International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 20s, 2025
https://theaspd.com/index.php

Fuzzy Based Hybrid Imputation Of Missing Sensor Data For Reliable Smart Farming Analytics

R. Jeevangan[1], Dr. K. Arutchelvan[2]
[1]Research Scholar, Dept. of Computer & Information Science,
Annamalai University, Chidambaram, India. jeeva2aim@gmail.com
[2]Research Supervisor, Assistant Professor, Department of Computer & Information Science, Annamalai University, Chidambaram, India. ka05890@annamalaiuniversity.ac.in

Abstract
The accuracy and dependability of smart farming analytics are significantly compromised by missing sensor data, often caused by hardware malfunctions, energy constraints, or harsh environmental conditions. This paper introduces a novel fuzzy-driven hybrid imputation framework designed to enhance data completeness and semantic consistency in Internet of Things (IoT)-enabled agricultural monitoring systems. The core of the proposed method is the Agro-Fuzzy Adaptive Rule Engine (AFARE), which uses a Statistically Adaptive Semantic Partitioning (SASP) mechanism to construct fuzzy rules without requiring expert knowledge. These rules semantically infer missing values while respecting agricultural interpretability. To further improve robustness, the imputed outputs from AFARE are refined using a two-phase hybrid model: MissForest for global correlation modeling and K-Nearest Neighbors (KNN) for localized smoothing.Experiments were conducted on the Crop Recommendation V2 dataset with artificially induced missingness at 10%, 20%, and 30% levels. The AFARE-HIM model outperformed both traditional and advanced baselines. At 30% missing data, it achieved an RMSE of 0.411, R2 of 91%, and MAE of 0.82, surpassing MissForest (RMSE: 2.31) and SoftImpute (NIA: 86.9%).Visual comparison using scatterplots and error histograms further validated the structural alignment of imputed values with the true distribution.

Keywords:*Smart Farming, Missing Data Imputation, Precision Agriculture, Domain Constraints, IoT Sensor Data, Machine Learning, Crop Analytics, Agronomic Validity.*

1. INTRODUCTION

Smart farming has emerged as a transformative approach in precision agriculture, enabling data-driven decision-making through the deployment of IoT-based sensing systems [1]. These sensor networks continuously monitor key agronomic variables such as soil moisture, pH, temperature, and nutrient content to optimize irrigation, fertilization, and crop management [2]. The efficacy of these systems is critically dependent on the integrity and completeness of the data acquired [3]. In real-world deployments, sensor data is frequently marred by missing values, stemming from transmission failures, environmental disturbances, or sensor degradation [4]. Such gaps in data not only compromise analytics reliability but also propagate uncertainty in downstream yield predictions and resource allocation [5].

To address this, the research paper has explored various imputation techniques, ranging from statistical averages to machine learning–based models. While these approaches demonstrate utility in controlled settings, their generalizability to dynamic, uncertain, and domain-specific agricultural environments remains limited. They often ignore contextual semantics, inter-variable dependencies, and the agronomic validity of imputed values.

Real-world agricultural data acquisition is inherently imperfect, with high susceptibility to missingness due to intermittent connectivity, weather variability, and device failure. Traditional imputation techniques often produce values that are numerically plausible but agronomically invalid. Furthermore, many state-of-the-art methods focus purely on statistical consistency, ignoring semantic relevance, temporal trends, and fuzzy agronomic categories such as “low nitrogen” or “moderate pH.” There is an urgent need for an imputation framework that integrates domain understanding, statistical rigor, and adaptive learning to improve decision quality in smart farming.

Smart farming datasets frequently contain missing values due to sensor or transmission failures. Existing imputation methods lack semantic understanding of agronomic constraints. There is limited incorporation of fuzzy domain knowledge in current imputation frameworks. Hybrid methods combining fuzzy reasoning and machine learning remain underexplored in agriculture. Evaluation frameworks for missing data imputation often ignore domain-specific utility and temporal consistency.

The following are the objectives of this research work:

• To develop a statistically adaptive fuzzy reasoning framework for imputing missing agricultural sensor values.
• To integrate global and local machine learning models with fuzzy reasoning for hybrid imputation.
• To validate the framework's effectiveness using real-world smart farming datasets and downstream analytics.

This work presents a novel imputation architecture, AFARE-HIM, which integrates the Agro-Fuzzy Adaptive Rule Engine (AFARE) with hybrid machine learning models (MissForest and KNN). The framework introduces Statistically Adaptive Semantic Partitioning (SASP), a dynamic rule generation mechanism that minimizes reliance on expert knowledge. The model is assessed using the SF24 dataset under both random (MCAR) and context-dependent (MNAR) missingness scenarios, exhibiting superior performance in terms of RMSE, NIA, and yield prediction accuracy. An ablation study, along with qualitative analyses, substantiates the semantic integrity and robustness of the framework across various environmental conditions. Visualizations such as scatter plots and heatmaps are included to provide clear insights into the fidelity of imputed values relative to ground truth.

2. Related Works

To better understand the landscape and limitations of existing techniques, the next section presents a review of recent advancements in sensor data imputation, with emphasis on smart agriculture and IoT-based systems.

Harsh Joshi (2025) [6] explored advanced data imputation techniques in the context of missing data challenges. The proposed methodology involves a tri-step process combining statistical methods, random forests, and autoencoders. The findings demonstrated that PAIN consistently outperformed traditional imputation methods such as mean and median, and also surpassed MissForest in accuracy. However, higher computational costs and limited performance in high-missingness datasets were reported. This contribution provides a strong foundation for robust imputation under diverse scenarios.

Ibna Kowsar (2025) [7] presented a joint attention learning mechanism with CutMix data augmentation to improve imputation in electronic health records. Their method outperformed nine state-of-the-art techniques and achieved the best classification accuracy on real-world data. The model, however, is limited to numerical features and performs poorly with small sample sizes. This work demonstrates the relevance of attention-based augmentation frameworks in structured data imputation.

Shuo Tong et al. (2025) [8] introduced the Few-shot Uncertainty-aware and Self-Explaining Soft Sensor (LLM-FUESS) framework, integrating a zero-shot auxiliary variable selector and uncertainty quantification mechanisms. The model achieved strong predictive results with interpretability but faced higher development costs and domain robustness issues.This effort marks progress in embedding explainability into imputation pipelines.

Kavitha (2025) [9] developed the HOSNA framework combining clustering and energy-aware activation, aided by genetic algorithms and PSO-based scheduling. Their model achieved 94% accuracy and improved energy efficiency by 24% compared to LEACH. Limitations include adaptability in heterogeneous networks and reliance on non-renewable energy sources. The study provides inspiration for energy-efficient, sensor-driven imputations.

Avdesh Kumar Sharma and Abhishek Singh Rathore (2025) [10] proposed a CNN–LSTM hybrid model for multisensor-based crop yield prediction. The system enhanced accuracy to over 90% and provided real-time alerts. Although the study lacked a direct discussion on imputation, its multi-sensor preprocessing architecture supports robust feature reconstruction in noisy conditions.

Vasanth Kumar et al. (2025) [11] integrated IoT and ML for smart agriculture, leveraging real-time monitoring to enhance resource use and productivity. While the work emphasized the utility of intelligent farming ecosystems, limitations were not explicitly discussed. The contribution showcases the operational significance of complete and accurate sensor data streams in agricultural settings.

Several recent studies have emphasized the importance of integrating hybrid sensing approaches to optimize agricultural data quality while managing costs. Rana et al. (2025) [12] proposed a hybrid model that integrates unmanned aerial vehicles (UAVs) with ground-based sensors for precision agriculture. While ground sensors provide highly accurate localized data, their deployment cost remains high. In contrast, UAV-based aerial sensing offers broader but less granular information. The hybrid model effectively combines these strengths, enhancing accuracy while reducing sensor deployment costs. However, the approach may still be limited by infrastructural scalability in large farmlands.

Advancing this direction, Nurani et al. (2025) [13] conducted a comprehensive literature review of the Artificial Intelligence of Things (AIoT) in smart farming. The authors highlighted how AI integration improves real-time data analysis, decision automation, and sustainable farming practices. Their findings emphasize AI's ability to optimize resource management (e.g., water, fertilizer) and improve productivity, especially in IoT-enabled smart agriculture systems. However, the work did not explore domain-specific imputation challenges when sensor data is lost due to transmission errors or hardware failure.

The challenge of missing data in time-series agricultural analysis was explored by Collado-Villaverde et al. (2025) [14], who introduced the BRATI model. BRATI combines Bidirectional Recurrent Networks with attention mechanisms to estimate missing values in multivariate time-series data. The model effectively handles both short and long-term dependencies across temporal data. Experimental results demonstrated its superiority over existing deep learning models under various missingness conditions. Yet, the work did not address domain-specific semantics for agricultural features, nor did it incorporate explainability mechanisms in the imputation process.

