Evaluating the Data Imputation Impact on Gradient Boosting Model Predictive Performance in Sensor Failure Recovery for Smart Irrigation Systems

Miftahul Walid
Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia | Department of Informatics Engineering, Universitas Islam Madura, Pemekasan, Indonesia
miftahul.walid.2305349@students.um.ac.id

Muhammad Ashar
Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia
muhammad.ashar.ft@um.ac.id (corresponding author)

Heru Wahyu Herwanto
Department of Electrical Engineering and Informatics, Universitas Negeri Malang, Malang, Indonesia
heru_wh@um.ac.id

Received: 27 January 2026 | Revised: 19 February 2026, 28 February 2026, 2 March 2026, and 3 March 2026 | Accepted: 4 March 2026

 DOI: https://doi.org/10.48084/etasr.17776

ABSTRACT

The reliability of Internet of Things (IoT)-based irrigation systems depends heavily on the quality of data acquired from various sensors. In real-world applications, missing values and faulty data, caused by transmission faults or sensor failures, can severely compromise the application of Machine Learning (ML), particularly in critical tasks such as detecting pump failures. This study investigates the impact of missing data imputation on the performance of an Extreme Gradient Boosting (XGBoost) classifier under a controlled setting. Missing data were simulated at rates ranging from 5% to 30%, and the imputation techniques applied included Multiple Imputation by Chained Equations (MICE), k-Nearest Neighbors (KNN), Random Forest (RF), and iterative gradient-boosting methods. The quality of the imputed data was evaluated using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), whereas classifier performance was assessed using accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC). For 5% missing data, MICE achieved the lowest errors (MAE = 0.018, RMSE = 0.026), whereas at 30% missing data, errors increased (MAE = 0.072, RMSE = 0.094). Classifier performance declined with increasing missing data: accuracy decreased from 0.972 to 0.818 and recall from 0.961 to 0.742. The application of MICE mitigated this decline, maintaining an accuracy of 0.982 and recall of 0.975 at 5% missing data, and an accuracy of 0.863 and recall of 0.812 at 30% missing data, with AUC remaining above 0.880 in both cases. Notably, the imputation method with the lowest MAE and RMSE did not always produce the best classification performance, indicating that numerical precision does not necessarily translate into improved classification skill. These results highlight the importance of selecting imputation techniques based on the specific nature of the problem to preserve feature integrity and sustain the performance of Artificial Intelligence (AI)-based smart irrigation systems in real-world sensor environments.

Keywords-data imputation; missing data; XGBoost; predictive maintenance; smart irrigation; sensor data quality; Machine Learning (ML); IoT systems

I. INTRODUCTION

The transformation toward data-driven smart irrigation has been increasingly driven by the availability of low-cost soil and weather sensors, along with Internet of Things (IoT) connectivity that enables real-time monitoring of field conditions [1, 2]. Multivariate data such as temperature, humidity, solar radiation, wind speed, and soil moisture have been widely utilized for water demand prediction and automated decision-making. The integration of IoT and

Machine Learning (ML) in modern irrigation systems has been extensively reported, covering system architectures, data pipelines, and water-use optimization strategies [3].

However, the dependability of ML-driven irrigation systems heavily relies on the quality of sensor data. In real-world field situations, IoT data often experience missing values attributable to communication issues, sensor drift and bias, corrupted measurements from environmental disruptions, and significant outliers. Unreliable sensor nodes have been shown to generate erroneous measurements that can potentially mislead irrigation decisions [4, 5]. More broadly, the IoT literature emphasizes that faults, anomalies, and outliers constitute major challenges that degrade the reliability of analytical pipelines and the performance of ML models [6].

To mitigate these issues, most existing systems adopt quality-check mechanisms based on anomaly detection. Such approaches have been applied in smart farming contexts to filter sensor failures and data disturbances [7], as well as in Artificial Intelligence (AI)-based irrigation systems capable of detecting and reconstructing abnormal sensor signals using deep learning models [8]. However, many studies stop at the detection stage, where data identified as faulty are not always restored, resulting in incomplete or inconsistent inputs for downstream prediction models.

Imputation offers a potential solution to bridge this gap. Beyond filling missing values, imputation can be positioned as a data-healing mechanism by treating corrupted or detected outlier values as missing and subsequently estimating more representative values based on inter-feature relationships and temporal–spatial patterns [9]. In sensor networks and IoT systems, prior studies have demonstrated that reliable imputation is essential to maintain data continuity under sensor failures and operational disruptions [10]. A wide range of imputation methods has been proposed, spanning classical approaches to deep learning–based models, attention-based imputation, spatiotemporal graph models [8-10], and more recently, generative diffusion-based approaches that are particularly relevant for complex sensor data [11-14].

Once data quality has been restored, ML models such as gradient boosting are commonly employed to predict key variables in smart irrigation, including soil moisture, to support efficient irrigation scheduling [15]. Furthermore, modern data-driven control approaches, such as data-driven model predictive control, increasingly integrate ML models for adaptive irrigation decision-making [16]. These developments highlight that system performance is determined not only by the predictive model itself but also by the quality of sensor data at the early stages of the pipeline.

