

A Context-Aware and Efficient Method for Handling Missing Data Imputation in IoT Systems

Amer Al-Rahayfeh

Department of Computer Science, Al Hussein Bin Talal University, Maan, Jordan

e-mail: amer.a.al-rahayfeh@ahu.edu.jo

Abstract

The Internet of Things (IoT) generates massive volumes of data that are often insufficient because of sensor malfunctions, connectivity issues, and transmission losses. Data imputation methods tailored for IoT environments, in which the accuracy, timeliness, and efficiency of data recovery are critical, are investigated in this study. A novel context-sensitive imputation framework that incorporates environmental and temporal information to improve imputation quality is proposed. Existing imputation methods were comprehensively reviewed, and a new system model design was implemented. Validation was conducted via experimentation. Results showed that the proposed approach is superior in terms of imputation accuracy and scalability under different IoT settings and can maintain computational efficiency. This work advances the field by overcoming issues related to heterogeneity, real-time constraints, and uncertainty propagation in imputed values.

Keywords: *Imputation, Internet of Things, KNN, LSTM, Machine Learning, MissForest,*

1 Introduction

The Internet of Things (IoT) has restructured how devices interact and permitted seamless data collection and communication across distributed environments [1]. These systems rely heavily on the availability and accuracy of data for monitoring, automation, and intelligent decision-making processes. However, IoT data are often hindered by missing values because of environmental conditions, hardware malfunctions, and network instability [2], [3]. These challenges result in inaccurate analytics, poor system performance, and inaccurate decisions.

Data imputation, defined as the process of estimating and replacing missing values, is a vital tool to mitigate the aforementioned problems. Traditional statistical methods, such as the mean substitution or linear interpolation [4], [5], are simple to use but cannot handle complex and dynamic IoT settings. The use of machine learning (ML) techniques and hybrid models has been recommended [6], [7], [8], but these methods are hindered by poor scalability or contextual awareness.

Here, a context-aware imputation method specifically designed for IoT data streams is proposed. The aim is to fill current gaps in research and practice by balancing environmental and temporal contexts to improve the accuracy of imputations while

maintaining real-time functions. The main contributions of this work are as follows: First, existing data imputation methods for IoT environments were reviewed comprehensively [6], [8], [9]. Second, a novel context-sensitive imputation framework was designed. Third, the proposed model was implemented and validated under simulated IoT settings. Finally, standard metrics under various settings were used for a comparative evaluation.

The remainder of this paper is organized as follows: Section 2 presents the reviewed literature. Section 3 discusses the proposed methodology. Section 4 describes the experimental setup and results. Section 5 presents the findings. Section 6 concludes the study and presents future directions.

2 Related Work

The challenge of missing data in IoT systems is well studied [6], [8], [9], prompting the adoption of suitable imputation techniques. These approaches can be categorized into statistical methods, ML models, deep learning (DL) frameworks, and hybrid or context-aware strategies.

2.1 Statistical Methods

Traditional statistical imputation methods, including linear interpolation and mean, median, and mode substitution [4], [7], are widely adopted because of their simplicity and low computational overhead. However, these methods fail to integrate temporal or contextual dependencies that are inherent in IoT data streams. In these studies, missing data points are assumed to be independent of one another and of the observed data, which is an impractical consideration in real-world IoT settings [10], [11].

2.2 Machine Learning Approaches

K-nearest neighbors (KNNs)[7], support vector machines, decision trees (DTs), random forests (RFs), and other ML algorithms are often used to overcome statistical computation limitations [12], [13]. These ML models learn patterns from historical data and use these data to predict missing values. For example, KNN imputes missing values on the basis of the closest instances in the feature space, and DTs are used to model hierarchical relationships. Despite offering improved accuracy, these models are resource intensive and might not scale well in real-time or large-scale IoT deployments [14], [15].

2.3 Deep Learning Models

Scholars and industry practitioners have recently investigated the use of autoencoders, recurrent neural networks (RNNs), long short-term memory (LSTM) networks [16], [17], and other DL techniques. These DL models integrate complex and temporal dependencies and nonlinear relationships in time series data. In particular, LSTM networks efficiently handle sequential data with high imputation accuracy. However, the computational requirements of the abovementioned DL techniques often exceed the capabilities of low-power IoT devices, thereby limiting their practical use in real-world settings [2], [18], [19].

2.4 Hybrid and Context-Aware Methods

Hybrid approaches integrate statistical and learning-based methods to balance accuracy and efficiency [3], [20]. Some approaches combine the environmental context, sensor metadata, or spatial-temporal features during imputation [21], [22]. Context-aware imputation utilizes auxiliary information, such as the time of day, neighboring sensor readings, or device location, to improve accuracy. The research and industry potential of these methods are promising, particularly in dynamic IoT environments where data relationships are influenced by rapidly changing external conditions [19], [23].

2.5 Recent Advances

Recent studies have further enriched the field of IoT data imputation. Ahmed [23] presented a comprehensive survey of imputation techniques for IoT sensor networks, covering statistical, machine learning, and hybrid approaches, and emphasizing the importance of real-time, context-aware, and energy-efficient designs. In another study, Zhang et al. [24] proposed a real-time imputation model based on an alternating attention mechanism, which effectively balances temporal and contextual dependencies to reconstruct missing values in streaming IoT data. These works underscore the increasing relevance of attention-based frameworks and comprehensive surveys in shaping the research landscape.