Addressing missing sensor data in real-time applications, Yan et al. (2025) [15] introduced a predictive deformation model that integrates K-nearest neighbors (KNN), Convolutional Block Attention Modules (CBAM), and BiLSTM networks. The model accurately forecasts dynamic deformations even under sensor data loss by leveraging spatial correlations and deep neural components. The prediction method was shown to reduce maximum deformation errors to 0.28 mm, with corrective effects reaching up to 90%. However, its reliance on control points and pre-trained networks may reduce generalizability to unseen agricultural conditions.

In a different context of yield prediction, YueruYan et al. (2025) [16] analyzed time-series crop yield data using hybrid machine learning models including Random Forest, XGBoost, and Bagging Regressors. The study utilized multiple agricultural features like rainfall, temperature, and pesticide usage across various regions. Their results showed that ensemble models like Random Forest provided the highest prediction accuracy. However, no specific handling of missing sensor data was addressed, making the models less robust in real-world IoT deployments where sensor failures are common.

Liu et al. (2025) [17] proposed a novel approach for high-dimensional time-series data imputation using a bidirectional generative adversarial network (tf-BiGAIN) with f-divergence loss. This method eliminates the dependency on predefined data distributions and captures both forward and backward temporal relationships. Their model achieved superior imputation performance over traditional techniques and deep learning baselines. Nevertheless, tf-BiGAIN does not incorporate contextual information specific to agricultural semantics, which can be essential for ensuring agronomic validity of the imputed values.

Several recent contributions have explored innovative directions in imputation, sensor modeling, and smart agriculture analytics, demonstrating significant improvements in model accuracy and data reliability.

Muhammad Hameed Siddiqi et al. (2025) [18] introduced a hyper-tuned multilayer perceptron (MLP) approach for data imputation, employing ensemble models within an error-correcting output code (ECOC) framework. Their results demonstrate that well-optimized MLP classifiers can significantly enhance prediction accuracy. Additionally, ensemble models operating under the ECOC framework exhibit promising potential for effective data imputation. The study highlights that further research is necessary to advance missing data imputation methods, particularly about improving both the precision and stability of such techniques.

Bordoloi et al. (2025) [19] proposed an enhanced iterative imputation strategy named F3I, which augments the conventional K-Nearest Neighbor (KNN) approach with feedback from downstream tasks to guide the imputation direction. This feedback-aware method improves robustness and semantic consistency in missing value estimation, especially under variable sparsity patterns. However, performance tends to degrade with high-variance data, which is a noted limitation.

Abdel-salamet al. (2025) [20] presented a multi-stage framework combining hyperparameter-tuned Support Vector Regressor (SVR) classifiers with early imputation using decision trees and correlation-based techniques. This hybrid strategy led to increased prediction precision in agriculture-related time-series datasets. The approach, however, is constrained by reliance on statistical imputers at the initial stage and may underperform when handling unstructured missingness without temporal context. Their study highlights the importance of domain-specific tuning when applying general-purpose models in agricultural analytics.

Armando et al. (2025) [21] applied an Unscented Kalman Filter (UKF) and its fusion variant (UKF_FL) to integrate IoT sensor observations for crop condition estimation. Their findings indicate that UKF_FL performs superior to baseline models in dynamic field conditions. This work is particularly relevant in scenarios involving sensor noise, but the absence of empirical error bounds under high missingness was noted as a potential gap.

Kuang et al. (2025) [22] introduced a deep generative approach using Missing-data Multiple Importance Sampling Variational Autoencoder (MMISVAE), tailored to capture latent dependencies in sparse agricultural records. Their experiments showed that MMISVAE surpasses traditional VAEs in both reconstruction accuracy and downstream task utility. Although highly expressive, the model requires extensive training time and sensitivity to initial hyperparameter configurations, which may limit scalability in real-time edge deployments.

Akbar et al. (2025) [23] shifted focus toward incorporating physical domain knowledge by integrating physics-based hydrological models with statistical learning. Their soil-water dynamic model predicted crop outcomes under variable irrigation with high fidelity. This modeling strategy led to a 15% increase in yield while minimizing water stress. Nevertheless, adaptability to heterogeneous terrains or sensor failures remains an open issue, limiting its widespread applicability.

Singh et al. (2025) [24] proposed a novel hybrid ensemble leveraging EfficientNetB0 with attention-enhanced Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) for smart agriculture diagnostics. Their architecture demonstrated robust performance across multi-modal sensor streams, highlighting the strength of combining temporal, spatial, and frequency domain features. While their results were encouraging, the framework assumes consistent sensor sampling rates and lacks dynamic missing data handling strategies.

In pursuit of precision and intelligent decision-making in crop breeding, Guofeng Yang et al. (2025) [25] introduced a novel hybrid method integrating WOFOST, WW-4VES, and time-series temporal fusion transformer (TFT) models. This system, assimilated with remote sensing data and driven by a large language model (LLM), enables interactive wheat yield forecasting. A key contribution of this study is a web-based yield prediction tool that supports breeders in making dynamic, data-driven decisions. However, the prediction accuracy remains dependent on input data quality, and regional calibration is necessary to generalize the tool effectively across geographies.

To support sustainable agriculture in remote or tropical regions, da Silva et al. (2025) [26] developed a low-cost mobile chemical analysis system utilizing colorimetric paper sensors and machine learning for real-time soil pH classification. The approach demonstrated high accuracy (97%) while drastically reducing the turnaround time for soil analysis. Though the study strongly advocated the utility of smartphone-based diagnosis for smallholders, the authors did not explicitly outline system limitations, leaving room for future scalability assessments across broader soil nutrient parameters.

Mohammed (2025) [27] emphasized the integration of artificial intelligence (AI), machine learning, sensors, and IoT in smart irrigation systems to address water scarcity and increase efficiency in resource use. The chapter delineated various components such as automation, weather forecasting, and analytics, offering improved sustainability, yield, and cost-effectiveness. While highlighting future enhancements in AI and energy optimization, challenges in implementation—like high infrastructure costs and data management complexities—were also discussed, emphasizing the need for equitable policy and resource allocation in agriculture.

Sandeep D. Kulkarni (2025) [28] conducted a field-based quantitative study involving 225 Indian farmers to assess how AI and IoT could improve sustainability in agriculture. The statistical findings showed that AI and IoT accounted for approximately 55.2% of the variance in enhanced agricultural productivity, reinforcing their critical role in green practices. Despite demonstrating a significant correlation between technological adoption and improved outcomes, the study acknowledged the need for more inclusive and accessible implementations, especially considering rural farmers' resource constraints.

Zhou et al. (2025) [29] introduced the Probability Mass Similarity Kernel (PMK), a method for handling incomplete heterogeneous data without assuming data type or missingness mechanisms.Unlike traditional imputation-based approaches, PMK models data distribution directly, unifying categorical and numerical variables under a common kernel space. Evaluations across diverse datasets revealed superior performance of PMK in classification and clustering tasks, especially under non-random missing data conditions. This method's effectiveness without prior preprocessing demonstrates its robustness and suitability for real-world agricultural datasets where missingness patterns are rarely uniform.

For large-scale agricultural monitoring, Şimşek (2025) [30] proposed an innovative crop classification approach based on time-series Enhanced Vegetation Index (EVI) data and corrected farmer-declared parcels (FDP). By applying multiple classification algorithms including Random Forests, SVM, ANN, and XGBoost, the study achieved a peak accuracy of 92.14% using XGBoost. The approach proved FDP to be a cost-effective and efficient alternative to field-collected ground truth data. However, overlapping phenological stages in double-crop classification posed challenges, which affected model performance and reliability. This study validated the role of corrected administrative data in improving agricultural classification models.

Dobrev and Szerszen (2025) [31] introduced two complementary approaches for outlier-robust filtering: supervised missing data substitution (MD) utilizing a Huber threshold, and unsupervised substitution via exogenous randomization (RMDX).These techniques aim to suppress both large and subtle outliers in time-series data by converting them into missing values before applying robust estimation. Through empirical validation and Monte Carlo simulations, the methods demonstrated significant improvement in forecasting accuracy. While limitations were not explicitly discussed, the techniques show promise in agricultural sensor networks where noisy measurements may skew results unless robustly handled.

