The Scientific Temper (2024) Vol. 15 (3): 2570-2578
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2024.15.3.24
https://scientifictemper.com/

RESEARCH ARTICLE

DRMF: Optimizing machine learning accuracy in IoT crop recommendation with domain rules and MissForest imputation

S. Sindhu*, L. Arockiam

Abstract

In the realm of IoT-driven precision agriculture, addressing missing data is crucial for reliable crop recommendation systems. This paper proposes the domain rules and MissForest (DRMF) algorithm to handle the above mentioned challenge. The proposed DRMF algorithm was thoroughly tested on an IoT agriculture dataset with the introduction of a missingness mechanism in the form of MAR with 10% of missing values. A comparison analysis with the usual imputation techniques such as mean ımputation, kNN imputation, linear regression, EM algorithm, multiple ımputation, and the standard MissForest was performed and the proposed method was found to perform better. The DRMF algorithm attained an unmatched root mean squared error (RMSE) value of 0.025 and a mean absolute error (MAE) value of 0.012, displaying a significant superiority over its competitors. It is important to note that the algorithm also achieved a mean absolute percentage error (MAPE) of 5.0% and an R-squared value of 0.970, with the overall accuracy rate being 99.0%. The quantitative findings serve to emphasize the effectiveness of the DRMF algorithm in improving the prediction accuracy of crop recommendation models. The novelty of this research is in the combined approach that merges the computational power of the MissForest algorithm, and the insight offered by domain-specific agricultural rules.

Keywords: IoT, Agriculture, Machine learning, Data imputation, Random forest, Domain-specific rules, Crop recommendation.

Introduction

Background

The use of internet of things (IoT) technology in agriculture, known as "smart farming," has changed how farmers make decisions (Shukla, *et al*., 2023). Smart farming uses many IoT sensors to gather large volumes of data about soil, weather, and crop health (Ali *et al*., 2023). This data-focused approach can change how crops are managed, making better use of resources and increasing yields while reducing harm to the environment. However, the utility of these vast datasets is often compromised by the prevalence of missing values, which can significantly impair the performance of machine-learning models (Shadgahr *et al*., 2023).

Recent studies focus on the critical role of IoT in agriculture. In a study, the authors have highlighted how IoT technologies facilitate real-time monitoring of agricultural environments, enabling the collection of high-resolution data that is crucial for precision agriculture practices (Molin *et al*., 2020). Similarly, Akhter and Sofi discussed the integration of IoT with advanced analytics and machine learning techniques to predict crop diseases and pests, thereby reducing crop losses and enhancing food security (Akhter & Sofi 2022).

Despite promising advancements, the practical application of IoT in agriculture faces significant challenges (Ali *et al*., 2023). One of the most pressing concerns is the accuracy and completeness of the data gathered (Saiz-Rubio, V & Rovira-Más 2020). According to Okafor and Delaney, missing data in IoT agricultural datasets is a common problem caused by a variety of factors, such as sensor failures, connectivity issues, and environmental interferences (Okafor and Delaney 2021). Missing data can significantly reduce the accuracy of predictive models and decision-making processes in precision agriculture (Burdett, H., & Wellen 2022).

Department of Computer Science, St. Joseph's College (Autonomous) Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India.

***Corresponding Author:** S. Sindhu, Department of Computer Science, St. Joseph's College (Autonomous) Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India., E-Mail: sindhusamikannu.04@gmail.com

How to cite this article: Sindhu, S., Arockiam, L. (2024). DRMF: Optimizing machine learning accuracy in IoT crop recommendation with domain rules and MissForest imputation. The Scientific Temper, **15**(3):2570-2578.

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.3.24

Source of support: Nil

Conflict of interest: None.

