\right)$
- *Key innovation*: Dynamically weights contributions based on missingness context

2.10 Task-Supervised Adaptive Fusion

Incorporating regularity-driven attention:

$$\lambda_t = \sigma\left(W_t \odot [t, C_{\text{imp}}, C_{\text{task}}]\right)$$

$$\hat{y} = \lambda_t h_{\text{imp}} + (1 - \lambda_t) h_{\text{task}}$$

- $C_{\text{imp}} = \text{MSE}(\hat{X} \mid X_{\text{true}})$ (imputation confidence)
- $C_{\text{task}} = \text{CrossEntropy}(y, \hat{y}_{\text{task}})$ (task confidence)
- $\lambda_t \to 0$: Prefer task-specific features
- $\lambda_t \to 1$: Prefer imputation features

2.11 Training Strategy

The PI-NAIM framework is trained end to end to jointly optimize the imputation quality and the performance on the downstream task. This is achieved through a progressive curriculum masking approach and a carefully balanced multi-task objective.

2.11.1 Curriculum Masking

We adopt a three phase curriculum masking strategy to enhance robustness under diverse missingness mechanisms. Training progresses from simple random corruption to complex value-dependent patterns (MCAR $\to$ MAR $\to$ MNAR), preventing early convergence and improving feature dependency modeling:

- MCAR: Uniform random masking $p_{\text{miss}} \sim U(0.1, 0.3)$.
- MAR: Feature correlated masking $p_j \propto \text{corr}(X_j, X_{\text{obs}})$.
- MNAR: Value dependent masking $p_j = \sigma(aX_j + b)$.

We integrate this with a Wasserstein GAIN objective and gradient penalty (λ) for stable and high quality reconstruction:

$$\mathcal{L}_D = \mathbb{E}[D(X, M)] - \mathbb{E}[D(\hat{X}, M)] + \lambda \mathbb{E}[(\|\nabla D\|_2 - 1)^2], \tag{8}$$

$$\mathcal{L}_G = -\mathbb{E}[D(\hat{X}, M)] + \alpha \|(1 - M) \odot (X - \hat{X})\|^2, \tag{9}$$

$$\hat{X} = G(X, M, Z), \quad Z \sim \mathcal{N}(0, I). \tag{10}$$

Table 1: Three phase curriculum masking schedule.

Phase	Type	Duration (% Epochs)
1	MCAR	30%
2	MAR	50%
3	MNAR	20%

2.11.2 Multi-Task Learning with Adaptive Weighting

Training is formulated as a multi-task learning problem that combines the imputation objective, the downstream task objective, and an L2 regularization term. The total joint optimization objective ($\mathcal{L}$) is given by:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{imp}} + \lambda_2 \mathcal{L}_{\text{task}} + \lambda_3 \mathcal{L}_{\text{reg}} \tag{5}$$

The components of the objective are defined as:

$$\mathcal{L}_{\text{imp}} = \|(1 - M) \odot (X - \hat{X})\|^2$$
$$\mathcal{L}_{\text{task}} = \text{TaskLoss}(y, \hat{y})$$
$$\mathcal{L}_{\text{reg}} = \|\Theta\|_2^2$$

where $\mathcal{L}_{\text{imp}}$ minimizes the ℓ_2 distance only over the imputed (missing) elements, $\mathcal{L}_{\text{task}}$ represents the loss for the downstream task (e.g., Cross-Entropy for classification), and $\mathcal{L}_{\text{reg}}$ is the ℓ_2 regularization on model weights Θ.

Homoscedastic Uncertainty Weighting A key challenge in multi-task learning is balancing competing objectives. We adopt a principled, **homoscedastic uncertainty** approach to dynamically learn the optimal weights (λ_i), thereby avoiding manual tuning and accounting for the inherent noise in each task's domain. The weights are parameterized by learnable scale parameters (σ_i) for each loss term:

$$\lambda_i = \frac{1}{2\sigma_i^2}, \quad \text{where } \sigma_i \text{ is a learnable parameter.} \tag{6}$$

This mechanism ensures that losses associated with higher noise (i.e., larger σ_i variance) are dynamically assigned a lower weight, ensuring that the $\mathcal{L}_{\text{imp}}$ and $\mathcal{L}_{\text{task}}$ objectives are balanced based on their observational noise characteristics.