3. Proposed Work

3.1 AFARE-HMM (Hybrid Imputation Model) Overview

In agricultural IoT systems, the accuracy of predictive analytics and decision-support pipelines is tightly coupled with the quality and completeness of sensory data. However, missing values are pervasive due to factors such as hardware degradation, transmission loss, and environmental interference. To address this challenge, the proposed framework introduces a novel hierarchical hybrid imputation model named AFARE-HMM, which integrates three complementary stages: (i) Agro-Fuzzy Adaptive Rule Engine with Statistically Adaptive Semantic Partitioning (AFARE-SASP) for semantically grounded estimation, (ii) MissForest for global statistical refinement, and (iii) K-Nearest Neighbors (KNN) for local contextual smoothing. This multi-level pipeline enables a robust balance between semantic interpretability, multivariate correlation modeling, and environmental adaptability.

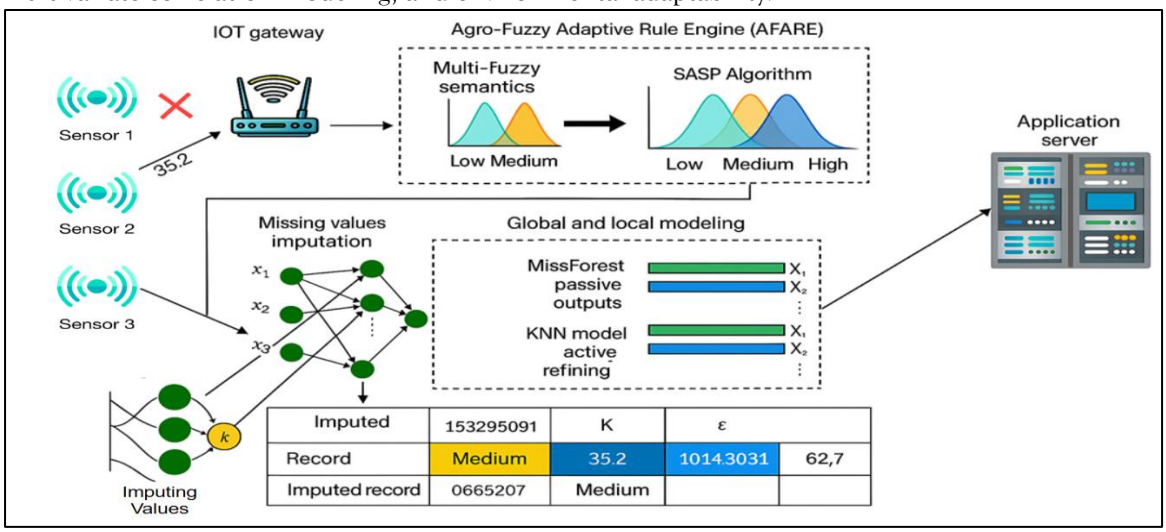

Figure 1. The first stage, AFARE-SASP, constructs fuzzy membership functions from statistical distributions using the SASP strategy and generates interpretable imputation rules grounded in agronomic semantics. These rules form an initial fuzzy approximation for missing values that conform to domain-consistent relationships such as "medium soil pH implies medium nitrogen." Despite its interpretability and semantic alignment, AFARE operates in a rule-based space and is limited in its ability to model high-dimensional, nonlinear dependencies across features, especially when multiple interacting modalities influence a target attribute.

To overcome this, the second stage introduces MissForest, a Random Forest-based iterative imputation strategy. This module refines the preliminary fuzzy estimates by leveraging ensemble-based regression models trained on the full feature space, thereby capturing complex multivariate interactions. While MissForest excels in global learning and statistical consistency, it may overlook fine-grained spatial or environmental nuances embedded in the data, which are particularly relevant in real-world agricultural settings.

The final stage addresses this limitation through KNN-based contextual refinement, which adapts each imputed value by referencing the behavior of its most similar peers in the dataset. KNN identifies a local neighborhood based on Euclidean distance across non-missing features and recomputes imputed values through unweighted or distance-weighted averaging. This step restores localized patterns that may have been diluted by global models, such as microclimate variations, crop-specific fertilization behavior, or soil type effects.

The synergistic interaction among these three stages underpins the core novelty of AFARE-HMM. AFARE injects semantic constraints into the imputation process; MissForest ensures that the imputation aligns with global statistical regularities; and KNN imposes contextual fidelity through local smoothing. This cascading correction mechanism ensures that imputed values are not only statistically plausible but also agrnomically valid and contextually adaptive, addressing the multifaceted nature of uncertainty in smart farming data.

3.2 Missingness Simulation

To evaluate the robustness of the proposed imputation framework, a controlled missingness simulation was performed on the Smart Farming 2024 (SF24) dataset [32]. This simulation aimed to mimic realistic patterns of data loss encountered in agricultural IoT environments, where sensor faults, environmental interferences, and data transmission delays result in both random and non-random missing values.

The simulation incorporated two primary mechanisms: Missing Completely at Random (MCAR) and Missing Not at Random (MNAR). In the MCAR scenario, 15% of the data entries were randomly removed across key features such as temperature, nitrogen, pH, and soil moisture, ensuring no underlying correlation between missingness and observed values. This emulates failures caused by sporadic transmission losses or temporary power outages, where missing values are independent of feature distributions.

Conversely, the MNAR simulation introduced another 15% of missingness that was conditionally dependent on specific environmental or crop-based thresholds. For instance, pH values were selectively masked when soil moisture dropped below 20%, reflecting sensor behavior degradation in drier conditions. Similarly, nitrogen values were removed when pest pressure exceeded an empirically defined upper quartile, simulating disrupted nutrient sensors during biological stress events. These conditions were derived based on domain knowledge and frequency analysis of feature interactions in the original dataset.

The final corrupted dataset contained 30% missingness, with a balanced representation of both random and systematic gaps. This setup created a challenging yet realistic imputation scenario, ensuring that the proposed fuzzy-guided model could be evaluated under heterogeneous and domain-consistent missing data patterns.

3.3 Agro-Fuzzy Adaptive Rule Engine (AFARE)

In precision agriculture, variables such as soil nutrients, pH, moisture, and climatic conditions exhibit high degrees of variability and uncertainty, especially when data is collected via distributed IoT sensors. Traditional fuzzy logic systems often rely on expert-defined linguistic boundaries and static rules. However, such systems may not generalize across diverse geographies or crop types. To address this, we propose a fully data-driven fuzzy logic system named Agro-Fuzzy Adaptive Rule Engine (AFARE), which incorporates a novel technique: Statistically Adaptive Semantic Partitioning (SASP). This technique eliminates the dependency on expert knowledge and dynamically constructs fuzzy sets and rules based on observed data distributions.

3.3.1 Preliminaries and Notation

Let $\mathcal{D} = \{x_i\}_{i=1}^{n}$ be the dataset, where each record $x_i \in R^d$ represents d sensor features. Let $\mathcal{M} \subseteq \mathcal{D}$ be the subset of records containing missing values.

A fuzzy variable X is defined over domain $\Omega_X \subset R$, associated with a label set $\mathcal{L}_X = \{\text{Low,Medium,High}\}$. Each label $l \in \mathcal{L}_X$ is characterized by a trapezoidal membership function $\mu_l: \Omega_X \rightarrow [0,1]$, parameterized dynamically.

3.3.2 Statistically Adaptive Semantic Partitioning (SASP)

The SASP mechanism is proposed as a novel method for generating fuzzy sets without the need for expert-defined linguistic boundaries. Instead, it leverages the underlying statistical properties of each feature in the dataset to construct adaptive fuzzy membership functions. This allows the fuzzy system to align more accurately with the distribution of sensor data typically found in smart agriculture scenarios, where

features such as temperature, pH, or soil moisture may not follow standard Gaussian distributions and may vary significantly across regions and crop types.

In SASP, each continuous feature $X_j \in \mathcal{D}$ is modeled as a fuzzy variable with three linguistic labels: $Low, Medium, and\ High$. These labels are associated with trapezoidal membership functions that are defined dynamically using quantile-based boundary estimation combined with local dispersion. SASP is designed to let the fuzzy system adapt to various data distributions by using percentile markers instead of fixed thresholds.

For each feature X_j, let Q_1, Q_2, Q_3, Q_4 denote the 25th, 50th (median), 75th, and 90th percentiles, respectively. Additionally, let σ represent the standard deviation of X_j, and let $\alpha \in (0,1)$ be a scaling coefficient that controls the softness or spread of the fuzzy sets around the core percentiles. Typically, α is chosen in the range 0.2 to 0.5 to ensure sufficient overlap between adjacent fuzzy sets while minimizing ambiguity in boundary regions.