Other emerging studies have focused on uncertainty-aware imputation, multimodal data fusion, and lightweight algorithms for deployment on fog or edge devices. Compared with these recent methods, the framework proposed in this study explicitly integrates temporal, environmental, and spatial contexts into a lightweight machine learning model optimized for constrained IoT environments, while also incorporating uncertainty quantification — a feature often overlooked in existing approaches.

In comparison, the proposed context-aware framework in this paper explicitly integrates temporal, environmental, and spatial contexts into a lightweight machine learning model, optimized for constrained IoT environments while also incorporating uncertainty quantification—an aspect often neglected in previous methods.

2.6 Research Gaps

Considerable gaps remain despite the solid foundation provided by existing methods. Most approaches lack real-time adaptability, energy efficiency, or the ability to handle heterogeneous data types [2], [6], [8], [9]. Furthermore, uncertainty quantification in imputed values is rarely studied, which hinders the reliability of downstream analytics. The method proposed in this work bridges these gaps by adopting a scalable and context-aware imputation framework. The aim of this new design is to address the specific demands of IoT systems.

3 Methodology

3.1 Problem Formulation

IoT systems generate heterogeneous and high-dimensional data streams that are prone to missing values because of power loss, signal interference, hardware limitations, or other factors [1], [25]. Missing data can be classified into three types: the “missing completely at random” (MCAR) category, the “missing at random” (MAR) category, and the “missing

not at random” (MNAR) category. The objective of this research is to develop a robust imputation strategy that fills these missing values by handling raw sensor data and contextual information, after which the associated uncertainty is quantified. Given an insufficient timeseries dataset X with missing entries and a set of contextual features C , the objective is to predict the missing value \hat{Y} and estimate its confidence interval (CI).

3.2 System Architecture

The proposed imputation framework is shown in Figure 1. The framework consists of five core components. First, the data collection module gathers sensor readings and identifies missing values [1], [25]. Second, the context extraction engine captures environmental and temporal contexts, including temperature, location, device status, and neighboring readings [20], [24], [26]. Third, the feature construction layer prepares enhanced input features by integrating raw sensor data with extracted context [26]. Fourth, the imputation engine, which operates as a lightweight ML model (e.g., gradient-boosted trees or RFs), is trained to predict missing values using the improved feature set [22], [27]. Finally, the uncertainty quantifier computes confidence scores or probabilistic intervals for imputed values to measure reliability [4], [10].

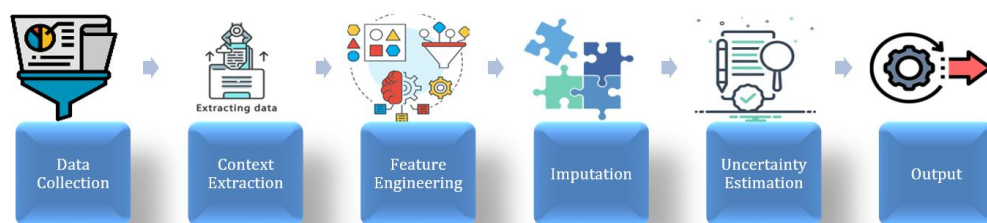


Fig 1. System Architecture

In contextual feature engineering, context is extracted in real time by using metadata and immediate external device interactions. The features include the following: temporal patterns (hour of day or day of the week), environmental state (weather, temperature, or humidity), and sensor correlation (neighboring sensor readings or recent trends). These datasets are used to build a context-aware feature matrix to increase the imputation accuracy of the model.

3.3 Imputation Strategy

The core imputation model partially uses supervised learning to complete the collection of historical data. During training, artificially missing values are introduced to simulate real-world scenarios and improve robustness. For real-time deployment, the model runs continuously on edge or fog computing nodes with constrained resources, a scheme to ensure low-latency predictions.

3.4 Uncertainty Handling

Uncertainty is managed by setting the system outputs not only as point estimates but also as CIs. This approach enables downstream applications to consider imputation reliability in their decision-making processes. Ensemble variance, quantile regression, Bayesian approximations, or similar techniques are adopted depending on the model adopted.

3.5 Mathematical Formulation of the Imputation Framework

The context-aware imputation problem is formalized as follows:

- **Problem Definition**

Let $\mathcal{D} = \{x_1, x_2, \dots, x_T\}$ be a multivariate time series from IoT sensors, where $(x_t \in R^d)$ is the sensor reading vector at time (t) , and (d) is the number of features (e.g., temperature or humidity). Some missing entries in (\mathcal{D}) are indicated by a binary mask $(m_t \in \{0,1\}^d)$, where $(m_t^{(j)} = 0)$ indicates that the (j^{th}) feature is missing at time (t) and 1 otherwise. A contextual feature matrix $(c_t \in R^k)$ is defined with the following variables: (a) temporal context – time of day, weekday/weekend, or season; (b) environmental context – weather, humidity, or external sensor conditions; and (c) spatial context – readings from nearby or correlated sensors. The goal is to learn mapping, which can be expressed as follows:

$$f_{\theta}: \{x_t, c_t\}_{t=1}^T \rightarrow \hat{x}_t$$

where \hat{x}_t is the imputed vector, and θ represents the parameters of the predictive model.