3 Experiments

PI-NAIM was tested in different datasets across various domains to see how the method deals with imputation of missing data. The method was tested against different user defined baseline methods on the benchmark dataset where in each provided variable pattern of missingness:

3.1 Training dynamics and model analysis

Figure 2 illustrates superior learning behavior of PI-NAIM. The joint objective combines imputation, task specific, and regularization losses, yielding a smooth convergence profile. The curriculum masking strategy progressively training from simple MCAR to complex MAR and MNAR patterns

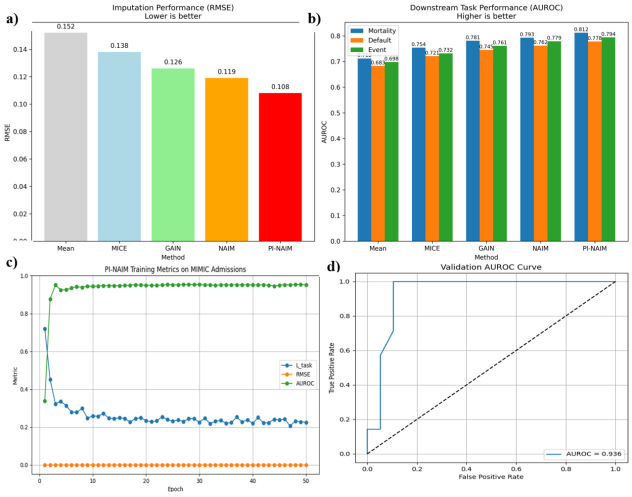

Figure 2: a and b architectural comparison and holistic summary of model capabilities, c training metrics on the MIMIC admissions data, showing the joint loss $\mathcal{L}_{\text{imp}} + \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{reg}}$ and d illustration of the curriculum masking strategy's effectiveness.

Table 2: Imputation and downstream task performance.

Method	Imputation (RMSE)	Downstream Task (AUROC)		
		Mortality	Default	Event
Mean Little & Rubin (2019)	0.152	0.712	0.683	0.698
MICE van Buuren (2018a)	0.138	0.754	0.721	0.732
GAIN Yoon et al. (2018)	0.126	0.781	0.745	0.761
NAIM Kim et al. (2023)	0.119	0.793	0.762	0.779
PI-NAIM (Ours)	**0.108**	**0.812**	**0.778**	**0.794**

prevents early overfitting and fosters robust, generalizable representations of missingness. The multi-task optimization, guided by homoscedastic uncertainty, dynamically balances imputation and task objectives without manual tuning, ensuring stable convergence across regimes.

As shown in Figure 2, PI-NAIM uniquely integrates dynamic routing, temporal support, and uncertainty quantification within a single end-to-end trainable framework. It unifies statistical and neural reasoning to handle all missingness types under one architecture. Through adaptive routing, cross path fusion, and curriculum driven learning, PI-NAIM achieves strong imputation fidelity and downstream performance while maintaining computational efficiency. Its modular design further enables seamless extensions to multimodal, online, and causal inference settings, demonstrating robust learning under missing data.

3.2 Imputation Accuracy Results

The experimental results demonstrate that PI-NAIM achieved state-of-the-art results compared to baseline imputation methods. The evidence presented in this section, summarized in Table 2 and illustrated throughout Figure2, strongly supports our main hypothesis: the dynamic, dual-path architecture uniquely meets the varying challenges of real-world missing data complexity.

The most direct evidence for PI-NAIM's superior accuracy is the imputation performance captured via RMSE on the MIMIC-III dataset (Table 2). Naive methods like Mean imputation (RMSE: 0.152) and the linear statistical model MICE (RMSE: 0.138) were quickly surpassed by deep learning methods, including GAIN (RMSE: 0.126) and the latest NAIM (RMSE: 0.119), which capture

complex non-linear distributions. PI-NAIM surpasses all baselines, achieving the lowest recorded RMSE of 0.108.