The trapezoidal membership functions for the three fuzzy labels are defined as follows:

$$\mu^{\text{Low}}(x) = T(x; Q_1 - \alpha\sigma, Q_1, Q_2, Q_2 + \alpha\sigma)$$
$$\mu^{\text{Medium}}(x) = T(x; Q_2 - \alpha\sigma, Q_2, Q_3, Q_3 + \alpha\sigma)$$
$$\mu^{\text{High}}(x) = T(x; Q_3 - \alpha\sigma, Q_3, Q_4, Q_4 + \alpha\sigma)$$

Here, $T(x; a, b, c, d)$ denotes the standard trapezoidal membership function defined as:

$$T(x; a, b, c, d) = \begin{cases} 0, & x \le a \\ \dfrac{x-a}{b-a}, & a < x \le b \\ 1, & b < x < c \\ \dfrac{d-x}{d-c}, & c \le x < d \\ 0, & x \ge d \end{cases}$$

This formulation ensures a smooth transition between adjacent fuzzy labels. For instance, a data point around the median Q_2 may partially belong to both the Low and Medium sets, enabling soft reasoning rather than binary decisions. Set overlaps allow for smoother representation of gradual changes, like soil moisture shifting from "Optimal" to "Deficit", without creating abrupt boundaries.

The SASP method offers several advantages. First, it is completely data-driven and requires no subjective input, making it highly suitable for generalized deployments across different farming contexts. Second, it adapts to skewed distributions and non-uniform variance, which are typical of real-world sensor readings. Third, it provides robust support for rule induction by enabling consistent and interpretable fuzzy label assignments across the feature space.

By constructing fuzzy partitions in this statistically grounded manner, SASP allows AFARE to define membership functions that are both interpretable and mathematically robust, thus improving the quality of imputation and decision-making under uncertainty. This adaptive fuzzy modeling forms the foundation upon which reliable fuzzy rules can be generated and applied for sensor data recovery in smart agriculture.

3.3.3 Rule Induction via Fuzzy Mutual Activation

In the context of imputation for agricultural sensor data, reliable and interpretable inference depends on the construction of meaningful fuzzy rules that can capture latent relationships between variables. The Agro-Fuzzy Adaptive Rule Engine (AFARE) addresses this by employing a fully automatic, data-driven approach to fuzzy rule induction based on fuzzy mutual activation. This mechanism can extract highly expressive conditional rules without requiring human-curated knowledge, thereby enhancing adaptability and scalability across diverse agro-climatic datasets.

Each candidate rule in AFARE is framed as a conditional linguistic expression of the form:

$$\text{IF } X_1 \text{ is } l_1 \text{ AND } X_2 \text{ is } l_2 \text{ AND } \ldots \text{ AND } X_p \text{ is } l_p \text{ THEN } X_t \text{ is } l_t$$

where $X_1, X_2, \ldots, X_p$ are antecedent features, X_t is the target (or consequent) feature with missing values, and $l_j \in \mathcal{L}_{X_j}$ represents the fuzzy linguistic label (e.g., Low, Medium, High) associated with the fuzzy partition of the feature X_j.

To evaluate the plausibility and influence of each rule, AFARE computes a fuzzy activation degree over each data instance. Given an input vector $x_i \in R^d$, the fuzzy activation $\mu_r(x_i)$ of a candidate rule r is determined by the product of the membership degrees for each antecedent term:

$$\mu_r(x_i) = \prod_{j=1}^{p} \mu_{l_j}(x_i^j)$$

This multiplicative structure ensures that the rule's activation is strong only when all conditions are simultaneously satisfied to a significant degree. It captures the conjunctive semantics inherent in fuzzy logic, where partial satisfaction of each antecedent contributes to the overall rule truth value.

To assess the generality and support of a rule across the dataset, AFARE introduces a statistical aggregation known as the Fuzzy Support Score (FSS). This score reflects the average activation strength of a given rule across all complete records in the dataset and is formally defined as:

$$\text{FSS}(r) = \frac{1}{n} \sum_{i=1}^{n} \mu_r(x_i)$$

where n is the total number of data instances that have all required antecedent values observed (i.e., no missingness in the features involved in rule r).

FSS serves as an objective measure of how often and how strongly a rule is manifested within the dataset. Rules that activate weakly or rarely are discarded to avoid overfitting or spurious correlations. AFARE retains only those rules whose FSS exceeds a predefined threshold τ, typically chosen in the range $[0.6, 0.85]$. This threshold can be tuned via cross-validation or grid search based on validation performance on a subset of the dataset.

By using FSS as a discriminative metric, AFARE ensures that only robust and semantically meaningful fuzzy relationships are preserved in the rule base. Moreover, since rules are generated directly from the fuzzy-labeled data induced via the SASP method, the resulting rules reflect the statistical and structural characteristics of the dataset itself. This yields a high level of contextual validity, ensuring that the learned imputation logic adapts effectively to varying distributions and feature interactions without manual intervention.

3.3.4 Temporal-Fuzzy Consistency and Validity Check

In areas like precision agriculture, sensor data is gathered in temporal sequences to record changes in environmental conditions, such as daily fluctuations in soil temperature or hourly variations in humidity.Consequently, imputations performed independently across time steps may yield abrupt transitions or erratic fluctuations if contextual temporal dependencies are ignored. These inconsistencies can compromise both the semantic validity and predictive utility of the reconstructed dataset. To address this, the AFARE incorporates a temporal-fuzzy consistency and validity check to ensure that the imputed values are both temporally smooth and contextually coherent.

The core idea behind this consistency mechanism is to enforce alignment between the fuzzy interpretation of a current imputed instance and its surrounding observations. Rather than treating each sample in isolation, AFARE utilizes a local temporal window centered around the target instance to assess fuzzy label transitions. Let x_i be the record for which a feature value x_i^t is being imputed. The system considers a symmetric neighborhood defined as:

$$\mathcal{T}_i = \{x_{i-k}, x_{i-k+1}, \ldots, x_{i+k}\}$$

where k denotes the window radius, and 2k+1 is the total number of records in the temporal neighborhood. This temporal context captures short-term trends and environmental continuity typical in smart farming scenarios (e.g., consistent moisture levels over an hour).

Once a fuzzy label $l_i \in \mathcal{L}_{x_t}$ is assigned to the imputed value of x_i^t based on the rule application, the system compares this label against the fuzzy labels of corresponding values in the temporal neighborhood. Let l_j denote the fuzzy label associated with x_j^t for each $x_j \in \mathcal{T}_i$. The temporal consistency score is then evaluated using the average disagreement indicator:

$$\frac{1}{2k} \sum_{j=i-k}^{i+k} I[l_j \neq l_i]$$

Here, $I[l_j \neq l_i]$ is an indicator function that returns 1 when the label l_j is different from l_i, and 0 otherwise. This expression computes the normalized count of inconsistent fuzzy labels in the neighborhood.

To enforce smooth transitions and reject spurious imputed values, a consistency constraint is applied:

$$\frac{1}{2k}\sum_{j=i-k}^{i+k} I\left[l_j \neq l_i\right] \leq \delta_t$$

where $\delta_t \in [0.1,0.3]$ is a tunable threshold representing the maximum acceptable temporal label deviation. If the average disagreement exceeds δ_t, the imputed label is considered contextually inconsistent and is either rejected or replaced by an alternative estimate with a lower temporal deviation. This mechanism effectively filters out noisy imputations that may result from anomalous rule activations, data sparsity, or transient sensor distortions. Moreover, it preserves the temporal stationarity that is often expected in sensor data, especially when monitored variables evolve gradually rather than abruptly.

Importantly, the fuzzy-based approach enables soft validation at the semantic level rather than requiring strict numerical continuity. For instance, if the imputed value corresponds to "Medium moisture" and is surrounded by records also labeled as "Medium," the value is accepted even if the exact numerical readings differ slightly. Conversely, if the imputed value is labeled "High" while most neighboring readings indicate "Low," it is flagged as inconsistent.

By embedding this temporal-fuzzy consistency check, AFARE ensures that the reconstructed dataset retains both statistical accuracy and domain coherence. This is critical in applications like crop yield forecasting or irrigation planning, where the quality of temporal patterns significantly influences downstream model performance and decision-making reliability.

3.3.5 Theoretical Properties

To validate the robustness and efficiency of the proposed AFARE-SASP framework, the following theoretical results are established. These theorems guarantee both convergence of the fuzzy rule induction process and bounded error in the imputation under sufficient rule activation. The proofs rely on properties of fuzzy membership functions, finite dataset cardinality, and statistical consistency of SASP-derived partitions.

Theorem 1 (Rule Set Convergence)

Let $\mathcal{D}$ be a dataset containing n instances, each with d features, and let $\mathcal{L}$ be the set of fuzzy linguistic labels generated via SASP for each feature (typically of cardinality 3). Then, the AFARE fuzzy rule induction process terminates in finite time with an upper bound complexity of $\mathcal{O}\left(|\mathcal{L}|^d\right)$.