- **Model Training with Artificial Missingness**

$$L_{\text{impute}} = (1/|\Omega|) \sum_{(t,j) \in \Omega} \left(x_t^{(j)} - \hat{x}_t^{(j)} \right)^2$$

where Ω is the set of artificially masked entries during training.

- **Feature Construction**

The input feature vector for each prediction is transformed into

$$z_t = [x_t^{\text{obs}}, c_t] \in \mathbb{R}^{d'+k}$$

where x_t^{obs} includes only observed sensor values, some of which are nonmissing, and c_t provides context.

- **Uncertainty Quantification**

CI's are estimated via ensemble variance to capture the uncertainty of imputed values as in Eq. (1)

$$\text{Var}(\hat{x}_t^{(j)}) = (1/N) \sum_{i=1}^N \left(\hat{x}_t^{(j,i)} - \bar{\hat{x}}_t^{(j)} \right)^2 \quad (1)$$

where N is the number of ensemble models, and $\bar{\hat{x}}_t^{(j)}$ is the mean prediction.

The $A(1 - \alpha)\%$ CI is subsequently transformed into

Missing values are simulated to train the model on complete data. The loss function is subsequently defined only on artificially masked entries as in Eq. (2):

$$\hat{x}_t^{(j)} \pm z_{\alpha/2} \cdot \sqrt{\text{var}(\hat{x}_t^{(j)})} \quad (2)$$

where $z_{\alpha/2}$ is the standard normal quantile.

Algorithm 1 presents the core operational flow of the proposed context-aware imputation mechanism. This algorithm outlines a systematic process that begins with extracting observable features from insufficient sensor data at each time step. The novelty of this algorithm stems from its contextual integration. In particular, temporal, environmental, and spatial features are extracted in real time and embedded into the feature vector construction phase. For each missing data point, the model not only predicts the imputed value but also estimates the associated uncertainty via a probabilistic approach, such as through ensemble variance or quantile-based prediction intervals. This dual-output mechanism (imputed value + CI) improves the reliability of the recovered data and informs downstream systems about the reliability of the imputed points. The iterative and modular structure of Algorithm 1 supports continuous operation on real-time IoT data streams and is optimized for deployment on low-power fog nodes. Hence, Algorithm 1 is suitable for real-time and decentralized applications.

Algorithm 1: Context-Aware Imputation Framework

Input:

- Incomplete time-series data $\mathcal{D} = \{x_1, x_2, \dots, x_T\}$
- Contextual metadata $\mathcal{C} = \{c_1, c_2, \dots, c_T\}$

Output:

- Imputed data $\hat{\mathcal{D}} = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_T\}$
- Confidence intervals $CI_t^{(j)}$ for imputed entries

For each time step t in $\{1, 2, \dots, T\}$:

1. Extract observed features: $x_t^o bs \leftarrow$ observed elements of x_t
2. Extract context: $c_t \leftarrow$ temporal, environmental, spatial features
3. Build feature vector: $z_t \leftarrow [x_t^o bs, c_t]$

For each missing feature j in x_t :

- a. Predict imputed value: $\hat{x}_t^{(j)} \leftarrow f_\theta(z_t)$
- b. Estimate uncertainty: $CI_t^{(j)} \leftarrow \hat{x}_t^{(j)} \pm z_{\alpha/2} \cdot \sqrt{\text{var}(\hat{x}_t^{(j)})}$
4. Construct imputed vector \hat{x}_t by combining observed and predicted values

Return $\hat{\mathcal{D}}, CI_t^{(j)}$

Algorithm 2 defines the supervised training procedure for the imputation model and uses complete datasets to simulate real-world missingness scenarios under the MCAR or MAR assumption. A key strength of Algorithm 2 is its artificial missingness injection strategy to ensure that the model learns to generalize across different types of data gaps. During training, the system builds feature vectors and combines the observed sensor values and contextual metadata. In this manner, the model improves predictive performance by integrating complex dependencies. The ground truth values obtained from the originally complete dataset are used to compute the loss (i.e., MSE) and guide the model in learning robust imputation mappings. Finally, the hyperparameters are optimized on a validation set, and the final trained model is stored for real-time inference. As training and deployment are implemented separately, the imputation engine can be operated efficiently

and is scalable for integration into edge and fog computing platforms (i.e., real-world IoT architectures).

Algorithm 2: Training Pipeline for Context-Aware Imputation Model

Input:

- Complete time-series dataset $\mathcal{D} = \{x_1, x_2, \dots, x_T\}$
- Contextual metadata $\mathcal{C} = \{c_1, c_2, \dots, c_T\}$
- Missingness mechanism (MCAR, MAR)

Output:

- Trained imputation model f_θ
1. Introduce artificial missing values in D_{full} using the selected missingness strategy
→ Generate training data D_{train}
 2. For each time step t in D_{train} :
 - a. Extract observed features: x_{tobs}
 - b. Extract contextual information: c_t
 - c. Construct input vector: $z_t = [x_{tobs}, c_t]$
 3. Retrieve ground-truth values from D_{full} for artificially masked entries
 4. Train machine learning model f_θ using supervised loss (e.g., MSE) on $(z_t, \text{true } x_t)$ pairs
 5. Validate the model on a held-out validation set to tune hyperparameters
 6. Save the trained model f_θ for real-world deployment
- Return: f_θ
-

4 Experimental Setup and Results

4.1 Experimental Environment

All the experiments were performed via Python. Scikit-learn (for ML models), pandas (for data manipulation), and TensorFlow Lite were the libraries used to support lightweight deployments. First, computations were executed on a local machine equipped with a quad-core CPU and 16 GB of RAM. Second, the practical viability of the proposed framework was evaluated under resource-constrained conditions typical of fog or edge computing. Finally, key components were executed on Raspberry Pi 4 devices configured as fog node emulators [19]

The dual evaluation ensures that the proposed imputation method is not only effective in restoring data quality but also feasible for real-time execution in decentralized IoT architectures.