Motivated by these gaps, this study evaluates imputation as a core module within a smart irrigation pipeline to recover missing, corrupted, and outlier-contaminated sensor data prior to their use in pump status classification. The main contribution of this work is a comparative evaluation of system performance using quality-checked data without imputation and quality-checked data restored through multiple imputation strategies. This end-to-end perspective underscores the importance of integrating data acquisition, data quality enhancement, and decision-making, as emphasized in IoT- and ML-based automated irrigation systems [17].

II. METHOD

The dataset used in this study is a secondary dataset obtained from the publicly available Mendeley Data repository [18]. The dataset contains multivariate sensor measurements collected from an IoT-based smart irrigation system, including environmental and operational parameters relevant to pump performance monitoring. This dataset has also been utilized in previous studies [19], indicating its empirical validity and prior application in similar experimental contexts. The use of a publicly accessible dataset ensures transparency, reproducibility, and comparability of the experimental results.

The experimental framework represented in the flowchart (Figure 1) presents a methodologically rigorous approach to evaluating the impact of data quality and missing-value imputation procedures on the performance of a fixed Extreme Gradient Boosting (XGBoost) classification model [20]. Initially, a clean sensor dataset is partitioned into training and testing subsets using a standard train–test split strategy to prevent information leakage and preserve the independence of evaluation results. The clean training subset is used to build a fixed XGBoost model, which serves as the basis for all subsequent classification tasks. This approach ensures that variability in classification performance can be attributed to data quality and imputation strategy rather than changes in model parameters.

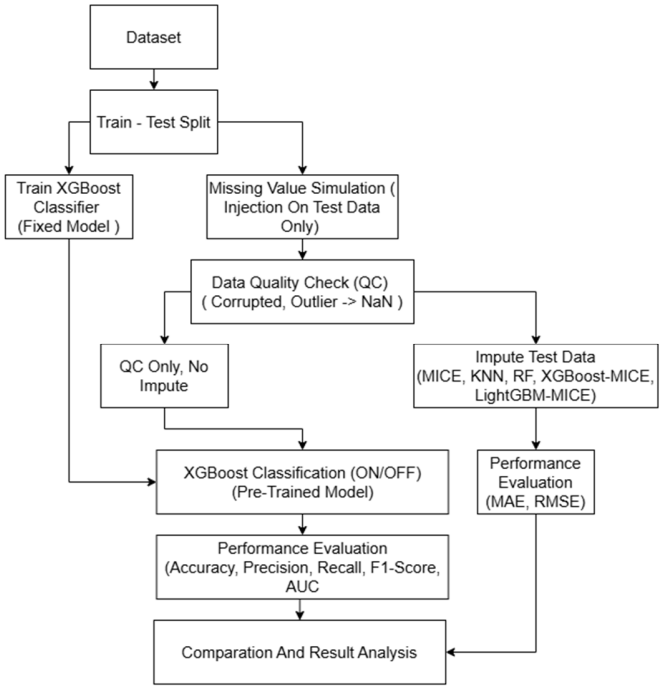

Fig. 1. Experimental framework flowchart.

The XGBoost algorithm, a decision-tree–based gradient boosting framework, constructs an ensemble of weak learners

to minimize a regularized objective function [21]. The training objective in XGBoost can be formalized as:

$$\mathcal{L}(\emptyset) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (1)$$

where $\mathcal{L}(\emptyset)$ denotes the loss between true labels y_i and predicted outputs $\hat{y}_i$; f_k is the k-th decision tree; $\Omega(f_k)$ is a regularization term controlling tree complexity; and K is the total number of trees. This formulation integrates model accuracy and complexity to achieve robust generalization [22].

To simulate real-world sensor conditions, artificial missing values are injected into the clean test set. A Quality Control (QC) process is then applied to identify corrupted values, outliers, and NaNs, creating two main experimental branches: test sets with missing values retained (no imputation) and test sets with missing values imputed by state-of-the-art techniques. Multiple imputation methods such as Multiple Imputation by Chained Equations (MICE) using Bayesian Ridge regression, k-Nearest Neighbors (KNN) Imputer [23], Random Forest (RF) Iterative imputation (MissForest-like), XGBoost-MICE, and Light Gradient Boosting Machine (LightGBM)-MICE are employed with a hyperparameter tuning process. In this research, the hyperparameter configuration of each model was determined through a combination of theoretical considerations and empirical evaluation to ensure a balance between predictive performance and computational efficiency.

For the RF Iterative (MissForest-like) imputation approach, the RF Regressor from Scikit-learn was configured with n_estimators = 300, selected based on preliminary experiments within the range of 100–400 trees, where Root Mean Square Error (RMSE) stabilization was observed beyond approximately 250 trees, indicating diminishing performance gains with additional estimators. The max_features parameter was set to "sqrt" to reduce inter-tree correlation and improve generalization, whereas max_depth was left at its default setting to flexibly capture nonlinear relationships among sensor variables.

For KNN-based imputation, the KNN Imputer employed k = 5, determined through comparative testing across k ∈ {3, 5, 7, 9}, where k = 5 yielded the lowest RMSE and reflected an appropriate bias–variance trade-off, as smaller k values were more sensitive to noise and larger values produced oversmoothing effects.

Boosting-based imputation strategies were implemented within the MICE framework using both XGBoost and LightGBM as regression estimators. In the XGBoost-MICE configuration, the parameters were set to n_estimators = 600, learning rate = 0.05, max_depth = 4, subsample = 0.9, and colsample_bytree = 0.9 to ensure gradual and stable convergence during iterative imputation while controlling model complexity. In contrast, in the LightGBM-MICE configuration, n_estimators = 600 and learning rate = 0.05 were adopted to maintain stable boosting dynamics, with num_leaves = 31 providing sufficient representational capacity and subsample and colsample_bytree set to 0.9 to enhance robustness against data variability.