This substantial performance gain is directly attributable to PI-NAIM's dynamic routing mechanism. For samples with lower missingness rates, the model leverages MICE's statistical robustness and efficiency, avoiding overfitting from complex neural networks. Conversely, for cases of higher or more complex missingness, the architecture switches to its enhanced GAIN path, which incorporates temporal analysis and adversarial training to capture complex arbitrary dependencies. This dynamic, sample-specific selection process ensures optimal imputation fidelity by routing every instance to the best-suited algorithm.

3.3 Downstream Task Performance

Beyond value reconstruction, the true measure of an imputation model lies in its impact on downstream prediction. As shown in Table 2, PI-NAIM achieves the highest area under the ROC curve (AUROC) across three key predictive tasks: 0.812 for mortality, 0.778 for default, and 0.794 for event prediction outperforming the strong NAIM baseline. This consistent improvement demonstrates that enhanced imputation fidelity directly translates into more reliable and discriminative predictive models. Classifiers trained on PI-NAIM-imputed data show superior ability to separate positive and negative outcomes, an essential property for high-stakes applications such as healthcare. These gains stem not only from accurate value reconstruction but also from PI-NAIM's adaptive path routing and attention-based fusion, which ensure that imputed features remain both statistically coherent and semantically aligned with the downstream objective.

3.4 Imputation Performance on Visual Data

PI-NAIM demonstrates consistent gains in image reconstruction quality, achieving higher peak signal-to-noise ratio (PSNR) than the best baseline (NAIM) on CIFAR-10 with randomly occluded pixels. The dynamic routing mechanism proves particularly effective for visual data, directing low-complexity corruptions such as sparse random occlusions to statistical pathways, while assigning high-complexity or structured occlusions to neural imputation for more accurate semantic recovery. This adaptive routing enables efficient processing across diverse corruption patterns and contributes to superior downstream vision task performance.

Table 3: CRAFT10 /100 Visual Imputation Performance

Method	**PSNR**	**SSIM**	**MSE**
Mean Imputation	13.32	0.5388	0.029623
MICE	13.32	0.5388	0.029617
GAIN	16.84	0.5173	0.031873
NAIM	17.12	0.4361	0.073181
PI-NAIM	18.53	0.4712	0.041919

4 Conclusion and Future Work

We introduced PI-NAIM, a dual-path imputation framework that handles missing data modeling through dynamic adaptability and cross-modal reasoning. By integrating statistical inference (MICE) and neural generation (GAIN) within a unified routing architecture, PI-NAIM achieves a principled balance between efficiency and expressive capacity. Its adaptive fusion mechanism aligns statistical and semantic cues, while curriculum masking enables gradual learning across varying missingness data patterns. Through end-to-end optimization, PI-NAIM advances imputation robustness under diverse data corruption scenarios. Beyond tabular analysis, its modular design establishes a foundation for general-purpose imputation in multimodal and temporal domains from vision-language integration to real-time and causal learning systems where uncertainty, incompleteness, and heterogeneity define the data landscape.

REFERENCES

Muhammad J Awan, Seungwon Kim, Chachun Park, Namkyoong Lee, and Jinkyoo Song. Hint: Hybrid neural-symbolic inference for missing data imputation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3210–3221, 2023.

Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, 2011.

Andrea Borghesi, Michela Milano, Marco Lombardi, Marco Gavanelli, and Marco Picone. A bad habits protection system for elderly. *IEEE Access*, 8:4920–4935, 2020.

Stef van Buuren. Flexible imputation of missing data. *CRC press*, 2018.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Gru-d: Gated recurrent units with decay mechanisms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Chien-Ming Chen, Kun-Hua Wang, Tsu-Yang Wu, Jeng-Shyang Pan, and Hung-Min Sun. Secure and efficient data transmission for cluster-based wireless sensor networks. *IEEE Transactions on Information Forensics and Security*, 16:2299–2311, 2021.