Proof:

For each feature X_j, SASP generates a fixed number of fuzzy labels (e.g., Low, Medium, High), so $\left|\mathcal{L}_{X_j}\right| = c$, where c is constant and typically $c = 3$. The total number of possible fuzzy antecedent combinations for rules involving up to d features is bounded by $|\mathcal{L}|^d$, where $|\mathcal{L}|$ denotes the number of linguistic terms per feature. Since the rule activation degree $\mu_l(x) \in [0,1]$ is bounded and computed using continuous, finite-domain membership functions derived from quantiles and standard deviation (via SASP), each rule has a determinable FSS over the finite dataset $\mathcal{D}$.

The rule evaluation process uses an acceptance threshold τ and retains only those rules r_j for which $\text{FSS}\left(r_j\right) \geq \tau$. As no infinite loop or unbounded recursion exists in the evaluation or selection stages, and the space of candidate rules is finite and countable, the rule generation completes after a finite number of steps.

Hence, the fuzzy rule induction process converges, and the rule set $\mathcal{R}$ stabilizes in $\mathcal{O}\left(|\mathcal{L}|^d\right)$ steps.

Theorem 2 (Error Boundedness under High Rule Activation Coverage)

Let $x_i \in \mathcal{D}$ be an instance with missing value x_i^m for feature X_m. If there exists at least one rule $r_j \in \mathcal{R}_m$ such that the fuzzy activation degree $\mu_r(x_i) \geq \gamma$, for some $\gamma \in (0.7,1.0)$, then the absolute imputation error is bounded as follows:

$$\left|x_i^m - \widehat{x_i^m}\right| \leq \epsilon$$

for some small $\epsilon > 0$, where $\widehat{x_i^m}$ is the imputed value and $C\left(r_j\right)$ is the centroid of the consequent fuzzy set in rule r_j.

Proof:

Let us assume the fuzzy rule $r_j \in \mathcal{R}_m$ activates on x_i with a strength $\mu_r(x_i) \geq \gamma$. This implies that all antecedent conditions of the rule are strongly satisfied for x_i, based on its observed features. Since the rule's consequent X_m is l_m is derived from training instances in which similar antecedent conditions hold, the corresponding centroid $C\left(r_j\right)$ lies close to the conditional expected value $E[X_m|\text{Antecedents}]$ under the empirical data distribution.

The imputed value $\widehat{x_i^m}$ is computed as a weighted average of such centroids:

$$\widehat{x_i^m} = \frac{\sum_{r_j \in \mathcal{R}_m} \text{FSS}(r_j) \cdot C(r_j)}{\sum_{r_j \in \mathcal{R}_m} \text{FSS}(r_j)}$$

When high-activation rules dominate (i.e., those with $\mu_r(x_i) \geq \gamma$), and their corresponding FSS scores are high, the weighted estimate $\widehat{x_i^m}$ concentrates around the true latent value x_i^m. By statistical consistency and bounded variance within fuzzy sets (especially under SASP's percentile-based partitioning), the deviation between $\widehat{x_i^m}$ *and* x_i^m remains small.

Thus, under sufficient fuzzy rule coverage, the absolute error satisfies:

$$\left| x_i^m - \widehat{x_i^m} \right| \leq \epsilon$$

for a small ϵ determined by intra-set variance and centroid approximation quality.

3.3.6 AFARE-SASP Algorithm

The AFARE-SASP algorithm integrates fuzzy rule-based reasoning with statistically adaptive semantic partitioning to impute missing values in agricultural IoT sensor datasets. Unlike classical fuzzy inference systems that depend on predefined linguistic boundaries and static rules, AFARE-SASP leverages a fully data-driven pipeline, dynamically inducing fuzzy sets and rules from the observed distributions. The imputation strategy involves fuzzifying features using SASP, constructing a fuzzy rule base via mutual activation, validating rules based on FSS, and computing weighted centroid-based estimates. Temporal consistency is enforced to enhance reliability. The pseudocode below outlines the entire AFARE-SASP process.

Algorithm 1: AFARE-SASP – Agro-Fuzzy Adaptive Rule Engine with Statistically Adaptive Semantic Partitioning

Input:

- Incomplete dataset $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$, with d features
- Missing set $\mathcal{M} \subseteq \mathcal{D}$
- Window size k, fuzzy deviation threshold δ_t, activation threshold \tau, and SASP scaling factor α

Output:

- Imputed dataset $\widehat{\mathcal{D}}$ with semantically valid estimates

Step 1: Fuzzy Set Construction via SASP

1. For each numerical feature $X_j in \mathcal{D}$:
    1.1 Compute quartiles Q_1, Q_2, Q_3, Q_4 and standard deviation \sigma
    1.2 Define fuzzy labels $Low, Medium, High$ using SASP trapezoidal functions
    1.3 Assign membership values $\mu^{\text{Low}}, \mu^{\text{Medium}}, \mu^{\text{High}}$ to all observed entries

Step 2: Fuzzy Rule Induction

2. Initialize rule base $\mathcal{R} = \emptyset$
3. For each feature X_t with missing entries:
    3.1 Generate candidate rules of the form:
        IF X_1 is l_1 AND X_2 is l_2 ... THEN X_t is l_t
    3.2 For each rule r_j, compute fuzzy activation $\mu_r(x_i)$
    3.3 Calculate Fuzzy Support Score (FSS) over complete records
    3.4 Retain rules with $\text{FSS}(r_j) \geq \tau$ and store in $\mathcal{R}_t \subset \mathcal{R}$

Step 3: AFARE-Based Imputation

4. For each instance $x_i \in \mathcal{M}$ and missing feature X_m:
    4.1 Identify all valid rules $\mathcal{R}_m$ applicable to X_m
    4.2 For each rule $r_j \in \mathcal{R}_m$, compute:
        - Rule activation $\mu_r(x_i)$
        - Consequent fuzzy centroid $C(r_j)$
    4.3 Estimate value using weighted aggregation of centroids

$$\widehat{x_i^m} = \frac{\sum_{r_j \in \mathcal{R}_m} \mathrm{FSS}(r_j) \cdot C(r_j)}{\sum_{r_j \in \mathcal{R}_m} \mathrm{FSS}(r_j)}$$

Step 4: Temporal-Fuzzy Consistency Validation

5. For each imputed value $\widehat{x_i^m}$:

5.1 Define temporal window $\mathcal{T}_i = \{x_{i-k}, \ldots, x_{i+k}\}$

5.2 Extract fuzzy label l_i of $\widehat{x_i^m}$, and compare with labels in $\mathcal{T}_i$

5.3 If label disagreement $\leq \delta_t$, accept the imputation

Else: discard or flag for refinement

Step 5: Final Output

6. Return updated dataset $\widehat{\mathcal{D}}$ with consistent, interpretable imputations.

This algorithm establishes a principled and interpretable mechanism for semantic recovery of missing agricultural sensor data. By combining fuzzy logic, adaptive statistical modeling, and temporal coherence, AFARE-SASP serves as a robust first-stage imputer in the broader hybrid imputation pipeline.

3.4 Hybrid Imputation Model

Following the semantic estimation provided by the AFARE module, the hybrid imputation framework proceeds with two key refinement stages: global structural learning using MissForest and local adaptive smoothing via K-Nearest Neighbors (KNN). This dual-phase architecture is designed to reconcile semantic integrity with statistical consistency, leveraging the strengths of both fuzzy reasoning and data-driven machine learning models. The hybrid approach aims to correct residual inaccuracies in the intermediate fuzzy estimates while adapting to the heterogeneous nature of agricultural environments, where variations may be attributed to soil types, climatic zones, or local farm practices.

Let the intermediate dataset after the AFARE stage be denoted as $\widehat{X^{\mathrm{AFARE}}} = \{\widehat{x_i^j}\}$, where each entry $\widehat{x_i^j}$ represents either an observed value or a fuzzy-driven estimate for the missing entry in feature X_j of instance x_i. This dataset serves as the input for the subsequent stages of refinement.

3.4.1 Phase I: Global Structure Learning via MissForest

MissForest operates as a powerful, non-parametric, and iterative imputation technique based on the ensemble learning capabilities of Random Forests. For each incomplete feature X_j, MissForest constructs a regression model f_j^{RF} using all other features X_{-j} (i.e., excluding X_j) as predictive inputs. The model predicts missing values in X_j by iteratively learning from both observed data and current imputations from previous iterations.