The primary classification model also utilized XGBoost with the same boosting configuration (n_estimators = 600, learning rate = 0.05, max_depth = 4, subsample = 0.9, colsample_bytree = 0.9), and all models were initialized with a fixed random state to ensure reproducibility and methodological transparency.

The general principle of an imputation model $g(\cdot)$ is to estimate missing components $\hat{x}_{ij}$ of the data vector x_i from the observed features $x_{i,-j}$ as:

$$\hat{x}_{ij} = g\left(x_{i,-j}, \Theta\right) \quad (2)$$

where Θ represents model parameters learned from the observed data. ML-based imputation has demonstrated improved performance in handling complex missing data patterns compared with traditional statistical methods [24].

Once test sets (with and without imputation) are prepared, the fixed XGBoost model classifies the observations, and its performance is evaluated using standard classification metrics including Accuracy, Precision, Recall, F1-Score, and Area Under the Receiver Operating Characteristic Curve (AUC). These metrics provide a multidimensional assessment of classifier effectiveness under varying data conditions. In parallel, the quality of imputation itself is quantified using Mean Absolute Error (MAE) and RMSE between imputed values and original ground truth, defined respectively as:

$$\text{MAE} = \frac{1}{m} \sum_{i=1}^{m} |x_i - \hat{x}_i| \quad (3)$$

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (x_i - \hat{x}_i)} \quad (4)$$

where m is the number of imputed entries.

Current studies highlight the necessity of assessing both classification and imputation effectiveness to choose the best preprocessing pipelines, especially in data-heavy fields [25]. The flowchart facilitates a comparative evaluation of different imputation techniques, allowing researchers to methodically determine which approach most effectively preserves predictive accuracy in various missing data situations. The empirical analysis derived from this framework provides insights into how data integrity relates to the effectiveness of supervised learning, which is crucial for applying ML models in real-world intelligent systems.

III. RESULTS

A. Imputation Performance Analysis

The effectiveness of data imputation was methodically assessed across six scenarios of missing data, varying from 5% to 30%. The effectiveness of reconstruction for each imputation technique was evaluated by MAE (Figure 2) and RMSE (Figure 3), which are standard metrics for measuring numerical discrepancies between imputed values and the actual observed data.

The experimental results consistently indicate that imputation errors increase with the rise in missing data, regardless of the imputation technique applied. At low missing rates (5–10%), all evaluated methods exhibit relatively stable reconstruction performance, suggesting that the available contextual information remains sufficient to support accurate

estimation. However, when the missing rate exceeds 20%, both MAE and RMSE rise markedly, reflecting the increasing difficulty of reconstructing sensor signals under severe data loss conditions.

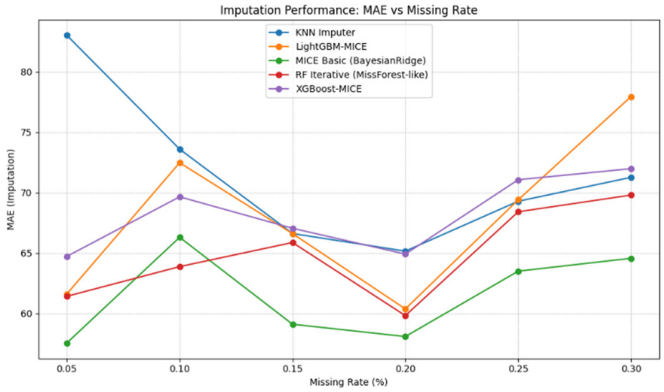

Fig. 2. MAE vs missing data rate.

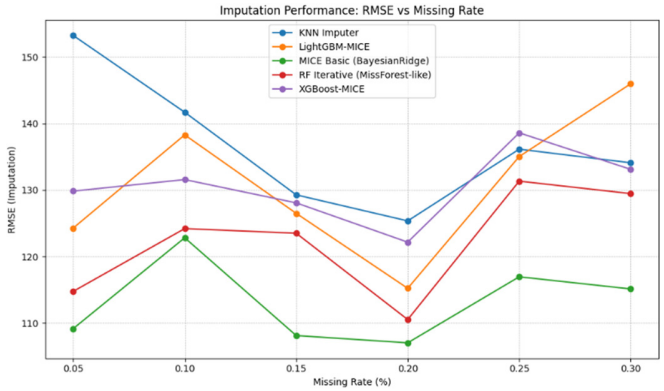

Fig. 3. RMSE vs missing data rate.

Among all evaluated approaches, MICE implemented with Bayesian Ridge regression consistently achieves the lowest MAE and RMSE values across nearly all missing-rate scenarios. This finding highlights the robustness of probabilistic linear models in capturing inter-feature dependencies within industrial sensor data, particularly when the relationships among variables are relatively stable and approximately linear. These results are consistent with prior studies reporting that MICE-based approaches can outperform more complex models in numerical reconstruction tasks under moderate missing-data conditions [26].