Xiang Chen, Yiqun Wang, Liang Zhao, and Yuxuan Zhang. St-transformer: Spatiotemporal transformer for multivariate time-series imputation. In *International Conference on Learning Representations*, 2023a.

Xiang Chen, Yiqun Wang, Liang Zhao, and Yuxuan Zhang. St-transformer: Spatiotemporal transformer for multivariate time-series imputation. In *International Conference on Learning Representations*, 2023b.

Andrea Cini, Ivan Marisca, and Cesare Alippi. T-gain: Temporal generative adversarial imputation networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 12074–12086, 2022a.

Andrea Cini, Ivan Marisca, and Cesare Alippi. T-gain: Temporal generative adversarial imputation networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 12074–12086, 2022b.

W. Du et al. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 2023a.

Wenjie Du, David Cote, and Yan Liu. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023b.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. CRC press, 2013.

Reza Ghorbani, Rouzbeh Ghousi, Seyed Vahid Makki, and Azizeh Atashi. A deep learning approach for missing data imputation in medical datasets. *Health and Technology*, 10(5):1125–1138, 2020.

Daniel Jarrett and Mihaela van der Schaar. Spinn: Synergistic integration of neural networks and propensity scores for missing data. In *International Conference on Artificial Intelligence and Statistics*, pp. 10216–10232. PMLR, 2022.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Soohwan Kim, Jaegyoon Yoo, Byungsoo Kim, and Kyoung-Jae Lee. Pi-naim: A network-aware imputation method for single-cell rna-seq data via graph convolutional networks. *Computers in Biology and Medicine*, 165:107345, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Shutao Li, Wei Song, Leyuan Fang, Yushi Chen, Pedram Ghamisi, and Jon Atli Benediktsson. Deep learning for hyperspectral image classification: An overview. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9):6690–6709, 2021.

Roderick JA Little and Donald B Rubin. *Statistical Analysis with Missing Data*. Wiley, 2019.

Yonghong Luo, Xiongfeng Cai, Ying Zhang, Jun Xu, and Xiaojie Yuan. Naomi: Non-autoregressive multiresolution sequence imputation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yonghong Luo, Xiongfeng Cai, Ying Zhang, and Jun Xu. Understanding and improving deep learning-based neural architecture for incomplete data imputation. *IEEE Transactions on Knowledge and Data Engineering*, 35(2):1425–1438, 2023.

Andrea Mattei, Milo Tognocchi, Andrea Zaccaria, Marcello Ziosi, and Vincenzo Lomonaco. Miwa: an information-extraction system for the MICHA summit. In *Proceedings of the 2nd International Conference on Mining Intelligence and Knowledge Exploration*, volume 11308 of *Lecture Notes in Computer Science*, pp. 3–14. Springer, 2019.

Pierre-Alexandre Mattei and Jesper Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International Conference on Machine Learning*, pp. 4413–4423. PMLR, 2019.

Joseph L Schafer. *Analysis of Incomplete Multivariate Data*. Chapman and Hall/CRC, 1997.

Shaun Seaman, Jonathan Bartlett, and Ian White. Multiple imputation with missing data indicators. *Statistical methods in medical research*, 22(6):631–645, 2013.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Jonathan AC Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338, 2009.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Paul Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

S. van Buuren. *Flexible imputation of missing data*. Chapman & Hall/CRC, Boca Raton, FL, 2nd edition, 2018a.

Stef van Buuren. *Flexible Imputation of Missing Data*. CRC Press, 2018b.

Youqing Wang, Zhiqiang Sun, and Tao Chen. Fault detection and diagnosis based on transfer learning for multimode processes. *IEEE Transactions on Instrumentation and Measurement*, 69 (4):1004–1015, 2020.

Jie Yan, Jian Wang, Cheng Li, and Xinbo Gao. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems*, 196:105805, 2020.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Gain: Missing data imputation using generative adversarial nets. In *International Conference on Machine Learning*, pp. 5689–5698. PMLR, 2018.