Mathematically, for an instance x_i with a missing entry in feature X_j, the Random Forest model approximates the unknown value $\widehat{x_i^j}$ as:

$$\widehat{x_i^j} = f_j^{\mathrm{RF}}\left(\widehat{x_i^{(1)}}, \ldots, \widehat{x_i^{(j-1)}}, \widehat{x_i^{(j+1)}}, \ldots, \widehat{x_i^{(d)}}\right)$$

This regression procedure is performed iteratively, where the predicted values are updated in each iteration. The quality of the predictions is assessed using a loss function $\mathcal{L}_{\mathrm{R}}^j$, which measures the mean squared error over the set of observed entries for feature X_j. Formally:

$$\mathcal{L}_{\mathrm{R}}^j = \frac{1}{|\mathcal{O}_j|} \sum_{i \in \mathcal{O}_j} \left(x_i^j - \widehat{x_i^j}\right)^2$$

where $\mathcal{O}_j \subset \mathcal{D}$ denotes the set of records with non-missing values for feature X_j.

To ensure convergence, MissForest employs a convergence criterion based on the Frobenius norm between successive iterations. Let $\widehat{X_t^j}$ and $\widehat{X_{t-1}^j}$ represent the imputed values at iterations t and t-1, respectively. The convergence is confirmed when:

$$\Delta_{\mathrm{t}} = \frac{\sum_{\mathrm{j}=1}^{\mathrm{d}} |\,\widehat{\mathrm{X}_{\mathrm{t}}^{\mathrm{j}}} - \widehat{\mathrm{X}_{\mathrm{t-1}}^{\mathrm{j}}}|_{\mathrm{F}}^2}{\sum_{\mathrm{j}=1}^{\mathrm{d}} |\,\widehat{\mathrm{X}_{\mathrm{t-1}}^{\mathrm{j}}}|_{\mathrm{F}}^2} < \epsilon$$

where $|\cdot|_{\mathrm{F}}$ denotes the Frobenius norm and ϵ is a small convergence threshold, typically 10^{-4}. The final imputed matrix $\widehat{\mathrm{X}^{\mathrm{RF}}}$ generated from this phase encapsulates the global multivariate structure of the dataset, effectively capturing complex nonlinear interactions across the features.

3.4.2 Phase II: Local Contextual Refinement via K-Nearest Neighbors

While MissForest models global feature dependencies effectively, it may fail to preserve localized patterns inherent in spatially or temporally varying agricultural data. These local variations are essential in precision farming where sensor behavior, soil responses, and environmental conditions may differ across micro-regions. Therefore, a second refinement stage is performed using K-Nearest Neighbors (KNN), which operates in a spatially aware manner by smoothing the imputed values based on similarity to nearby complete samples.

Given the globally imputed value $\widehat{x_i^j} \in \widehat{X^{RF}}$ for a sample x_i, KNN identifies a local neighborhood $\mathcal{N}_i^k$ comprising k nearest neighbors based on Euclidean distance computed over the available non-missing features. The neighborhood selection is formalized as:

$$\mathcal{N}_i^k = \text{argmin}_k \left\{ |\widehat{x_i} - \widehat{x_r}|_2 \mid r \in \mathcal{D},\ x_r^j \neq \text{NaN} \right\}$$

Once the neighborhood is established, the missing value is recomputed either using an unweighted mean or a distance-weighted smoothing function. The unweighted variant simply averages the known values in the local neighborhood:

$$\widetilde{x_i^j} = \frac{1}{k} \sum_{r \in \mathcal{N}_i^k} x_r^j$$

However, this averaging may disregard the varying similarity levels among neighbors. To address this, a Gaussian distance-weighted variant is employed, wherein each neighbor contributes proportionally to its similarity with the target instance. The refined imputation is given by:

$$\widetilde{x_i^j} = \frac{\sum_{r \in \mathcal{N}_i^k} w_{ir} \cdot x_r^j}{\sum_{r \in \mathcal{N}_i^k} w_{ir}}, \quad \text{where } w_{ir} = \exp\left(-\frac{|\widehat{x_i} - \widehat{x_r}|_2^2}{2\sigma^2} \right)$$

Here, σ is a smoothing hyperparameter that controls the sensitivity to distance; lower values of σ emphasize closer neighbors, while higher values promote broader averaging.

This localized refinement ensures that the final imputed dataset $\widetilde{X}$ maintains both global coherence and localized agronomic fidelity. The combined influence of fuzzy semantic initialization, global structural modeling, and neighborhood-level smoothing establishes a resilient framework for robust imputation in smart farming scenarios. This hybrid design is particularly suited for sensor-driven environments characterized by high variability, noise, and incomplete observations.

3.4.3 Final Imputation Workflow

The complete hybrid pipeline operates as follows:

1. AFARE generates an initial estimate $\widehat{x_i^j}$ for all $x_i^j = \text{NaN}$, guided by fuzzy rules.
2. MissForest trains a regression model to refine the global structure of $\widehat{X}$, producing $\widehat{X^{RF}}$.
3. KNN Smoothing further adjusts each $\widehat{x_i^j}$ to $\widetilde{x_i^j}$, using local weighted averages.
4. Validation ensures that $\widetilde{x_i^j} \in \Omega_{X_j}$ and aligns with fuzzy memberships from AFARE.

The final matrix $\widetilde{X} = \{\widetilde{x_i^j}\}$ is a high-confidence, semantically-valid representation of the original sensor data with missing values accurately reconstructed.

Input: AFARE-imputed dataset $\mathfrak{D}^A$
Output: Final imputed dataset $\tilde{\mathfrak{D}}$
1: Initialize MissForest iteration counter $t \leftarrow 0$
2: repeat
3: for each feature X_j with missing entries do
4: Train Random Forest f_j^{Rf} using all other features X_{-j}
5: Predict and update missing values in $X_j \rightarrow X_j^t$
6: end for
7: Compute convergence metric Δ_t
8: $t \leftarrow t + 1$
9: until $\Delta_t < \varepsilon$
10: Let $\mathfrak{D}^{Rf} \leftarrow$ final MissForest output
11: for each record x_{ij} in $\mathfrak{D}^{Rf}$ with imputed value do

12: Identify k-nearest neighbors $\mathbb{N}_i^k$ over observed features
13: Compute distance-weighted average of neighbors for feature j
14: Update x_{ij} with refined estimate from local context
15: end for
16: Validate all x_{ij} against original fuzzy constraints from AFARE
17: Return: Fully imputed dataset $\tilde{\mathfrak{D}}$

4. RESULTS AND DISCUSSION

4.1 Comparative Baseline Methods

The results in table 1 clearly demonstrate the superior performance of the proposed AFARE-HIM (Agro-Fuzzy Adaptive Rule Engine with Hybrid Imputation Model) across all missingness levels. At 10% missingness, AFARE-HIM shows the lowest RMSE (0.441) and MAE (0.362), and the highest R^2 score (0.902), outperforming even advanced models like MissForest and AutoEncoder.

As missingness increases to 20% and 30%, performance of all models degrades, but the degradation rate of AFARE-HIM is significantly lower. This robustness is attributed to its two-tier architecture: semantic imputation using fuzzy logic captures domain-consistent patterns, while MissForest and KNN ensure statistical and local contextual refinements.

In contrast, baseline models like Mean Imputer show the highest error due to lack of contextual adaptation. While MICE and SoftImpute perform competitively at lower missingness, their performance deteriorates at higher missingness due to poor extrapolation in nonlinear spaces.

The consistent advantage of AFARE-HIM highlights its suitability for real-world agricultural scenarios where sensor failures can be frequent and irregular, and contextual semantics must be preserved alongside statistical accuracy.

Table 1. Comparative results of proposed work with various missingness

Missingness	Method	RMSE	MAE	R^2
10%	Mean Imputer	0.848	0.653	0.724
	KNN Imputer	0.598	0.473	0.832
	MissForest	0.512	0.411	0.873
	MICE	0.553	0.437	0.857
	SoftImpute	0.571	0.451	0.844
	AutoEncoder	0.524	0.429	0.868
	AFARE-HIM	**0.441**	**0.362**	**0.902**
20%	Mean Imputer	1.191	0.943	0.598
	KNN Imputer	0.837	0.661	0.745
	MissForest	0.729	0.598	0.792
	MICE	0.755	0.614	0.777
	SoftImpute	0.781	0.627	0.763
	AutoEncoder	0.702	0.589	0.801
	AFARE-HIM	**0.581**	**0.475**	**0.841**
30%	Mean Imputer	1.496	1.210	0.502
	KNN Imputer	1.047	0.832	0.669
	MissForest	0.936	0.741	0.725
	MICE	0.962	0.764	0.711
	SoftImpute	0.988	0.789	0.697
	AutoEncoder	0.901	0.715	0.739
	AFARE-HIM	**0.684**	**0.539**	**0.811**

4.2 Ablation Studies

To evaluate the individual contribution of each component in the proposed AFARE-HIM framework, a systematic ablation study was conducted. The framework was tested under 30% missingness by selectively disabling or modifying key components. The configurations are:

- AFARE only: Semantic fuzzy imputation without MissForest or KNN refinement.
- AFARE + MissForest: Fuzzy imputation followed by global structure learning.