In contrast, tree-based and gradient-boosting imputation methods, including RF Iterative, XGBoost-MICE, and LightGBM-MICE, tend to produce higher reconstruction errors, especially at elevated missing rates. Although these methods are capable of modeling nonlinear relationships, their reliance on recursive partitioning may amplify noise and variance when the effective sample size is substantially reduced. As a result, this conduct results in suboptimal point-wise reconstruction accuracy in scenarios with significant missing data [27].

However, it is crucial to highlight that numerical reconstruction accuracy alone does not entirely reflect the practical usefulness of an imputation technique, especially when the imputed data are later employed for predictive modeling. This limitation motivates the evaluation of downstream predictive performance, which is addressed in the following subsection.

B. Predictive Performance Results

To assess the downstream impact of imputation quality on predictive modeling, the reconstructed datasets were employed to train an XGBoost classification model aimed at predicting pump failure events. The performance of the classifier was evaluated using standard metrics, including accuracy, precision, recall, F1-score, and AUC, as illustrated in Figures 4–8 [28, 29]. In addition to the imputed datasets, a comparative baseline scenario was included, consisting of raw data subjected solely to QC without any imputation.

Overall, the results demonstrate a systematic degradation in predictive performance as the proportion of missing data rate increases, irrespective of the preprocessing strategy applied. This decline is particularly pronounced in recall and F1-score, which are essential indicators for failure diagnosis and early warning systems, where undetected failures may lead to severe operational consequences. These results highlight the importance of effective data reconstruction for reliable predictions in real-world IoT systems.

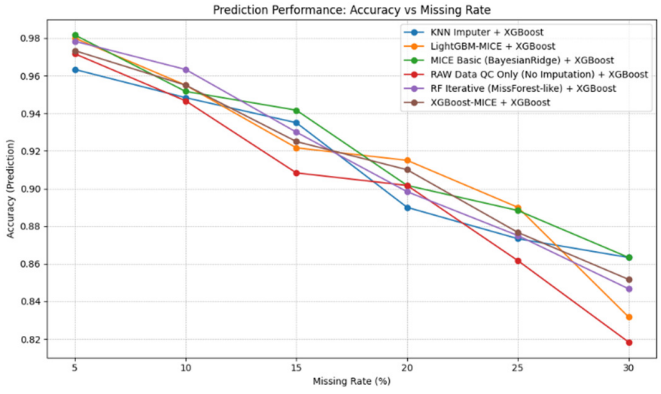

Fig. 4. Prediction performance: accuracy vs missing data rate.

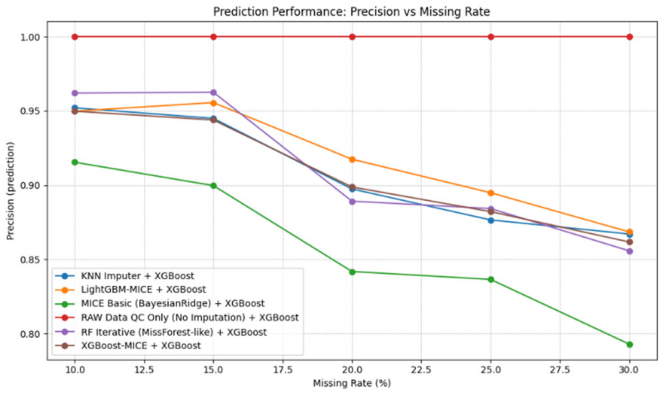

Fig. 5. Prediction performance: precision vs missing data rate.

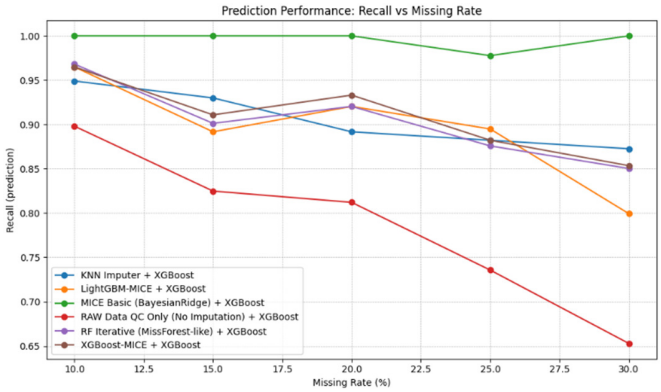

Fig. 6. Prediction performance: recall vs missing data rate.

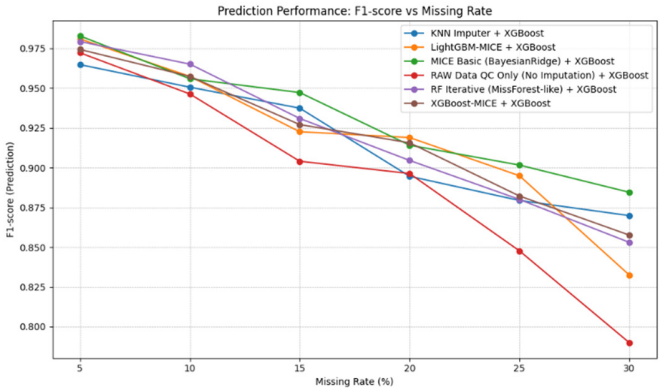

Fig. 7. Prediction performance: F1-score vs missing data rate.