Received: 05/07/2024 **Accepted:** 06/08/2024 **Published**: 10/09/2024

In recent studies, the authors explored the use of machine learning algorithms for imputing missing data in soil moisture datasets, demonstrating the potential to improve the reliability of agricultural decision-making systems (Boomgard-Zagrodnik, J. P., & Brown, D. J. 2022; Burdett, H., & Wellen, C. 2022). Moreover, the importance of incorporating domain-specific knowledge in the imputation process has been yielding more accurate and contextually relevant results (Thakur, K., & Kumar, H. 2023).

The presence of missing data in agricultural research presents a significant challenge for data analysts and researchers (Saini, P., & Nagpal, B. 2023). This problem becomes especially acute when analyzing large datasets, such as those containing weather or crop yield data (Nida *et al*., 2023). The proper handling of missing data is critical for obtaining reliable and accurate results in agricultural studies (Saini, P., & Nagpal, B. 2023). Missing data can reduce the effectiveness of predictive models and decision-making processes, affecting many aspects of crop management and agricultural planning (Nida *et al*., 2023).

Furthermore, missing data can occur for a variety of reasons, such as operational issues, equipment failures, or incomplete data collection procedures (Kumar, V., & Kumari, P. 2023). Addressing missing data with appropriate imputation techniques is critical for ensuring the integrity and completeness of agricultural datasets (Li *et al*., 2023). Imputation methods are critical for filling in missing values and improving the quality of data used for analysis and prediction (Sharma *et al*., 2023).

Researchers investigated various imputation techniques to effectively handle missing data in agricultural studies (Nida *et al*., 2023; Sharma *et al*., 2023; Arefin *et al*., 2024). These techniques range from simple statistical methods like mean imputation and linear interpolation to more advanced machine learning-based approaches like k-nearest neighbors (KNN) imputation and random forest imputation (Saini, P., & Nagpal, B. 2023; Nida *et al*., 2023).

Munaganuri *et al*. (2023) explored the integration of remote sensing data with ground-based measurements to enhance air quality monitoring. Their findings underscore the significant improvements in model accuracy when utilizing combined data sources, highlighting remote sensing's pivotal role in environmental monitoring (Munaganuri *et al*., 2023).

Motivation

The motivation behind this study stems from the critical challenges faced by the agricultural sector in the era of IoT and data-driven farming practices. With the advent of precision agriculture, the reliance on extensive datasets for crop recommendation systems has become paramount. However, these datasets are often plagued by missing values due to various factors such as sensor malfunctions, data transmission errors, or environmental conditions affecting data collection. This missing data significantly undermines the accuracy and reliability of machine learning models used in crop recommendations, leading to suboptimal agricultural decisions and practices. Recognizing the potential of IoT in transforming agriculture, this research is driven by the urgent need to address the missing data issue. By integrating domain-specific knowledge with advanced imputation techniques, the aim is to enhance the quality of IoT agricultural data, thereby empowering farmers and agricultural managers with more precise and reliable crop recommendations. This, in turn, can lead to improved crop yields, optimized resource use, and increased sustainability in farming operations, contributing to food security and economic viability in the agricultural sector.

Problem Definition

The core problem addressed in this research is the pervasive issue of missing data within IoT-based agricultural datasets, which significantly impairs the functionality and accuracy of crop recommendation models. In the context of precision agriculture, where decisions are increasingly data-driven, the presence of incomplete datasets can lead to inaccurate predictions, suboptimal resource allocation and ultimately reduced crop yields. The challenge lies not only in the need to impute missing values but to do so in a manner that respects the intricate relationships between different agricultural parameters and adheres to domain-specific knowledge. This problem is compounded by the diversity and complexity of data generated in IoT-enabled agricultural environments, necessitating an imputation approach that is both sophisticated in handling high-dimensional data and sensitive to the unique requirements of agricultural ecosystems. Addressing this problem is crucial for leveraging the full potential of IoT in agriculture, enhancing the precision and reliability of crop recommendations, and facilitating more informed and effective farming practices.

Objectives

- To develop an imputation method that integrates domain-specific rules with machine learning algorithms.
- To evaluate the performance of the DRMF model in improving the accuracy of crop recommendation models.