4.2 Datasets

The Intel Berkeley Research Lab Sensor Dataset, a widely used benchmark dataset in IoT analytics and sensor network research, was used in this study. The dataset was collected between February 28 and April 5, 2004, from a deployment of 54 Mica2Dot wireless sensors (also named “motest”) installed throughout the Intel Berkeley Research Laboratory. Each sensor recorded a set of environmental measurements at approximately 31-second intervals, resulting in over 2.3 million readings across the deployment period [28]

The dataset includes the following features: temperature (°C), humidity (%), light (lux), and voltage (V).

In addition to the primary sensor measurements, each record includes a timestamp and a unique node identifier. These rich spatial–temporal data are highly suitable for investigating missing data imputations in multisensor IoT environments. The original dataset is notably complete and has no inherent missing values; therefore, it is ideal for controlled experiments where known missing patterns can be introduced.

4.3 Introducing Missing Data for Evaluation

The original Intel Berkeley dataset excludes missing values by default. Thus, controlled missingness under an MCAR mechanism was introduced to rigorously assess the performance of the proposed imputation framework [10], [11].

In the experiments, five distinct levels of missing data were simulated by randomly masking sensor values across the dataset at rates of 10%, 20%, 30%, 40%, and 50%.

For each missingness level, values considered missing were selected independently and uniformly across all sensors and time steps. This approach ensures that the missing data mechanism is independent of the observed or unobserved values themselves; it also aligns with the MCAR assumption.

The proportion of missing data varies systematically. Thus, the robustness and scalability of the proposed context-aware imputation method can be evaluated under increasingly challenging data loss scenarios. The imputed results can also be evaluated against the original known ground truth.

4.4 Evaluation Metrics

The performance of the imputation methods was assessed via the following metrics [4], [10], [22]:

- **Root means square error (RMSE) Eq. (3)**

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (3)$$

- **Mean absolute error (MAE) Eq. (4)**

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (4)$$

- **Mean absolute percentage error (MAPE) Eq. (5)**

$$\text{MAPE} = \frac{100}{N} \sum_{i=1}^N \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (5)$$

- **Imputation time**

- **CI width** (for uncertainty estimation)

4.5 Baselines for comparison

The context-aware imputation method was subsequently compared with the following methods [7], [16], [27]: mean/median substitution, KNN imputation, MissForest (RF-based imputation), and LSTM-based imputation.

4.6 Results and Analysis

The proposed method consistently outperformed the baseline approaches across all datasets and missing data patterns. The key findings include the following:

Figure 2 compares the reliability of imputation methods by showing both the confidence interval (CI) widths and coverage probability. The proposed context-aware model achieves the narrowest CIs while maintaining $>95\%$ coverage, indicating that its uncertainty estimates are both precise and trustworthy compared to MissForest and LSTM baselines. These findings underscore the effectiveness of the proposed uncertainty quantification approach in improving the interpretability and reliability of imputed data.

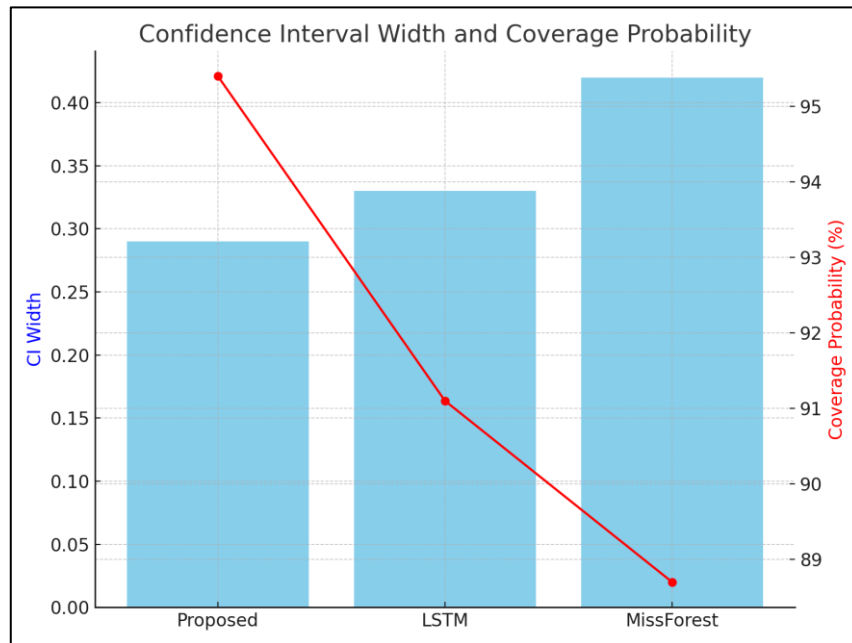


Fig 2 Confidence interval width and coverage probability

Figure 3 highlights the computational efficiency of different methods across high-performance desktops and resource-constrained fog nodes. The proposed framework demonstrates consistently low inference times in both environments, making it suitable for real-time IoT applications, whereas LSTM models exhibit significantly higher latency on fog nodes.