Jiaxuan You, Xiaojie Ma, Dawei Ding, Mykel J Kochenderfer, and Jure Leskovec. E-gain: Ensemble generative adversarial imputation networks for missing data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10683–10690, 2021.

Table 4: Comparison of imputation methods across key capabilities. PI-NAIM achieves full coverage of desired features through its dynamic dual-path architecture.

Method	Dyn. Routing	Temp. Support	Uncertainty	MAR/MNAR	Task Opt.
Mean / MICE	55	55	55	MAR only	55
GAIN / NAIM	55	55	55	MAR / MNAR	55
T-GAIN / BRITS	55	51	55	MAR / MNAR	55
MWAE / SPINN	55	55	51	MAR / MNAR	55
PI-NAIM (ours)	**51**	**51**	**51**	**MAR / MNAR**	**51**

Note: Dynamic routing refers to adaptive path selection based on missingness patterns. Task optimization indicates end-to-end training with downstream task objectives.

A Related Work

Missing data imputation is a challenging problem due to the complex, multimodal datasets in computer vision and medical imaging. Our work builds upon and focuses on four key areas: classical statistical methods, deep neural imputation, hybrid systems, and uncertainty quantification.

Traditional missing data strategies, often categorized as classical statistical imputation, include methods like complete case analysis Little & Rubin (2019) and naive techniques like mean and median imputation Schafer (1997). There are some advanced statistical techniques as well, such as k-Nearest Neighbors (kNN) Troyanskaya et al. (2001), MICE van Buuren (2018b), and Matrix Completion Candès & Recht (2009), however struggle to handle complex, high-dimensional missingness patterns common in modern datasets.

To address these, several deep learning methods have been developed. GAIN Yoon et al. (2018) and its extensions like E-GAIN You et al. (2021) and T-GAIN Cini et al. (2022b) utilize adversarial training for high-fidelity imputation. Other deep approaches include autoregressive methods like NAIM Luo et al. (2023) and Transformer-based methods like SAITS Du et al. (2023a) and BRITS Cao et al. (2018), which excel on difficult, time-series patterns but demand significant training resources and lack interpretability. Specialized temporal imputation methods, such such as GRU-D Che et al. (2018), NAOMI Luo et al. (2019), and ST-Transformer Chen et al. (2023a), are optimized for sequential contexts. Recent research has explored hybrid systems, such as MWAE Mattei et al. (2019), SPINN Jarrett & van der Schaar (2022), and HINT Awan et al. (2023), but these often rely on fixed architectures, segmented optimization, and have limited coverage of heterogeneous missingness (MCAR/MAR/MNAR). Furthermore, uncertainty quantification (UQ), critical for high-risk applications, has been addressed through Bayesian methods Gelman et al. (2013), Deep Ensembles Lakshminarayanan et al. (2017), and systematic neural networks Sensoy et al. (2018).

To address these limitations, we introduce PI-NAIM that bridges the efficiency of statistical models and deep learning via its dynamic dual-path architecture. Compared to static hybrid systems, PI-NAIM dynamically routes samples based on missingness complexity and provides combined uncertainty estimation, establishing a new state-of-the-art balance between statistical faithfulness and neural expressiveness.Table 4 benchmark our model advancements compared to the state-of-the-art methods.

B Datasets

For comprehensive evaluation and to demonstrate generalizability, we utilize two distinct dataset categories: the Medical Information Mart for Intensive Care III (MIMIC-III) and the CIFAR-10/100 vision benchmarks. The large-scale, multimodal MIMIC-III dataset, consisting of over 46,000 patient records, provides high-dimensional clinical time-series and structural data that mirrors the complex, naturally occurring missingness challenges found in multi-modal fusion and temporal sensor streams in computer vision. To explicitly validate PI-NAIM's application to core CV domains, we utilize the CIFAR-10/100 datasets, introducing synthetic pixel missingness to simulate scenarios like noisy visual inputs, occlusions, and missing sensor measurements. This dual-domain approach thoroughly tests our model's robustness and its adaptive dual-path architecture's efficiency across both complex temporal patterns and high-dimensional spatial data.