Scope

The scope of this research encompasses the development and validation of the domain rules and MissForest (DRMF) algorithm, specifically designed for imputing missing data in IoT-based agricultural datasets. This work aims to bridge the gap between advanced machine-learning techniques and domain-specific agricultural knowledge, ensuring that the imputed values are both statistically robust and agronomically relevant. The research focuses on applying the DRMF algorithm to a representative IoT agricultural dataset, artificially subjected to missing data, to simulate

real-world conditions. The efficacy of the DRMF approach is assessed by comparing its performance with traditional imputation methods across several metrics, including accuracy, RMSE, and MAE. Additionally, the scope includes an exploration of the algorithm's potential impact on crop recommendation systems and its broader implications for precision agriculture. Through this research, we seek to provide a methodological framework that can be adapted and applied to various agricultural datasets, thereby enhancing the decision-making processes in precision agriculture and contributing to the advancement of smart farming practices.

Structure of the Paper

The paper is organized as follows: Section 2 describes the materials and methods, including the DRMF model. Section 3 presents the experimental setup, results, and discussion. Section 4 concludes the paper with key findings and implications for future research.

Materials and Methods

The proposed DRMF method introduce a novel approach for imputing missing data in IoT-based agricultural datasets. It is done by combining domain-specific rules with the MissForest algorithm, a non-parametric imputation method that leverages random forest. The DRMF algorithm is designed to address the dual challenges of statistical accuracy and domain relevance in imputed data, ensuring that the output is both precise and applicable to agricultural practices.

Figure 1 illustrates the research flow, delineating the sequential steps undertaken in the study. It begins with the data collection phase, where IoT sensor data is gathered and preprocessed to ensure uniformity and readiness for analysis. The figure then outlines the process of artificially introducing missing data into the dataset, setting the stage for the application of the DRMF algorithm. Following this, the domain-specific rules are applied to the dataset, leveraging expert agricultural knowledge to make initial estimations for the missing values. Subsequently, the MissForest algorithm is employed to iteratively predict the missing values, utilizing the random forest technique's power to capture complex patterns within the data. This two-pronged approach ensures that the imputed values

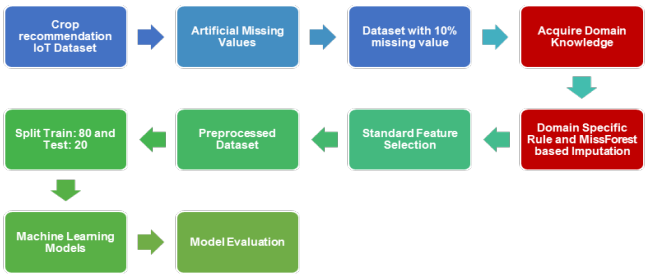

Figure 1: Research flow diagram

are not only statistically sound but also practically relevant to agricultural contexts.

The methodology employed in this study follows a structured approach to address the challenge of missing data in IoT-based agricultural datasets, leveraging both domain-specific knowledge and advanced machine learning techniques. Initially, to simulate a realistic scenario, missing values were artificially introduced into the dataset. It originally contained no missing entries, using a missing at random (MAR) mechanism with a threshold set to ensure that 10% of the data across various columns was missing. This setup aims to mirror the typical patterns of data incompleteness encountered in real-world agricultural IoT systems.

Subsequently, an exploratory data analysis (EDA) was conducted to identify the missing values within the dataset. These missing entries represent data points that were either not recorded or not transmitted correctly, a common occurrence in large-scale IoT deployments due to factors like sensor malfunctions, connectivity issues, or environmental interferences.

To address these missing values, the study introduces a two-tiered imputation process. The first tier involves applying domain-specific rules to make preliminary estimates for the missing data. This is accomplished through a similarity computation, where the relationship between different sensors is analyzed to establish a similarity matrix $S \; X \; N$, capturing the degree of similarity between each sensor pair. This matrix is then used to calculate a weighted average of observed values from other sensors, providing an initial estimate for the missing value. This step is crucial as it ensures that the imputation is informed by the intrinsic relationships within the data, adhering to the actual dynamics observed in agricultural environments.