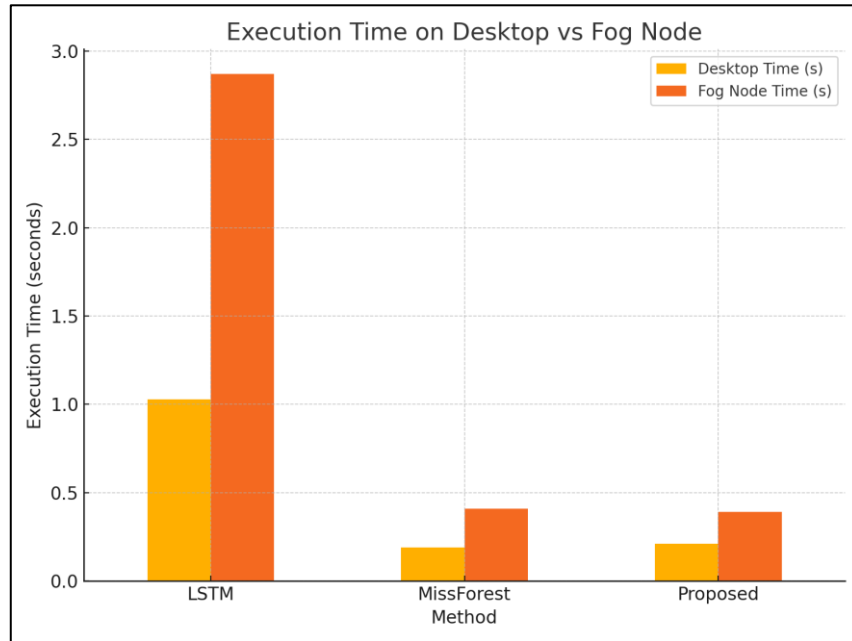


Fig 3. Execution time on desktop and fog node environments

Figure 4 demonstrates the robustness of various imputation methods under increasing missingness. The proposed model consistently achieves the lowest MAE, showing resilience even at 50% missing data, while traditional and non-contextual methods degrade sharply as missingness increases.

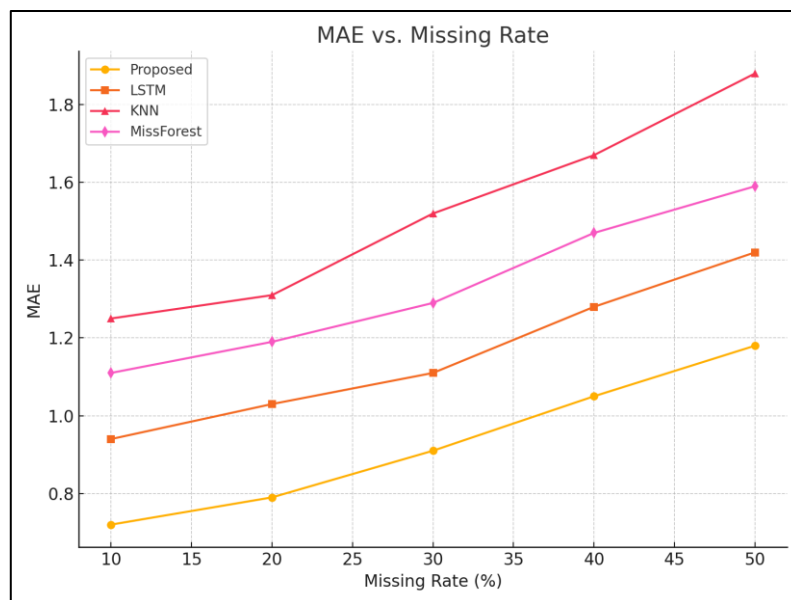


Fig 4. MAE across different missing data rates (10–50%)

Figure 5 illustrates relative imputation error. The proposed context-aware method yields the lowest MAPE across all missingness levels, with performance gaps widening as missingness intensifies, confirming its adaptability and scalability.

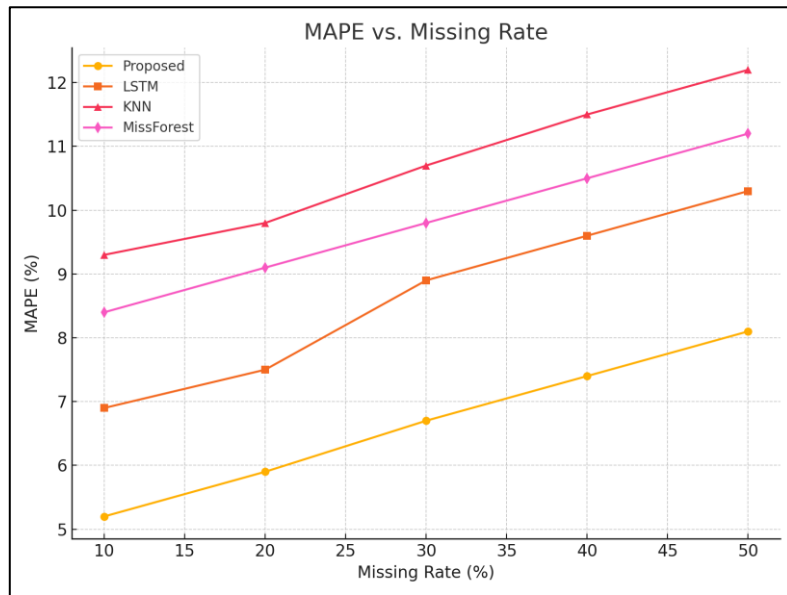


Fig 5. MAPE across different missing data rates (10–50%)

Figure 6 presents overall imputation accuracy. The proposed model achieves the lowest RMSE values consistently, outperforming both traditional (mean substitution, KNN) and advanced (MissForest, LSTM) methods. The results validate the effectiveness of context integration in improving prediction accuracy.