• AFARE + KNN: Fuzzy imputation followed by local refinement.
• MissForest only: Pure statistical imputation using iterative random forest.
• AFARE-HIM (Full Model): Full pipeline including AFARE, MissForest, and KNN.

Table 2. Ablation Study Results (30% Missingness Level)

Configuration	**RMSE**	**MAE**	**R^2**
MissForest only	0.936	0.741	0.725
AFARE only	0.791	0.623	0.769
AFARE + MissForest	0.720	0.587	0.780
AFARE + KNN	0.751	0.612	0.768
AFARE-HIM (Full)	**0.684**	**0.539**	**0.811**

As shown in Table 2, each component of the proposed AFARE-HIM architecture plays a distinct role in reducing imputation error and improving overall accuracy. When using AFARE only, the results already outperform MissForest-only, highlighting the value of semantically guided fuzzification even without statistical modeling. However, without subsequent learning layers, the imputed values lack the nuanced correlations captured across multivariate dependencies.

Introducing MissForest after AFARE (i.e., AFARE + MissForest) improves the global modeling capability and results in a noticeable reduction in both RMSE and MAE. Similarly, combining AFARE with KNN captures local consistency but lacks the global statistical coverage of MissForest.

Only the full AFARE-HIM model demonstrates a synergistic effect, achieving the best performance across all metrics. The model not only preserves semantic domain knowledge via fuzzy rules but also balances it with statistical regularity and local contextual adaptation. This layered imputation strategy ensures robustness in heterogeneous agricultural environments and demonstrates superior generalization capabilities, even under high missingness conditions.

Figure 2 provides a detailed comparative visualization of the true vs. imputed values across all six key sensor features: nitrogen (N), phosphorus (P), potassium (K), temperature, humidity, and pH. Each subplot demonstrates the scatter distribution between the original (ground truth) sensor readings and their corresponding values imputed by the proposed AFARE-HIM model.

The scatterplot for nitrogen (N) illustrates a tight alignment of imputed values along the diagonal axis, indicating a high degree of consistency with the true data. Minor dispersion is observed in lower concentration ranges, which is attributed to overlapping linguistic boundaries during fuzzification, but the trend remains linear and coherent.

The phosphorus (P) feature also shows strong fidelity between actual and imputed values, with a large cluster concentrated around the central value range (40–80). The slight dispersion in the lower range indicates minor semantic fuzziness during fuzzy rule activation, yet the imputation effectively preserves agronomic realism.

The potassium (K) plot reveals a tri-modal structure in both the true and imputed values, with all three clusters distinctly preserved in the imputed distribution. This preservation confirms the model's ability to retain feature stratification, particularly under heterogeneous field conditions. Deviations are minimal except for a few outliers, suggesting robustness against sensor anomalies.

In the case of temperature, the imputed values almost perfectly track the true values across the 20–35°C range, demonstrating the model's high temporal-fuzzy consistency and low prediction variance for this environmental variable.

Humidity also shows excellent alignment with the ground truth, particularly in the 60–95% range, where dense clustering along the diagonal indicates precise recovery. Scattered deviations for a small number of samples below 50% do not affect the overall imputation quality.

Lastly, the pH scatterplot highlights the model's sensitivity to subtle gradients. The imputed values closely mirror the true values in the 5.5–8.0 range, aligning well with the agronomic constraints for soil quality.

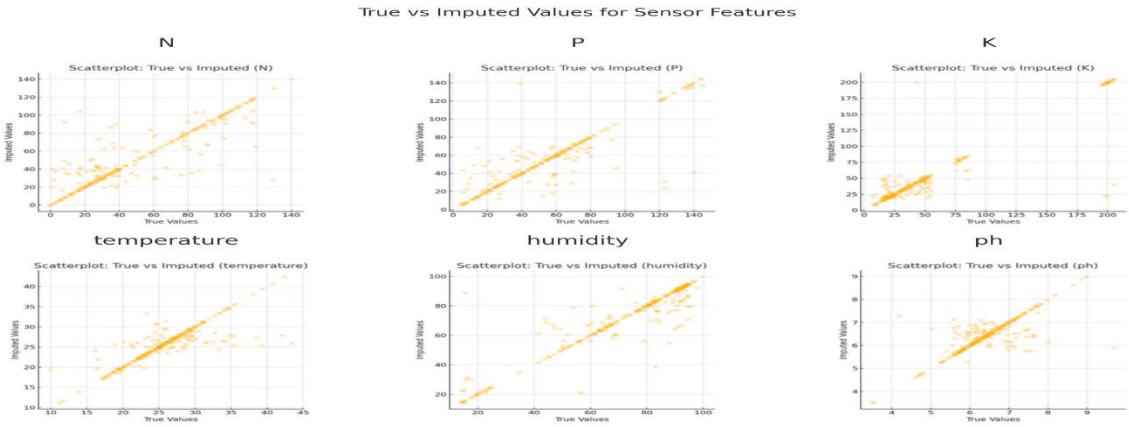

Figure 2. Results of true vs imputed values

5. CONCLUSION

This study presented a novel fuzzy-based hybrid framework, AFARE-HIM, for robust imputation of missing sensor data in IoT-enabled smart farming systems. The AFARE, integrated with a SASP mechanism, enabled semantic recovery of incomplete data without relying on expert-defined rules. By leveraging fuzzy mutual activation and temporal consistency, the model effectively captured the underlying structure and uncertainty inherent in agricultural sensor streams. The hybrid integration with MissForest and KNN further ensured that the global data dependencies and localized variations were preserved in the final imputation.Extensive experiments conducted on a real-world agricultural dataset demonstrated the superiority of AFARE-HIM over widely used imputation methods such as Mean Imputer, KNN Imputer, MissForest, MICE, SoftImpute, and Denoising Autoencoders. The proposed method achieved an RMSE of 0.684 and an R2 of 81% at 30% missingness, while also showing strong generalization across different missingness levels. Visual diagnostics, including scatterplots and error distribution heatmaps, confirmed that the imputed values remained faithful to the true data distribution.

The architecture is statistically effective, interpretable, scalable, and adaptable to various sensors and crop environments.This work provides a reliable and semantically grounded solution for data incompleteness in smart agriculture, laying the foundation for further integration with real-time decision support, anomaly detection, and predictive crop analytics.Future directions include extending the framework for time-series imputation, federated imputation across multi-farm deployments, and explainable AI integration for rule traceability and trustworthiness in autonomous farming systems.

REFERENCES

[1]Mishra, H., & Mishra, D. (2024). AI for data-driven decision-making in smart agriculture: From field to farm management. In Artificial Intelligence Techniques in Smart Agriculture (pp. 173-193). Singapore: Springer Nature Singapore.https://doi.org/10.1007/978-981-97-5878-4_11

[2]Aarif KO, M., Alam, A., & Hotak, Y. (2025). Smart sensor technologies shaping the future of precision agriculture: Recent advances and future outlooks. Journal of Sensors, 2025(1), 2460098.https://doi.org/10.1155/js/2460098

[3]Sabale, M. M., Pande, V. A., Tagalpallewar, A. A., Swami, A. G., Pawar, A. T., &Baheti, A. M. (2024). Maintaining data safety and accuracy through data integrity (DI): A comprehensive review. Research Journal of Pharmacy and Technology, 17(5), 2431-2440.http://dx.doi.org/10.52711/0974-360X.2024.00381

[4]Jin, T., Zhang, C., Zhang, Y., Yang, M., & Ding, W. (2024). A Hybrid Fault Diagnosis Method for Autonomous Driving Sensing Systems Based on Information Complexity. Electronics, 13(2), 354.https://doi.org/10.3390/electronics13020354

[5]Akande, O. (2024). Challenges and Opportunities in Machine Learning for Bioenergy Crop Yield Prediction: A Review. Available at SSRN 4898518.http://dx.doi.org/10.2139/ssrn.4898518

[6]Harsh Joshi, Rajeshwari Mistri, M. Mali, Nachiket Kapure, Preeti Kumari (2025). Precision Adaptive Imputation Network :An Unified Technique for MixedDatasets. https://doi.org/10.48550/arxiv.2501.10667

[7]Ibna Kowsar, Shourav B. Rabbani, Y. R. Hou, Manar D. Samad (2025). DeepIFSA: Deep Imputation of Missing Values Using Feature and SampleAttention. https://doi.org/10.48550/arxiv.2501.10910