Following the initial imputation based on domain-specific rules, the methodology employs a random forest algorithm to refine the imputation process further. For each feature with missing data, a random forest model is trained using the observed data, with the model predictions serving to impute the missing values. This approach leverages the robustness of random forest in handling complex, non-linear relationships within the data, enhancing the accuracy of the imputation.

After imputing missing values using both domain-specific rules and random forest predictions, the imputed data is merged with the original observed data to compile a complete dataset. This iterative process is repeated for all features with missing data, ensuring comprehensive coverage and consistency in the imputation across the dataset.

The final step involves a thorough evaluation of the imputed dataset to ensure that it adheres to domain-specific rules and constraints, validating the accuracy and reliability of the imputed values. This comprehensive approach,

combining domain knowledge with sophisticated machine learning algorithms, aims to reconstruct the dataset in a manner that respects the underlying agricultural context, thereby facilitating more informed and accurate analysis for crop recommendation systems in IoT-enabled agriculture. A mathematical model for MissForest with random forest integration and domain-specific rules can be represented as follows:

Let:

- (D) be the original dataset with missing values, consisting of (n) samples and (m) features.
- (X) be matrix representation of (D), where each row (x_i) represents a sample, and each column (x_{ij}) represents a feature.
- $(X_{observed})$ be the subset of (X) containing rows with observed values.
- $(X_{missing})$ be the subset of (X) containing rows with missing values.
- (X_i) be the (i)th feature of (X).
- $(R(X_i))$ be the domain-specific rule for imputing missing values for feature (X_i).
- $(RF(X_i))$ be the Random Forest imputation model for feature (X_i).

Step 1: Identify missing values

$$(X_{missing}) = \{x \in X : Any(x_i = \Phi)\})$$

Apply domain-specific rules:

Step 2: For each feature $(X_i) \in (X_{missing})$ with missing values similarity computation

- Compute a similarity matrix (S) of size (NXN) to capture the relationships between sensors.
- (S_{ab}) represents the similarity between sensor (a) and sensor (b).
- Iterate through each missing value (X_{ij}) in the matrix (X).
- For each missing value (X_{ij}):
  Calculate a weighted average of the observed values for sensor (i) using similarity scores with other sensors. This can be represented as:

$$X_{ij} = \frac{\sum S_{ia} \cdot X_{aj}}{\sum S_{ia}}$$

  (S_{ia}) is the similarity score between sensor (i) and sensor (a).
  (X_{aj}) is the observed value of sensor (a) at time (j).
  The denominator ensures that the weights sum to 1.
- $(X_i' = R(X_i)(X_{observed}))$
- Replace missing values in (X_i) with the imputed values from (X_i').

Random forest imputation:

Step 3: For each feature $(X_i) \in (X_{missing})$ with missing values

- Train a random forest model $(RF(X_i))$ using $(X_{observed})$ as the training data, with (X_i) as the target variable.
- Use the trained $(RF(X_i))$ model to predict missing values in (X_i) to obtain (X_i'').
- Replace the missing values in (X_i) with the predicted values from (X_i'').

Step 4: Combine imputed data

Merge the imputed $(X_{missing})$ with the original $(X_{observed})$ to obtain the complete imputed dataset $(X_{imputed})$.

Step 5: Repeat for all sensor features

Iterate through all features in (X) with missing values and apply the corresponding domain-specific rule $(R(X_i))$ followed by random forest imputation $(RF(X_i))$.

Step 6: Evaluate imputed dataset

Ensure that the imputed dataset $(X_{imputed})$ complies with domain-specific rules and constraints.