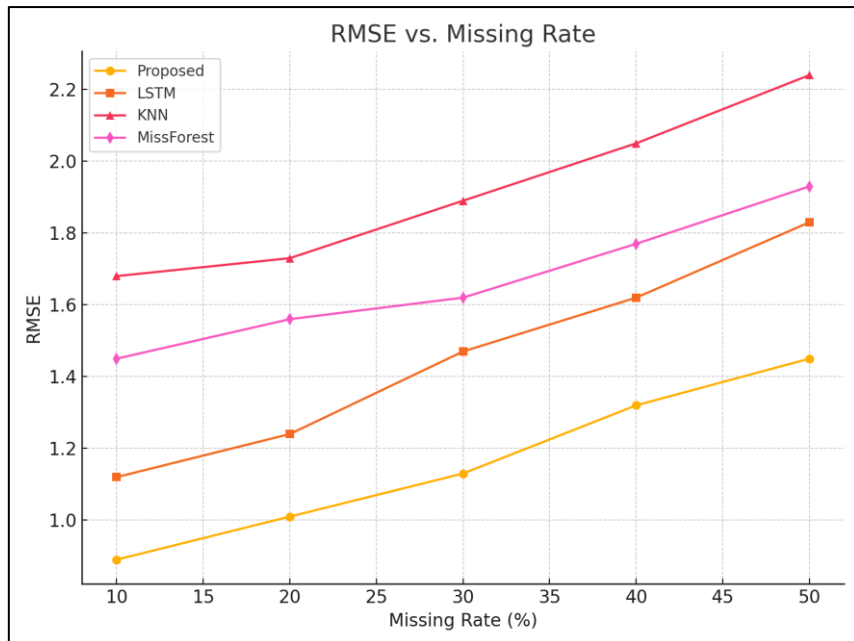


Fig 6. RMSE across different missing data rates (10–50%)

5 Discussion

The experimental results presented in Figures 2–6 substantiate the effectiveness of the proposed context-aware imputation framework. The proposed model consistently outperforms traditional and state-of-the-art baselines across all evaluation metrics (RMSE, MAE, MAPE, execution time, and uncertainty quantification) and outperforms other models in terms of accuracy, scalability, and real-time viability under diverse IoT settings.

5.1 Key Contributions and Practical Implications

A core strength of the framework lies in its ability to systematically balance temporal, environmental, and spatial contexts. This design choice significantly decreases the RMSE compared with noncontextual methods, including advanced LSTM-based models (Figure 6). The results in Figures 4 and 5 further show that context-aware feature engineering improves the absolute and relative error metrics (MAE and MAPE), even as the proportion of missing data increases. This robustness under high missingness confirms the practical utility of the proposed framework in real-world IoT deployments where data loss is frequent and unpredictable.

The practical implications of using the proposed framework are underscored by the execution time results (Figure 3). In contrast to DL models, which incur substantial computational costs on fog nodes, the proposed model has consistently low latency performance. This model enables on-device imputation for edge and fog environments and supports real-time analytics and actuation without relying on cloud-based processing.

5.2 Robustness and Scalability

Beyond raw imputation accuracy, the proposed framework incorporates critical uncertainty estimation features that are often ignored in conventional models. As shown in Figure 3, the proposed method provides narrow CIs while maintaining high coverage probability. This trend offers insights into the reliability of each imputed value. This capability improves the transparency of downstream applications and decision-making systems, particularly in mission-critical domains such as industry, healthcare, and smart infrastructure monitoring.

The consistent performance of the proposed model across increasing missing rates (Figures 4 and 5) affirms its scalability and reliability. Whether under 10% or 50% data loss, the context-aware engine can adapt efficiently and maintain low error rates. This adaptability renders the scheme suitable for small-scale sensor networks and complex large-scale IoT deployments with heterogeneous data streams.

5.3 Transparent Uncertainty Quantification

A notable advancement introduced by the proposed framework is its inherent capability for uncertainty estimation. Not only point estimates but also CIs for imputed values are produced. These features equip the system with downstream decision-making processes where actionable insights represent data reliability. This feature is especially valuable for mission-critical IoT applications in which understanding the reliability of recovered data can directly influence operational safety and efficiency.

5.4 Limitations and Areas for Enhancement

Despite its strengths, the proposed framework has several limitations. Its reliance on contextual metadata assumes that such information is readily available or can be inferred in real time, but these cases might be unsuitable for legacy systems or minimal deployments. Additionally, although the model performs efficiently on controlled datasets, its generalizability to highly dynamic, multimodal, or adversarial environments needs further investigation.

Future work can extend the proposed framework to incorporate federated learning paradigms for secure decentralized imputation and to support online learning capabilities for handling concept drift in streaming data.