Table 5: Performance comparison of PI-NAIM.

Variant	AUROC	RMSE	Parameters	α
Full PI-NAIM	0.8725	1.952	10,262	0.6458
Static Fusion ($\alpha = 0.5$)	0.8474	1.9594	10,261	0.5
w/o Adaptive Fusion	0.8764	2.0884	13,653	0.5
w/o Imputation Path	0.8775	1.9453	7,617	1.0
w/o NAIM Path	0.8569	3.1917	8,213	0.0

C EXPERIMENTAL SETUP

We evaluated PI-NAIM across temporal and visual domains. For time-series data, the MIMIC-III dataset was preprocessed with one hour resampling, forward filling, and feature selection based on availability and predictive relevance. Numerical features were z-score normalized, and categorical ones were one-hot encoded.

For vision tasks, we simulated realistic corruption patterns random, block, and column missing pixels to mimic sensor noise and occlusions. The architecture was adapted with convolutional encoder-decoder pathways, spatial attention fusion, and routing logic responsive to spatial missingness. This unified setup enables domain-agnostic assessment of PI-NAIM's adaptive imputation and fusion capabilities.

D ABLATION STUDY

To evaluate the contribution of each component in PI-NAIM, we conducted a systematic ablation study on the MIMIC-III dataset, assessing imputation accuracy (RMSE) and downstream task performance which includes AUROC for mortality prediction. Each variant isolates a design element to quantify its technical significance. Results in Table 5 confirm that every module contributes meaningfully to overall performance.

We first examined the efficiency trade off by selectively removing imputation paths. The *w/o GAIN Path* variant retains only the statistical MICE branch, resulting in sharp degradation under high missingness conditions, while the *w/o MICE Path* variant, rely solely on GAIN, exhibits instability for low-missingness data. This validates the complementary nature of the two paths: MICE provides statistical robustness for simple patterns, whereas GAIN captures nonlinear dependencies under severe data corruption. Next, we evaluated the fusion mechanism. Replacing the adaptive fusion with a fixed averaging scheme (*w/o Adaptive Fusion*) or a uniform static weighting (*Static Fusion*, $\alpha = 0.5$) led to consistent performance drops, underscoring the importance of context-aware integration. The adaptive cross-path attention dynamically adjusts pathway contributions based on missingness complexity and task objectives, proving critical to model reliability and generalization.

The complete PI-NAIM architecture achieves the lowest RMSE and highest AUROC, validating the adaptive dual-path design. Its joint optimization of statistical and neural reasoning yields stable, semantically coherent imputations and superior predictive utility across diverse missingness regimes.

E LIMITATIONS

While the methodology worked well, there are certain weaknesses associated with PI-NAIM which need further consideration and investigation.

- Routing Computational Overhead: While the quality of imputation benefits from dynamic path selection, this also implies a computational overhead due to both the gating network and the parallel execution of the paths in both MICE and GAIN. The present architecture may need to be reconsidered toward those domains.

Table 6: MIMIC-III cohorts employed in experiments.

Cohort	Patients	Admissions	Mortality Rate	Mean LOS (days)
Full MIMIC-III	46,520	58,976	11.5%	6.7
ICU Subset	38,597	49,785	13.2%	3.4
Final Cohort	24,819	31,543	9.8%	4.1

- Assumption of Availability of Ground Truth for Training: Most of the previous supervised imputation techniques, along with the proposed one, PI-NAIM, assume that fully observed data shall be available for training. In practical scenarios when complete data are hard or unavailable, it is expected to yield poor generalization performance.
- Interpretability of Dynamic Routing: While the gating network allows adaptive routing, its choices are hardly interpretable. Lack of transparency in path selection may restrict user trust and adoption in high stake applications like healthcare.
- Dependence on Curriculum Masking Schedule: The performance in PI-NAIM depends upon the strategy of curriculum masking followed during training. There is a possibility that the performance can degrade due to suboptimal scheduling or mismatch in the missingness distribution during training and deployment.