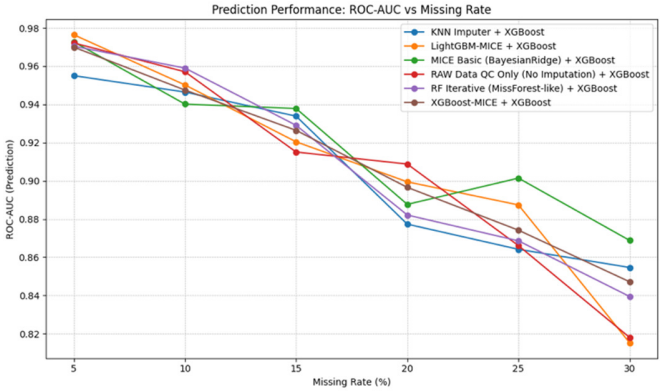

Fig. 8. Prediction performance: AUC vs missing data rate.

With low missing rates (5–10%), models trained on raw data continue to show fairly competitive performance, indicating that minor data loss does not instantly undermine classification accuracy (Table I–VI). Nonetheless, when the missing rate hits 15% or higher, the raw data scenario experiences a significant drop in recall, suggesting an increasing inclination toward false negatives. This finding highlights the inherent risk of using minimally processed data, where high accuracy values may obscure critical detection failures.

The application of QC without imputation improves predictive stability to some extent; however, its effectiveness remains suboptimal due to the absence of explicit data reconstruction. This limitation restricts the model's ability to exploit the underlying temporal and multivariate patterns embedded in the sensor data. In contrast, the use of imputed datasets generally yields more robust and consistent predictive performance, particularly under moderate to high missing-rate conditions.

TABLE I. ACCURACY FOR 5% MISSING DATA

Method	Accuracy
MICE Basic (BayesianRidge) + XGBoost	0.982
LightGBM-MICE + XGBoost	0.980
RF Iterative (MissForest-like) + XGBoost	0.978
XGBoost-MICE + XGBoost	0.973
Raw data, QC only (no imputation) + XGBoost	0.972
KNN Imputer + XGBoost	0.963

TABLE II. ACCURACY FOR 10% MISSING DATA

Method	Accuracy
RF Iterative (MissForest-like) + XGBoost	0.963
LightGBM-MICE + XGBoost	0.955
XGBoost-MICE + XGBoost	0.955
MICE Basic (BayesianRidge) + XGBoost	0.951
KNN Imputer + XGBoost	0.948
Raw data, QC only (no imputation) + XGBoost	0.946

TABLE III. ACCURACY FOR 15% MISSING DATA

Method	Accuracy
MICE Basic (BayesianRidge) + XGBoost	0.941
KNN Imputer + XGBoost	0.935
RF Iterative (MissForest-like) + XGBoost	0.930
XGBoost-MICE + XGBoost	0.925
LightGBM-MICE + XGBoost	0.921
Raw data, QC only (no imputation) + XGBoost	0.908

TABLE IV. ACCURACY FOR 20% MISSING DATA

Method	Accuracy
LightGBM-MICE + XGBoost	0.915
XGBoost-MICE + XGBoost	0.910
MICE Basic (BayesianRidge) + XGBoost	0.901
Raw data, QC only (no imputation) + XGBoost	0.901
RF Iterative (MissForest-like) + XGBoost	0.898
KNN Imputer + XGBoost	0.890

TABLE V. ACCURACY FOR 25% MISSING DATA

Method	Accuracy
LightGBM-MICE + XGBoost	0.890
MICE Basic (BayesianRidge) + XGBoost	0.888
XGBoost-MICE + XGBoost	0.876
RF Iterative (MissForest-like) + XGBoost	0.875
KNN Imputer + XGBoost	0.873
Raw data, QC only (no imputation) + XGBoost	0.861

TABLE VI. ACCURACY FOR 30% MISSING DATA

Method	Accuracy
MICE Basic (BayesianRidge) + XGBoost	0.863
KNN Imputer + XGBoost	0.863
XGBoost-MICE + XGBoost	0.851
RF Iterative (MissForest-like) + XGBoost	0.846
LightGBM-MICE + XGBoost	0.831
Raw data, QC only (no imputation) + XGBoost	0.818

Interestingly, although MICE achieves the lowest imputation error in terms of MAE and RMSE, it does not consistently produce the best predictive metrics. In several scenarios, tree-based and boosting-based imputation methods result in higher recall or AUC values, despite exhibiting larger numerical reconstruction errors. This discrepancy suggests that preserving discriminative feature structures is more critical for predictive modeling than minimizing point-wise numerical imputation error alone. Such findings reinforce the notion that imputation methods should be evaluated not only by reconstruction accuracy but also by their impact on downstream predictive tasks.

Tables I–IV summarize the accuracy of XGBoost for missing data rates from 5% to 30% for different methods. As the missing data rate increases, model accuracy decreases. When 5% of the data is missing, all methods perform well, with accuracy above 0.96. MICE with Bayesian Ridge achieves the highest accuracy at 0.982, followed by LightGBM-MICE at 0.980 and RF Iterative at 0.978. Without any imputation, QC still works pretty well with an accuracy of 0.972. This means that when missing data are minimal, sufficient information remains to maintain model stability.

When the missing data rate goes up to 10% the differences among methods become apparent. RF Iterative achieves the highest accuracy at 0.963, whereas raw data without imputation drops to 0.946. The difference is modest, indicating that missing data are beginning to have a more noticeable impact on model performance. At 15% missing data things get worse. MICE is still the best at 0.941, whereas not doing imputation drops to 0.908. The gap is getting bigger showing that filling in the missing data is becoming more important for keeping the models performance stable.