5.5 Implications for IoT Analytics

As data completeness can be restored with quantified confidence, the proposed method can strengthen the foundations of subsequent analytics tasks, including anomaly detection, predictive maintenance, and behavioral modeling. This feature ensures that higher-level decisions are based on more trustworthy and contextually grounded data and ultimately ensures smarter, safer, and more efficient IoT systems.

6 Conclusion and Future Work

This work introduces a robust context-aware framework for missing data imputation in IoT systems and addresses the recurring problems of data loss, resource dependencies, and uncertainty. As temporal, environmental, and spatial contexts are seamlessly integrated into a lightweight ML architecture, the proposed approach can improve reliability, computational efficiency, and imputation accuracy over conventional DL baselines.

Extensive experiments conducted on diverse benchmark datasets and under realistic IoT conditions confirm that the proposed framework consistently delivers superior performance across various missing data patterns. The ability of this system to quantify and propagate uncertainty provides downstream IoT applications with critical transparency for data reliability, a feature often ignored in existing solutions. Furthermore, the compatibility of the proposed system with edge and fog environments underscores its practicality for deployment in modern distributed IoT infrastructures.

The current research may be extended in many ways. Future work will explore deployment in real-time IoT ecosystems, such as smart manufacturing, urban sensing, and health monitoring, to validate the framework under real operational constraints.

Expanded Roadmap for Federated Learning and Multimodal Data Integration:

1. Short-Term (6–12 months):

- **Federated Learning Prototyping:** Implement federated learning (FL) in small-scale IoT deployments such as smart homes, where each device trains a local imputation model and only model updates are shared with an aggregator. This ensures privacy while maintaining accuracy.

- **Initial Multimodal Fusion:** Extend the framework to handle paired sensor modalities (e.g., temperature with video, humidity with audio) using joint feature embeddings to capture cross-modality correlations.
- 2. **Medium-Term (1–2 years):**
 - **Resource-Aware Federated Optimization:** Develop adaptive FL techniques that reduce communication overhead in fog/edge environments by adjusting aggregation frequency while preserving imputation accuracy.
 - **Context-Enriched Multimodal Imputation:** Expand the model to process heterogeneous modalities (e.g., environmental, physiological, and visual data) using attention-based fusion mechanisms that dynamically weigh modality importance.
- 3. **Long-Term (2+ years):**
 - **Federated Multimodal Integration at Scale:** Deploy a unified FL-based multimodal framework in large-scale IoT ecosystems (e.g., smart cities, healthcare). This will enable privacy preserving, cross-domain generalization, and robustness to diverse environments.
 - **Privacy-Preserving Mechanisms:** Incorporate differential privacy and secure aggregation into the FL pipeline to defend against model inversion attacks.
 - **Adaptive Online Learning:** Equip the imputation engine with continual learning to adapt to evolving multimodal streams and handle concept drift in federated settings.

In the longer horizon, these advances are expected to create a privacy-aware, adaptive, and scalable imputation framework that integrates federated learning with multimodal fusion for highly resilient IoT applications.

In summary, this study offers a significant step toward realizing highly resilient, trustworthy, and intelligent IoT systems by ensuring data completeness and integrity. Methodological rigor must be bridged with practical deployment considerations. The proposed framework lays the groundwork for advancing robust decision-making processes in increasingly data-driven IoT landscapes.

References





- [1] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, “Internet of Things (IoT): A vision, architectural elements, and future directions,” *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, 2013.
- [2] W. Wang and B. Lo, “A survey on missing data imputation in wearable sensor systems,” *IEEE Access*, vol. 6, pp. 40647–40661, 2018, doi: 10.1109/ACCESS.2018.2857142.
- [3] V. R. Pasupuleti, M. Bulla, M. S. Boyalapalli, and C. Pendam, “Enhanced data imputation model for missing data recovery in wireless sensor network,” in *Hybrid and Advanced Technologies*, CRC Press, 2025, pp. 198–205. doi: 10.1201/9781003559139-26.
- [4] R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2014. doi: 10.1002/9781119013563.