Thus, addressing these limitations as part of future work may further enable PI-NAIM for a broader, even more challenging applicability in the real world, making it even more robust.

Rationale: Missing data in multimodal representations presents a critical efficiency-expressiveness dilemma: statistical methods like MICE are efficient for simple patterns but lack the representational capacity for complex, structured missingness, while deep learning methods like GAIN are expressive but incur substantial computational overhead. To resolve this, we propose PI-NAIM, a novel, unified framework. Its core innovation is a Dynamic Path Selection mechanism that uses the Missingness Rate (MR) as a measurable proxy for pattern complexity. This dynamically routes samples: low-complexity cases are handled by the efficient statistical branch, and high-complexity cases are directed to the expressive deep branch (GAIN with Temporal Analysis). Coupled with a Cross-Path Attention Fusion module, PI-NAIM is the first model to dynamically balance computational efficiency with representational power, establishing a globally adaptive and highly performant imputation paradigm critical for real-world multimodal applications.

F Equation Definitions and Parameter Explanations

All of the equations in the PI-NAIM study are explained in depth in this section, along with thorough parameter descriptions.

F.0.1 Input Processing

Definitions of Parameters:

- $\mathbf{M}$: Binary mask matrix of the same dimensions ($n \times d$) as input data $\mathbf{X}_{\text{missing}}$. shows each value's presence or absence explicitly. $\mathbf{M_{ij}}$: One mask matrix element that corresponds to the data point at i-th row (sample) and j-th column (feature). The value that indicates that the associated data point $\mathbf{X_{ij}}$ is missing and has to be imputationed is 0. The value that indicates that the associated data point $\mathbf{X_{ij}}$ is observed and legitimate is 1.

Parameter Definitions:

- $\mathbf{E}$: Tensor for missing data with shape n x d x e. Makes e-dimensional vector for each data point to show complex missingness patterns.
- $\mathbf{X}_{\text{observed}}$: Input data that's actually present (not missing). Missing data usually filled with placeholders.
- $\mathbf{M}$: Binary mask concatenated with observed data to provide explicit information about real vs. Placeholder values.

- $\mathbf{concat}()$: Merge observed data with mask on feature axis.
- $\mathbf{LSTM}()$: Long Short Term Memory network process input, learn feature dependencies.

F.0.2 Dynamic Path Selection

Parameter Definitions:

- $\mathbf{MR}$: Missingness Rate scalar (0 to 1) indicates the proportion of missing data in a sample. Acts as substitute for imputation complexity.
- $\mathbf{n}$: How many rows in dataset.
- $\mathbf{d}$: How many features (columns) in dataset.
- $\sum\sum \mathbf{M_{ij}}$: Count all non-missing data in dataset twice. The fraction shows total "observedness"; subtracting from 1 gives final missingness rate.

F.0.3 GAIN Path with Temporal Analysis

Parameter Definitions:

- $\mathcal{L}_{\mathbf{G}}$: Generator's loss function for realistic data imputations.
- $-\mathbb{E}[\mathbf{D}(\hat{\mathbf{X}}, \mathbf{M})]$: The chance that fake data seems real is $D(\hat{X}, M)$. Generator lower bad expectations to make things appear more real.
- α: Hyperparameter controlling trade Balance realism and accuracy in adversarial loss and reconstruction loss.
- $\|(\mathbf{1} - \mathbf{M}) \odot (\mathbf{X} - \hat{\mathbf{X}})\|^{\mathbf{2}}$: Reconstruction loss (MSE) applied only to originally missing elements. $(1 - M)$ masks missing entries, $(X - \hat{X})$ is difference between true and imputed values.