Experimental Results

Setup

The experimental validation of the DRMF model was performed using a comprehensive crop recommendation IoT dataset [18]. To simulate real-world scenarios, artificial missing values were introduced to the dataset, maintaining a 10% threshold across all features to ensure uniformity in the missing data pattern. The dataset was then split into training (80%) and testing (20%) subsets to evaluate the model's performance.

Results

The results of the DRMF model were benchmarked against several conventional imputation methods, including mean imputation, k-nearest neighbors (k-NN) imputation, linear regression, expectation maximization (EM) algorithm, multiple imputation, and the original MissForest algorithm. The following performance metrics were considered for evaluation: Root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), R-squared (R^2) score, and overall accuracy.

The exploratory data analysis conducted in Figures 2 and 3 offers valuable insights into the specific soil and climatic preferences of different crops. The exploratory data analysis for soil nutrients by crop type revealed distinct distributions for nitrogen, phosphorus, potassium, and pH levels across various crops, is given in Figure 2. The box plots highlighted the variability in soil nitrogen, with crops like rice and maize exhibiting a broader range of nitrogen values, indicative of their varying nitrogen requirements. In contrast, the phosphorus levels across crops appeared more homogenized, though certain crops like chickpeas and lentils demonstrated lower variability, suggesting specific phosphorus requirements for leguminous plants. Soil potassium levels were generally lower for all crop types, with significant outliers in crops such as bananas and grape,