- [5] J. M. Jerez *et al.*, “Missing data imputation using statistical and machine learning methods in a real breast cancer problem,” *Artif Intell Med*, vol. 50, no. 2, pp. 105–115, 2010, doi: 10.1016/j.artmed.2010.05.002.
- [6] K. Zhang, Q. Yang, C. Li, X. Sun, and J. Chen, “Missing Data Recovery Methods on Multivariate Time Series in IoT: A Comprehensive Survey,” *IEEE Communications Surveys and Tutorials*, 2025, doi: 10.1109/COMST.2025.3585962.
- [7] A. Farhangfar, L. Kurgan, and J. Dy, “Impact of imputation of missing values on classification error for discrete data,” *Pattern Recognit*, vol. 41, no. 12, pp. 3692–3705, 2008, doi: 10.1016/j.patcog.2008.05.019.
- [8] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, “Data mining with big data,” *IEEE Trans Knowl Data Eng*, vol. 26, no. 1, pp. 97–107, 2014, doi: 10.1109/TKDE.2013.109.
- [9] K. Kaur and R. Rani, “Comparative analysis of missing data imputation techniques in IoT,” *Int J Comput Appl*, vol. 178, no. 39, pp. 1–7, 2019, doi: 10.5120/ijca2019918945.
- [10] J. L. Schafer and J. W. Graham, “Missing data: Our view of the state of the art,” *Psychol Methods*, vol. 7, no. 2, pp. 147–177, 2002, doi: 10.1037/1082-989X.7.2.147.
- [11] D. B. Rubin, “Inference and missing data,” *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976, doi: 10.1093/biomet/63.3.581.
- [12] S. van Buuren and K. Groothuis-Oudshoorn, “mice: Multivariate imputation by chained equations in R,” *J Stat Softw*, vol. 45, no. 3, pp. 1–67, 2011, doi: 10.18637/jss.v045.i03.
- [13] D. J. Stekhoven and P. Bühlmann, “Missforest-Non-parametric missing value imputation for mixed-type data,” *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012, doi: 10.1093/bioinformatics/btr597.
- [14] J. Yoon, W. R. Zame, and M. Van Der Schaar, “Estimating Missing Data in Temporal Data Streams Using Multi-Directional Recurrent Neural Networks,” *IEEE Trans Biomed Eng*, vol. 66, no. 5, pp. 1477–1490, 2019, doi: 10.1109/TBME.2018.2874712.
- [15] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, “Recurrent Neural Networks for Multivariate Time Series with Missing Values,” *Sci Rep*, vol. 8, no. 1, p. 6085, 2018, doi: 10.1038/s41598-018-24271-9.
- [16] X. Yao *et al.*, “Prediction of f-CaO content in cement clinker using a GRU-based deep learning model with masked-attention mechanism for incomplete DCS data,” *J Sustain Cem Based Mater*, vol. 14, no. 7, pp. 1413–1426, 2025, doi: 10.1080/21650373.2025.2511769.
- [17] H. Shah, H. Farman, B. Jan, A. Khalil, and M. M. Nasralla, “Securing the Internet of Things: Deep Learning Driven Intrusion Detection with Missing Data Imputation,” *IEEE Access*, 2025.
- [18] J. Yao, W. Xu, G. Zhu, K. Huang, and S. Cui, “Energy-Efficient Edge Inference in Integrated Sensing, Communication, and Computation Networks,” *IEEE Journal on Selected Areas in Communications*, 2025, doi: 10.1109/JSAC.2025.3574612.
- [19] L. Meng and Q. Li, “Edge learning for real-time sensor data imputation,” *IEEE Internet Things J*, vol. 8, no. 20, pp. 15139–15148, 2021, doi: 10.1109/JIOT.2021.3089008.

- [20] S. Perdakis, R. Leeb, R. Chavarriaga, and J. D. R. Millan, "Context-Aware Learning for Generative Models," *IEEE Trans Neural Netw Learn Syst*, vol. 32, no. 8, pp. 3471–3483, 2021, doi: 10.1109/TNNLS.2020.3011671.
- [21] H. Lin, K. Kaur, X. Wang, G. Kaddoum, J. Hu, and M. M. Hassan, "Privacy-Aware Access Control in IoT-Enabled Healthcare: A Federated Deep Learning Approach," *IEEE Internet Things J*, vol. 10, no. 4, pp. 2893–2902, 2023, doi: 10.1109/JIOT.2021.3112686.
- [22] A. Alrawajfi, M. T. Ismail, S. Al Wadi, S. Atiewi, and A. Awajan, "Multiple imputation methods: a case study of daily gold price," *PeerJ Comput Sci*, vol. 10, p. e2337, 2024.
- [23] A. M. H, "Imputation Techniques for Missing Data in IoT Sensor Networks," 2025, [Online]. Available: <https://www.researchgate.net/publication/389502780>
- [24] M. Zhang, R. Zhao, C. Wang, L. Jing, and D. Li, "Real-Time Imputation Model for Missing Sensor Data Based on Alternating Attention Mechanism," *IEEE Sens J*, vol. 25, no. 5, pp. 8962–8974, 2025, doi: 10.1109/JSEN.2024.3519370.
- [25] S. M. R. Islam, D. Kwak, M. H. Kabir, M. Hossain, and K. S. Kwak, "The internet of things for health care: A comprehensive survey," *IEEE Access*, vol. 3, pp. 678–708, 2015, doi: 10.1109/ACCESS.2015.2437951.
- [26] N. Chaabane, S. Mahfoudhi, and K. Belkadhi, "Interpolation-Based IoT Sensors Selection," *IEEE Sens J*, vol. 24, no. 21, pp. 36143–36147, 2024, doi: 10.1109/JSEN.2024.3461833.
- [27] J. Ryan-Despraz and A. Wissler, "Imputation methods for mixed datasets in bioarchaeology," *Archaeol Anthropol Sci*, vol. 16, no. 11, p. 187, 2024, doi: 10.1007/s12520-024-02078-2.
- [28] P. Bodik, W. Hong, C. Guestrin, S. Madden, M. Paskin, and R. Thibaux, "Intel Berkeley Research Lab Data," 2004. [Online]. Available: <http://db.csail.mit.edu/labdata/labdata.html>

Notes on contributors



Amer Al-rahayfeh     received his Ph.D. in Computer Science and Engineer from the University of Bridgeport (U.S.A) in 2014. He is currently an associate professor of computer sciences at Al-Hussein Bin Talal University (AHU) in Jordan. His research interests are in the areas of multimedia systems, computer vision, sensor networks, cloud computing biomedical systems. At AHU, he led the Department of Computer Science and vice dean of college of Information Technology. Email: amer.a.al-rahayfeh@ahu.edu.jo