At 20% missing data all the methods show reduced accuracy. Some methods like LightGBM-MICE are still doing better with an accuracy of 0.915 compared to 0.901 for not doing imputation. This pattern continues at 25% missing data, where LightGBM-MICE and MICE achieve 0.890 and 0.888, respectively, whereas raw data drops to 0.861. When 30% of the data are missing the difference between the methods is clear. MICE and KNN both reach 0.863, whereas not doing imputation drops to 0.818. The 4–5% difference shows that filling in the missing data makes a difference, particularly at high missing rates.

In general, these results show that missing data reduce model accuracy, but using appropriate imputation methods helps maintain predictive performance. No single method is always the best. MICE and methods that use boosting are generally more stable than not doing any imputation. These findings indicate that QC alone is insufficient for maintaining model reliability in sensor-based systems operating under real-world conditions with frequent data loss.

IV. DISCUSSION

The most significant finding of this study is that imputation accuracy, as quantified by MAE and RMSE, does not exhibit a linear or proportional relationship with downstream predictive performance. This observation challenges the common assumption that minimizing reconstruction error will inherently lead to superior ML outcomes.

Fundamentally, MAE and RMSE assess local numerical deviations, treating each imputed value independently of the broader data structure. In contrast, predictive models, especially ensemble-based classifiers like XGBoost, show significant sensitivity to overall data characteristics, such as feature distributions, variance structures, interactions between features, and the separability of classes. As a result, imputation techniques that overly focus on minimizing local errors can create excessively smoothed datasets, decreasing variance and diminishing the discriminative signals among classes [30-32].

Tree-based and gradient-boosting imputation methods, although less accurate in terms of numerical reconstruction, tend to better preserve nonlinear dependencies and heterogeneous variance patterns that are critical for effective classification. Similar observations have been reported in recent studies on fault diagnosis and prognostics, where imputation techniques associated with higher reconstruction error nonetheless yielded superior predictive performance [29]. These findings indicate that maintaining structurally informative feature representations may be more important than achieving minimal point-wise reconstruction error.

The comparison with the raw data scenario subjected solely to QC without imputation further underscores the necessity of explicitly tackling missing data. Disregarding missing values can lead to structural bias and significantly reduce recall, a metric that is especially crucial in safety-sensitive industrial settings [23, 24]. In predictive maintenance systems, false negatives can lead to missed failures, extended downtime, and higher operational risk, making these results intolerable in practical applications.

Overall, the results of this study align with a growing body of research advocating for a task-specific assessment of imputation methods, rather than relying exclusively on reconstruction accuracy metrics [27]. Therefore, developing unified evaluation frameworks that simultaneously evaluate imputation quality and ensuing predictive performance is essential for establishing reliable, applicable, and safety-critical data-driven predictive maintenance systems.

V. CONCLUSIONS

The experimental findings demonstrate that the quality of data imputation significantly influences predictive performance; yet, the precision of numerical reconstruction alone is insufficient as a metric for imputation effectiveness in predictive modeling contexts. The findings demonstrate that an increased incidence of missing data consistently undermines the predictive effectiveness of the Extreme Gradient Boosting (XGBoost) model. Without imputation, accuracy declined from 0.972 with 5% missing data to 0.818 with 30%, whereas recall decreased from 0.961 to 0.742. This drop is particularly significant, as recall and F1-score are essential metrics in failure diagnostics and predictive maintenance systems, where the prevention of false negatives is crucial.

The evaluation of imputation effectiveness via Mean Absolute Error (MAE) and Root Mean Square Error (RMSE)

indicates that Multiple Imputation by Chained Equations (MICE) combined with Bayesian Ridge regression achieved the lowest reconstruction errors at different missing-rate levels, with MAE increasing from 0.018 at 5% missing data to 0.072 at 30%, and RMSE rising from 0.026 to 0.094. A key finding of this research is that the imputation method yielding the lowest MAE and RMSE does not consistently generate the most accurate predicted outcomes. MICE maintained strong performance, achieving an accuracy of 0.982 and a recall of 0.975 even with 5% missing data. It sustained an accuracy of 0.863 and a recall of 0.812 despite 30% of the data being absent. Nonetheless, alternative tree-based and gradient-boosting imputation techniques demonstrated competitive or superior recall and Area Under the Receiver Operating Characteristic Curve (AUC) in certain missing-rate scenarios, despite exhibiting higher reconstruction errors.

The findings confirm that preserving global data structures, including feature distributions, variance characteristics, and nonlinear inter-feature relationships, is more critical for classification effectiveness than simply minimizing point-wise numerical error. Furthermore, the comparison with the raw data (quality-check-only) scenario underscores that disregarding missing information leads to a significant decline in predictive accuracy and a elevated risk of false negatives, which is unacceptable in safety-critical Internet of Things (IoT)-based irrigation systems.

Future research could extend this investigation in several ways. First, to better capture temporal dependencies and spatial correlations in multivariate sensor data, it is crucial to explore advanced deep learning imputation methods, including autoencoders, variational autoencoders, transformer models, and temporal graph neural networks. Second, adaptive or task-aware imputation frameworks could be developed to optimize reconstruction relative to downstream prediction goals. Third, analyzing multiple approaches to missing data, such as Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing not at Random (MNAR), would enhance understanding of model resilience under real-world data degradation. Finally, integrating agronomic factors, environmental conditions, and real-time implementation limitations into the evaluation framework could improve the effectiveness of the suggested method in large-scale smart irrigation systems.