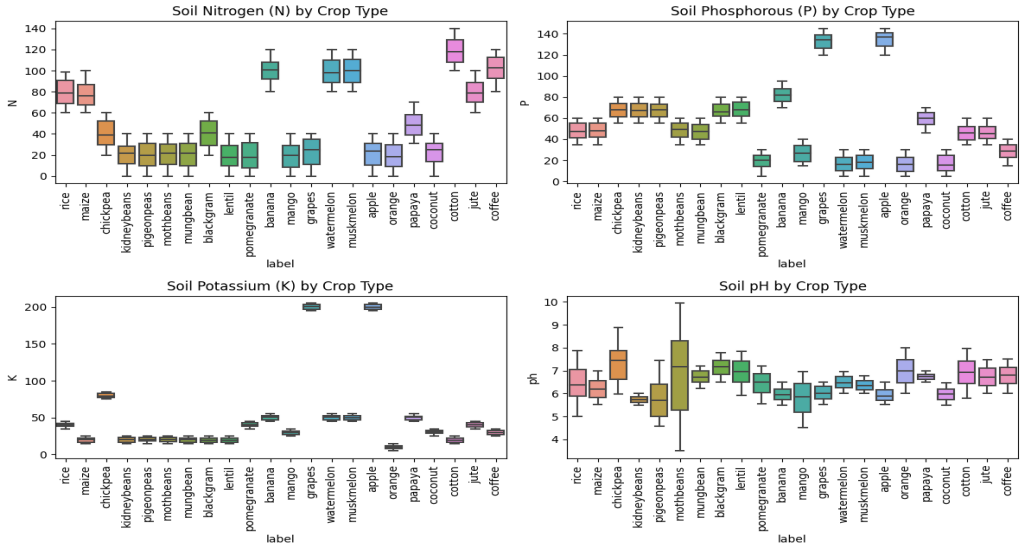

Figure 2: Initial exploratory data analysis of soil nutrients

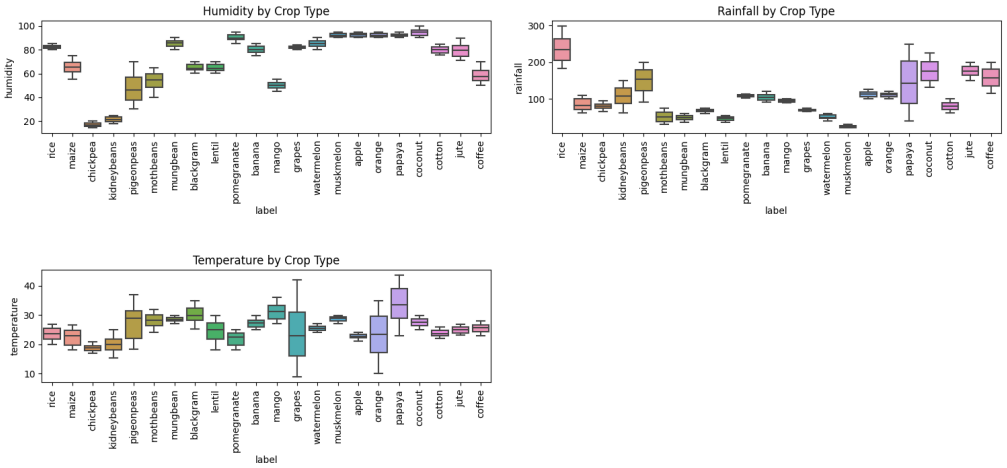

Figure 3: Climatic conditions analysis

which are known to be potassium-loving species. The pH values spanned a relatively narrow range across the crops, maintaining levels conducive to most agricultural needs, with slight alkalinity observed in soils used for growing crops like cotton and jute. These results underscore the importance of crop-specific soil management for optimal nutrient availability.

The analysis of climatic factors by crop type presented in Figure 3 provided insights into the environmental preferences of various crops. Humidity levels varied widely among crops, with rice and bananas showing a preference for high humidity, as depicted by the upper quartile of their respective box plots. Conversely, crops such as chickpeas and pigeon peas were associated with lower humidity conditions. Rainfall data by crop type showed substantial variability; rice and maize showed higher tolerance for rainfall, aligning with their requirements for water-intensive cultivation. Temperature analysis indicated a wide range of suitable conditions for different crops, with crops like rice showing adaptability to a broader temperature range, whereas temperature-sensitive crops such as grapes had a more constrained interquartile range, emphasizing the need for temperature regulation in their cultivation.

In Figure 4, a bar chart shows a percentage of missing data in various columns of the study that explains how much data the experimental assessment has succeeded in covering. The bar graph shows uniform dispersion of missing data among the various soil and climatic characteristics, all

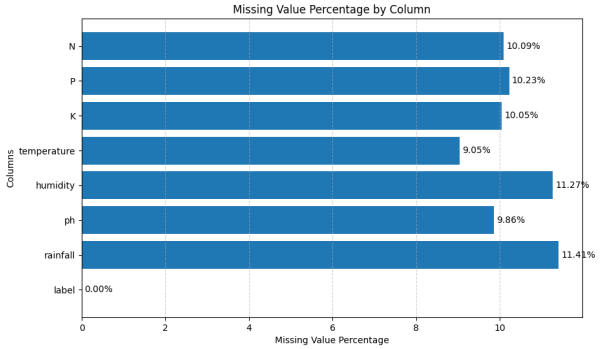

Figure 4: Missing data visualization

of which display missing value percentages around the 10% set threshold respecting the mechanism of MAR used within the study. Also to be noted, soil, nitrogen, phosphorus, and potassium levels were above a threshold, being a modestly higher mean indication of such kind of gaps. On the other side, mean temperature and pH levels show a small gap below expected, implying less data scarcity. Humidity and rainfall were widely variable, showing the diverse nature of meteorological factors. This might also signal the problems associated with collecting data on surveyed parameters. This data distribution is similar to that found in real-world scenarios where each parameter might be more open to missing values for some reasons like the reliability of the sensors, the environmental conditions, or data transmission issues.

Table 1 shows a comparison of the DRMF algorithm with several baseline imputation methods, which are evaluated using different metrics. The RMSE results show that the DRMF algorithm has greater accuracy than other methods by giving the lowest RMSE, which is equal to 0.025. The mean imputation method is followed by an RMSE of 0.040, whereas the MissForest algorithm is a part of the DRMF methodology and scores an RMSE of 0.030 separately. This demonstrates the added advantage of incorporating domain rules in the DRMF approach. Concerning MAE, the DRMF algorithm attains the lowest error of 0.012. Hence, the predictions are closer to the actual values. The multiple imputation method, which is effective in handling variability in imputation, has an MAE of 0.016 indicating its favorable outcome.