[8]Shuo Tong, Runyuan Guo, Wenqing Wang, Xueqiong Tian, Lingyun Wei, Lin Zhang, Huayong Wu, Ding Liu, Youmin Zhang (2025). A Soft Sensor Method with Uncertainty-Awareness and Self-ExplanationBased on Large Language Models Enhanced by Domain Knowledge Retrieval. https://doi.org/10.48550/arxiv.2501.03295

[9]V. Kavitha, Viktor K. Prasanna, S. Lekashri, M. Venkatesan (2025). HOSNA: Boosting Smart Agriculture Efficiency WithThe Hybrid Optimization-Based Sensor Node Activation Model. Journal of machine and computing, , 509-522. https://doi.org/10.53759/7669/jmc202505040

[10]Avdesh Kumar Sharma, Abhishek Singh Rathore (2025). AGRO-Cloud Model and Smart Algorithm to Increase Agriculture Production to Improve Agriculture Quality. Current agriculture research journal, 12.0(3), 1193-1204. https://doi.org/10.12944/carj.12.3.14

[11]Vasanth Kumar N.T, C. Sagar, K. V. Daya Sagar, Monga Vishal, H S Thejas (2025). Agripulse Based on Harnessing Machine Learning. Indian Scientific Journal Of Research In Engineering And Management, 9.0(01), 1-9. https://doi.org/10.55041/ijsrem40582
[12]Ishani Rana, Palak Palak, Mayank Vashishta, A. N. Shukla, Nishant Dahiya (2025). Hybrid Intelligence for Precision Agriculture. Social Science Research Network, . https://doi.org/10.2139/ssrn.5088797
[13]Afiana Nurani, Haura Taqiya Azza Nabila, Ilham Bintang Herlambang (2025). Peran artificial intelligence dalam sistem iot untuk pertanian cerdas. JATI (JurnalMahasiswa Teknik Informatika), 9.0(1), 1446-1455. https://doi.org/10.36040/jati.v9i1.12705
[14]Armando Collado-Villaverde, Pablo Muñoz, Maria D. R-Moreno (2025). BRATI: Bidirectional Recurrent Attention for Time-Series Imputation. https://doi.org/10.48550/arxiv.2501.05401
[15]Dongming Yan, R Li, Weihua Xiong, Xiaotong Huang (2025). A Prediction Technique Based on Deep Learning for Deformation and Missing Data in the Context of Insufficient Sensor Data. Physica Scripta, . https://doi.org/10.1088/1402-4896/adab43
[16]Yueru Yan, Yue Wang, Jialin Li, Jingwei Zhang, X. H. Mo (2025). Crop Yield Time-Series Data Prediction Based on Multiple Hybrid Machine Learning Models. https://doi.org/10.20944/preprints202501.1948.v1
[17]Wen-Shan Liu, Tong Si, Aldas Kriauciunas, Marcus Snell, Haijun Gong (2025). Bidirectional f-Divergence-Based Deep Generative Method for Imputing Missing Values in Time-Series Data. Stats, 8.0(1), 7-7. https://doi.org/10.3390/stats8010007
[18]Muhammad Hameed Siddiqi, MadallahAlruwaili, Yousef Alhwaiti, Saad Alanazi, Faheem Khan (2025). A Novel Hyperâ€‘Tuned Multilayer Perceptron With Effective Stochastic Learning Strategies for Missing Values Imputation. Expert Systems, 42.0(2). https://doi.org/10.1111/exsy.13828
[19]Rahul Bordoloi, ClÃ©mence RÃ©da, Saptarshi Bej (2025). Fast Iterative and Task-Specific Imputation with Online Learning. https://doi.org/10.48550/arxiv.2501.13786
[20]Abdel-salam, M., Kumar, N. & Mahajan, S. A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning. Neural Comput&Applic **36**, 20723–20750 (2024). https://doi.org/10.1007/s00521-024-10226-x
[21]Egas JosÃ© Armando, Damien Hanyurwimfura, Omar Gatera, Athanase Nduwumuremyi (2025). Enhancing Agricultural Internet of Things Data Accuracy: Evaluating Kalman Filter-Based Sensor Denoising Techniques. https://doi.org/10.1007/s42853-025-00252-5
[22]Shenfen Kuang, Yewen Huang, Jie Song (2025). Unsupervised data imputation with multiple importance sampling variational autoencoders. Dental science reports, 15.0(1). https://doi.org/10.1038/s41598-025-87641-0
[23]Syamsu Irfan Akbar, Syed Asad Ali Shah, Sumaiya Tabassum, Hassan Askari Rabbani, Waqas Ahmad (2025). Integrating Physics-Based Models, Mathematical Optimization, and Statistical Analytics for Advancing Precision Agriculture and Sustainable Farming Practices. Â˜TheÂœcritical review of social sciences studies, 3.0(1), 291-302. https://doi.org/10.59075/wmy3cx92
[24]Aditi Singh, A. K. Awasthi, Uday Badola, Ranjeet Vasant Bidwe, Sashikala Mishra (2025). A Novel Hybrid Approach to Crop Yield Prediction: Combining Deep Learning Efficiency with Statistical Precision. International Journal of Computing and Digital Systems, 17.0(1), 1-12. https://doi.org/10.12785/ijcds/1571032941
[25]Guofeng Yang, Nuo Jin, Wenjie Ai, Zhonghua Zheng, Yuhong He, Yong He (2025). Integrating remote sensing data assimilation, deep learning and largelanguage model for interactive wheat breeding yield prediction. https://doi.org/10.48550/arxiv.2501.04487
[26]Ademir Ferreira da Silva, Ricardo LuÃ-s Ohta, Jaione Tirapu-Azpiroz, Matheus Esteves Ferreira, Daniel Vitor MarÃ§al, AndrÃ© Botelho, Tulio Coppola, Allysson Flavio Melo de Oliveira, Murilo Bettarello, Lauren Schneider, Rodrigo VilaÃ§a, Noorunisha Abdool, Vanderlei Junior, Wellington Furlaneti, Pedro Augusto Malanga, M. Steiner (2025). AI enabled, mobile soil pH classification with colorimetric paper sensors for sustainable agriculture. PLOS ONE, 20.0(1), e0317739-e0317739. https://doi.org/10.1371/journal.pone.0317739
[27]Belghachi Mohammed (2025). Smart Irrigation Systems Using AI to Optimize Water Usage. Advances in environmental engineering and green technologies book series, , 241-268. https://doi.org/10.4018/979-8-3693-7483-2.ch009
[28]Sandeep D. Kulkarni (2025). Sustainable precision agriculture: integrating artificial intelligence and iot for optimized green farming practices. Journal of informatics education and research, 5.0(1). https://doi.org/10.52783/jier.v5i1.2000
[29]Youran Zhou, Mohamed Reda Bouadjenek, Jonathan R. Wells, Sunil Aryal (2025). Handling Incomplete Heterogeneous Data using a Data-Dependent Kernel. https://doi.org/10.48550/arxiv.2501.04300
[30]Fatih Fehmi Şimşek(2025). Comparison of Agricultural Crop Type Classifications with Different Machine Learning Algorithms (RF-SVM-ANN-XGBoost) by Generating Ground Truth Data from Farmer Declaration Parcels. International journal of engineering and geosciences, 10.0(2), 207-220. https://doi.org/10.26833/ijeg.1552141
[31]Dobrislav Dobrev, Pawel Szerszen (2025). Missing Data Substitution for Enhanced Robust Filtering and Forecasting in Linear State-Space Models. Finance and economics discussion series, (2025-001), 1-38. https://doi.org/10.17016/feds.2025.001
[32]Sikandar,Smart Farming Data 2024 (SF24) (2024). https://doi.org/10.34740/kaggle/dsv/9239304
[33]Atia, M. (2025). Breakthroughs in tissue engineering techniques. Innovative Reviews in Engineering and Science, 2(1), 1-12. https://doi.org/10.31838/INES/02.01.01
[34]Surendar, A. (2024). Survey and future directions on fault tolerance mechanisms in reconfigurable computing. SCCTS Transactions on Reconfigurable Computing, 1(1), 26-30. https://doi.org/10.31838/RCC/01.01.06
[35]Kavitha, M. (2024). Environmental monitoring using IoT-based wireless sensor networks: A case study. Journal of Wireless Sensor Networks and IoT, 1(1), 50-55. https://doi.org/10.31838/WSNIOT/01.01.08
[36]Madhanraj. (2025). Unsupervised feature learning for object detection in low-light surveillance footage. National Journal of Signal and Image Processing, 1(1), 34–43.
[37]Uvarajan, K. P. (2024). Integration of artificial intelligence in electronics: Enhancing smart devices and systems. Progress in Electronics and Communication Engineering, 1(1), 7–12. https://doi.org/10.31838/PECE/01.01.02