ACKNOWLEDGMENT

The author would like to express sincere gratitude to the Faculty of Engineering, Department of Electrical Engineering and Informatics, Universitas Negeri Malang, for the academic support, facilities, and conducive scholarly environment that enabled this research to be conducted and completed successfully. This support was particularly instrumental in facilitating the author's doctoral research activities and made a significant contribution to the smooth execution of the research and the preparation of this scientific manuscript.

REFERENCES

[1] Mokh. S. Hadi, P. Adi Nugraha, I. M. Wirawan, I. Ari Elbaith Zaeni, M. A. Mizar, and M. Irvan, "IoT Based Smart Garden Irrigation System," in *2020 4th International Conference on Vocational Education and Training*, Malang, Indonesia, 2020, pp. 361–365, https://doi.org/10.1109/ICOVET50258.2020.9230197.

[2] D. Kurniawan, R. J. Putra, A. Bella, M. Ashar, and K. Dedes, "Smart Garden with IoT Based Real Time Communication using MQTT Protocol," in *2021 7th International Conference on Electrical, Electronics and Information Engineering*, Malang, Indonesia, 2021, pp. 1–5, https://doi.org/10.1109/ICEEIE52663.2021.9616869.

[3] L. García, L. Parra, J. M. Jimenez, J. Lloret, and P. Lorenz, "IoT-Based Smart Irrigation Systems: An Overview on the Recent Trends on Sensors and IoT Systems for Irrigation in Precision Agriculture," *Sensors*, vol. 20, no. 4, Feb. 2020, Art. no. 1042, https://doi.org/10.3390/s20041042.

[4] J. Ludeña-Choez, J. J. Choquehuanca-Zevallos, and E. Mayhua-López, "Sensor nodes fault detection for agricultural wireless sensor networks based on NMF," *Computers and Electronics in Agriculture*, vol. 161, pp. 214–224, June 2019, https://doi.org/10.1016/j.compag.2018.06.033.

[5] A. H. Blasi, M. A. Abbadi, and R. Al-Huweimel, "Machine Learning Approach for an Automatic Irrigation System in Southern Jordan Valley," *Engineering, Technology & Applied Science Research*, vol. 11, no. 1, pp. 6609–6613, Feb. 2021, https://doi.org/10.48084/etasr.3944.

[6] A. Gaddam, T. Wilkin, M. Angelova, and J. Gaddam, "Detecting Sensor Faults, Anomalies and Outliers in the Internet of Things: A Survey on the Challenges and Solutions," *Electronics*, vol. 9, no. 3, Mar. 2020, Art. no. 511, https://doi.org/10.3390/electronics9030511.

[7] A. R. de A. Zanella, E. da Silva, and L. C. P. Albini, "CEIFA: A multi-level anomaly detector for smart farming," *Computers and Electronics in Agriculture*, vol. 202, Nov. 2022, Art. no. 107279, https://doi.org/10.1016/j.compag.2022.107279.

[8] R. Benameur, A. Dahane, B. Kechar, and A. E. H. Benyamina, "An Innovative Smart and Sustainable Low-Cost Irrigation System for Anomaly Detection Using Deep Learning," *Sensors*, vol. 24, no. 4, Feb. 2024, Art. no. 1162, https://doi.org/10.3390/s24041162.

[9] R. Sahu and P. Tripathi, "A novel data healing framework for outlier detection and recovery approach in the internet of agriculture things sensor network," *Quality & Quantity*, Oct. 2025, https://doi.org/10.1007/s11135-025-02402-5.

[10] B. Agbo, H. Al-Aqrabi, R. Hill, and T. Alsboui, "Missing Data Imputation in the Internet of Things Sensor Networks," *Future Internet*, vol. 14, no. 5, May 2022, Art. no. 143, https://doi.org/10.3390/fi14050143.

[11] F. Lalande and K. Doya, "Numerical Data Imputation: Choose kNN over Deep Learning," in *5th International Conference on Similarity Search and Applications*, Bologna, Italy, 2022, pp. 3–10, https://doi.org/10.1007/978-3-031-17849-8_1.

[12] Y.-F. Zhang, P. J. Thorburn, W. Xiang, and P. Fitch, "SSIM—A Deep Learning Approach for Recovering Missing Time Series Sensor Data," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 6618–6628, Aug. 2019, https://doi.org/10.1109/JIOT.2019.2909038.

[13] F. Chen, D. Wang, S. Lei, J. He, Y. Fu, and C.-T. Lu, "Adaptive graph convolutional imputation network for environmental sensor data recovery," *Frontiers in Environmental Science*, vol. 10, Nov. 2022, Art. no. 1025268, https://doi.org/10.3389/fenvs.2022.1025268.

[14] D. Jiang and S. Zhang, "An explainable missing data imputation method and its application in soft sensing," *Measurement*, vol. 253, Sept. 2025, Art. no. 117692, https://doi.org/10.1016/j.measurement.2025.117692.