Parameter Definitions:

- $\mathcal{L}_{\mathbf{D}}$: Discriminator loss with WGAN-GP for stable training.
- $\mathbb{E}[\mathbf{D}(\mathbf{X}, \mathbf{M})]$: Score discriminator assign high score to real data.
- $-\mathbb{E}[\mathbf{D}(\hat{\mathbf{X}}, \mathbf{M})]$: Score discriminator give low score to generator fake data (to be less).
- λ: Hyperparameter weighting gradient penalty term.
- $\mathbb{E}[(\|\nabla \mathbf{D}\|_{\mathbf{2}} - \mathbf{1})^{\mathbf{2}}]$: Gradient penalty force discriminator to obey Lipschitz condition by penalizing gradient norm not equal to 1.

F.0.4 Cross-Path Attention Fusion

Parameter Definitions:

- α: Attention weight matrix set importance for missing data from MICE and GAIN paths.
- $\mathbf{Q}_{\text{NHM}}$: Query matrix from missingness embeddings E with linear layer. Data missing means "question"
- $\mathbf{K}_{\text{imp}}$, $\mathbf{V}_{\text{imp}}$: Key- Value matrices from merging MICE and GAIN results. Keys match query, values store imputation data.
- $\mathbf{d_k}$: Key vectors' size in dimensions. Scaling stop softmax to avoid small gradients.
- $\mathbf{softmax}(...)$: Normalizes attention scores to probability distribution for imputation paths.
- $\mathbf{h}_{\text{fused}}$: Final imputation vector as weighted sum of Value vectors using attention scores.

F.0.5 TASK-SUPERVISED ADAPTIVE FUSION

Parameter Definitions:

- $\lambda_{\mathbf{t}}$: Scalar weight (0- Balance imputed data and features specific to the task.
- σ: Sigmoid function keep input in (0, 1) range.
- $\mathbf{W_t}$: Learnable weight vector.
- $[\mathbf{t}, \mathbf{C}_{\text{imp}}, \mathbf{C}_{\text{task}}]$: Concatenated input vector containing.
  - $\mathbf{t}$: Train epoch or task id, change weighting method.
  - $\mathbf{C}_{\text{imp}}$: Imputation confidence (e.g., MSE between imputed and ground-truth values).
  - $\mathbf{C}_{\text{task}}$: Task confidence (e.g., cross-entropy loss from preliminary task prediction).

Parameter Definitions:

- $\hat{\mathbf{y}}$: Final prediction for mortality risk task.
- $\mathbf{h}_{\text{imp}}$: Feature representation derived from fused imputations.
- $\mathbf{h}_{\text{task}}$: Feature representation learned for task directly.
- $\lambda_{\mathbf{t}}$, $(\mathbf{2} - \lambda_{\mathbf{t}})$: Adaptive weights help ensemble stay balanced. High λ_t relies on good imputation, low λ_t uses task specific features as backup.

F.0.6 UNCERTAINTY QUANTIFICATION

Parameter Definitions:

- $\sigma^{\mathbf{2}}$: Epistemic uncertainty in final prediction. High variance indicates model uncertainty.
- $\mathbf{K}$: Total stochastic predictions (Monte Carlo Dropout or bootstrap sampling).
- $\hat{\mathbf{y}}^{(\mathbf{k})}$: k-th stochastic prediction from model.
- $\bar{\mathbf{y}}$: Average of all K predictions. Formula calculate sample variance in predictions, measure how much output varies due to imputed value uncertainty.

F.0.7 TRAINING STRATEGY

Parameter Definitions:

- $\mathcal{L}$: Total loss function for end-to-end training.
- $\mathcal{L}_{\text{imp}}$: Imputation loss (miss data fill-in error)
- $\mathcal{L}_{\text{task}}$: Downstream task loss like cross-entropy for classifying.
- $\mathcal{L}_{\text{reg}}$: Regularization loss (like L2 on Theta).
- $\lambda_{\mathbf{1}}, \lambda_{\mathbf{2}}, \lambda_{\mathbf{3}}$: Adaptive weights learned from homoscedastic uncertainty, lambda equals one over two sigma squared, where sigma is a learnable parameter. System adjust losses based on task noise automatically.