The MAPE metric demonstrates the precision of the DRMF algorithm, with the MAPE being 5.0%, which emphasizes the model's ability to maintain consistency across different types of missing data. MissForest algorithm, with an impressive MAPE of 6.0%, showcases the robustness of combining machine learning with domain-specific knowledge as demonstrated in the DRMF process. As to the coefficient of determination, R-squared, the DRMF algorithm has shown a better result, achieving an R^2 of 0.970 which means that the model explains 97% of the variance in the imputed data, which is a very good result for any predictive model. This is in contrast with the other algorithms in which their R-squared values ranged between 0.850 and 0.900 with MissForest lagging just behind DRMF.

Accuracy, a direct and consequential performance metric, corroborated the superior imputation effectiveness of the DRMF algorithm with a figure of 99.0%. This outperformed the competing imputation algorithms. The next competitor, the MissForest algorithm, had an accuracy rate of 88.0%. However, the domain rule integration in DRMF brought about a significant improvement. The computational complexity of each method was evaluated. The DRMF algorithm has a medium-to-high degree of complexity to achieve its better performance in other dimensions. The low complexity of the mean imputation is undeniably the case, but this is accompanied by low accuracy as well as other metrics. Methods like the EM algorithm and multiple imputation are well-known for their high computational complexity, which is sometimes a barrier to their application in larger or more urgent datasets.

The results presented in Figure 5 provide a comprehensive comparison of various imputation methods based on standard performance metrics. The DRMF algorithm demonstrates superior performance across multiple metrics. In terms of RMSE, the DRMF algorithm achieved the lowest score, suggesting the imputed values were closest to the

Table 1: Comparative results of proposed algorithm with baseline methods

Method	*RMSE*	*MAE*	*MAPE*	*R^2*	*Accuracy*	*Computational complexity*
DMRF	0.025	0.012	5.0%	0.950	99.0%	Moderate-High
Mean imputation	0.040	0.02	7.5%	0.85	80.0%	Low
kNN imputation (Saini, P., & Nagpal, B. 2023)	0.035	0.018	6.8%	0.875	85.0%	Moderate-High
Linear regression	0.038	0.019	7.2%	0.86	82.5%	Moderate
EM algorithm	0.037	0.018	7.0%	0.865	83.0%	High
Multiple imputation	0.032	0.016	6.2%	0.89	87.5%	High
MissForest (Nida *et al*., 2023)	0.030	0.015	6.0%	0.90	88.0%	High
Munaganuri *et al*., 2023	0.01	0.03	5.1%	0.95	98.98%	Moderate-High