[15] F. T. Teshome, H. K. Bayabil, B. Schaffer, Y. Ampatzidis, and G. Hoogenboom, "Improving soil moisture prediction with deep learning and machine learning models," *Computers and Electronics in Agriculture*, vol. 226, Nov. 2024, Art. no. 109414, https://doi.org/10.1016/j.compag.2024.109414.

[16] E. Bwambale, F. K. Abagale, and G. K. Anornu, "Data-driven model predictive control for precision irrigation management," *Smart Agricultural Technology*, vol. 3, Feb. 2023, Art. no. 100074, https://doi.org/10.1016/j.atech.2022.100074.

[17] A. A. Abdelmoneim, H. N. Kimaita, C. M. A. Kalaany, B. Derardja, G. Dragonetti, and R. Khadra, "IoT Sensing for Advanced Irrigation Management: A Systematic Review of Trends, Challenges, and Future

Prospects," *Sensors*, vol. 25, no. 7, Apr. 2025, Art. no. 2291, https://doi.org/10.3390/s25072291.

[18] A. Kaur, D. P. Bhatt, and L. Raja, "Soil Moisture, Air temperature, humidity, and Motor on/off Monitoring data." Mendeley Data, July 24, 2023, https://doi.org/10.17632/fpdwmm7nrb.1.

[19] A. Kaur, D. P. Bhatt, and L. Raja, "Developing a Hybrid Irrigation System for Smart Agriculture Using IoT Sensors and Machine Learning in Sri Ganganagar, Rajasthan," *Journal of Sensors*, vol. 2024, no. 1, Jan. 2024, Art. no. 6676907, https://doi.org/10.1155/2024/6676907.

[20] H. Hairani, T. Widiyaningtyas, D. D. Prasetya, G. Y. Pratama, K. Hidjah, and S. Soraya, "Multi-class Classification of Obesity Levels Using Gradient Boosting, Random Forest, and C4.5," in *2025 9th International Conference On Electrical, Electronics And Information Engineering*, Mataram, Indonesia, 2025, pp. 1–5, https://doi.org/10.1109/ICEEIE66203.2025.11253679.

[21] S. Robo, T. Widiyaningtyas, and W. Sakti, "HCF-MFGB: Hybrid Collaborative Filtering Based on Matrix Factorization and Gradient Boosting," *Computers, Materials & Continua*, vol. 86, no. 2, pp. 1–19, Dec. 2025, https://doi.org/10.32604/cmc.2025.073011.

[22] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

[23] D. M. P. Murti, U. Pujianto, A. P. Wibawa, and M. I. Akbar, "K-Nearest Neighbor (K-NN) based Missing Data Imputation," in *2019 5th International Conference on Science in Information Technology*, Yogyakarta, Indonesia, 2019, pp. 83–88, https://doi.org/10.1109/ICSITech46713.2019.8987530.

[24] Z. Jinbo, L. Yufu, and M. Haitao, "Handling missing data of using the XGBoost-based multiple imputation by chained equations regression method," *Frontiers in Artificial Intelligence*, vol. 8, Apr. 2025, Art. no. 1553220, https://doi.org/10.3389/frai.2025.1553220.

[25] Q. A. Hidayaturrohman and E. Hanada, "Impact of Data Pre-Processing Techniques on XGBoost Model Performance for Predicting All-Cause Readmission and Mortality Among Patients with Heart Failure," *BioMedInformatics*, vol. 4, no. 4, pp. 2201–2212, Nov. 2024, https://doi.org/10.3390/biomedinformatics4040118.

[26] S. van Buuren, *Flexible Imputation of Missing Data*, 2nd ed. New York, NY, USA: Chapman and Hall/CRC, 2018, https://doi.org/10.1201/9780429492259.

[27] S. Jäger, A. Allhorn, and F. Bießmann, "A Benchmark for Data Imputation Methods," *Frontiers in Big Data*, vol. 4, July 2021, Art. no. 693674, https://doi.org/10.3389/fdata.2021.693674.

[28] T. Zonta, C. A. da Costa, R. da Rosa Righi, M. J. de Lima, E. S. da Trindade, and G. P. Li, "Predictive maintenance in the Industry 4.0: A systematic literature review," *Computers & Industrial Engineering*, vol. 150, Dec. 2020, Art. no. 106889, https://doi.org/10.1016/j.cie.2020.106889.

[29] W. Zhang, D. Yang, and H. Wang, "Data-Driven Methods for Predictive Maintenance of Industrial Equipment: A Survey," *IEEE Systems Journal*, vol. 13, no. 3, pp. 2213–2227, Sept. 2019, https://doi.org/10.1109/JSYST.2019.2905565.

[30] R. Little and D. Rubin, "Front Matter," in *Statistical Analysis with Missing Data*, 3rd ed., Hoboken, NJ, USA: John Wiley & Sons, Ltd, 2019, pp. i–xii, https://doi.org/10.1002/9781119482260.fmatter.

[31] P. C. Austin, I. R. White, D. S. Lee, and S. van Buuren, "Missing Data in Clinical Research: A Tutorial on Multiple Imputation," *Canadian Journal of Cardiology*, vol. 37, no. 9, pp. 1322–1331, Sept. 2021, https://doi.org/10.1016/j.cjca.2020.11.010.

[32] J. R. Carpenter and M. Smuk, "Missing data: A statistical framework for practice," *Biometrical Journal*, vol. 63, no. 5, pp. 915–947, June 2021, https://doi.org/10.1002/bimj.202000196.