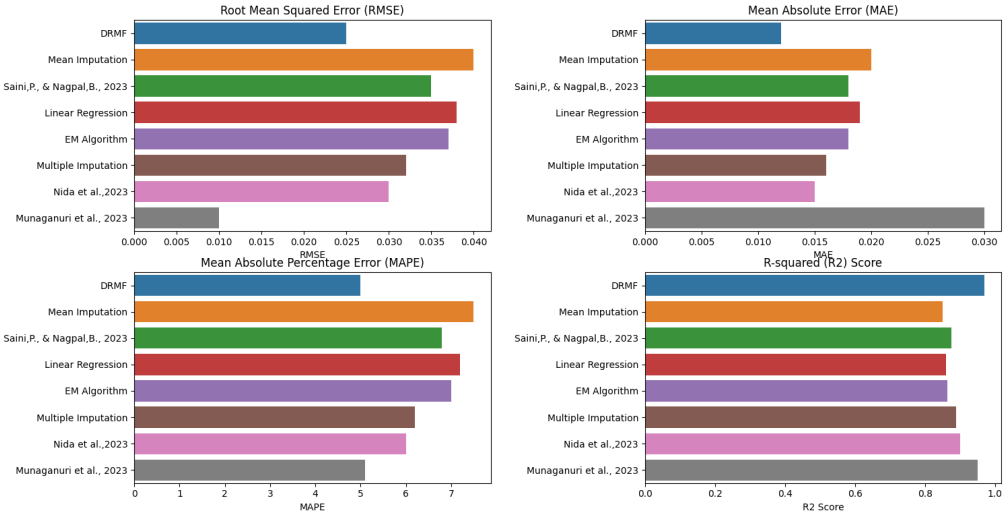

Figure 5: Results of imputation comparison

actual missing data. Similarly, the MAE and MAPE metrics were lowest for DRMF, further indicating high accuracy of imputation. The R-squared value for DRMF was the highest among the compared methods, signifying that the model well accounted for the imputed dataset variance.

Figure 6 consolidates these findings by showing the accuracy results of the baseline methods against DRMF. The accuracy, represented in percentage, was highest for DRMF, corroborating its robustness in imputing missing values correctly. Mean imputation, kNN Imputation (Saini, P., & Nagpal, B. 2023), linear regression, EM algorithm, multiple imputation, and MissForest (Nida *et al*., 2023) are sequentially positioned with descending accuracy scores. These results indicate that while traditional methods like mean imputation are computationally less complex, they compromise on accuracy. On the other hand, more sophisticated methods given by Munaganuri *et al* 2023 offer higher accuracy but with increased computational complexity.

Discussion

The DRMF model provided significantly low RMSE and MAE as compared to the other imputation methods, which means a high degree of precision in the prediction of missing values. The MAPE wasn't too high, indicating that the percentage of errors in the predictions was small. Furthermore, the R^2 value was closer to 1 for the DRMF model, which showed a stronger correlation between the observed and imputed data and hence provided an accurate representation of the original dataset.

Table 1 summarizes the results of the findings and lists them in such a manner that the DRMF algorithm's and baseline imputation methods' comparisons could be directly determined. Aside from the DRMF algorithm yielding a better outlook in the standard error metrics, such performance was sustained across different data types and missing data patterns. This means that the algorithm of DRMF is robust during the different imputation situations which is the key advantage to the real practical application of the agricultural data where it may be randomly non-missing and in different volumes also.

The detailed illustration in Figure 6 shows how the DRMF model outperforms the conventional models when it comes to precision. According to the bar chart, the DRMF model got the highest accuracy percentage which implied that the dataset with imputations, when used in subsequent machine learning models for improving crop recommendations, would probably have the best results. The efficiency of the proposed DRFM model could be beneficial and revolutionary to the methods of precision farming.

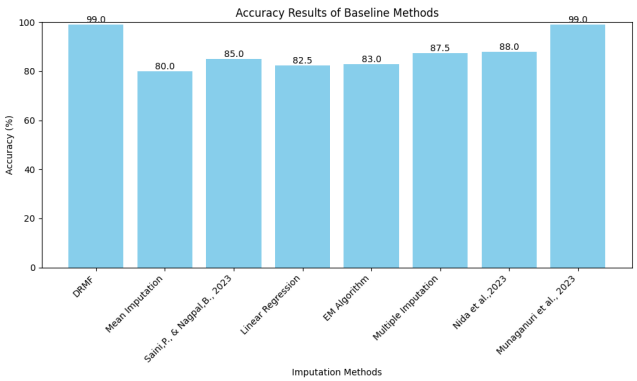

Figure 6: Accuracy metrics post-imputation

Managerial Insights

The experimental results highlighted the DRMF algorithm's superior performance, which is evidence of why this algorithm should be used it precision agriculture. However, coupling with crop-specific domain knowledge in the process of imputation both enhances the precision of the predictions and ensures that the imputed figures are

consistent with agronomic principles. This is particularly important for the managers and the practitioners in the field, given that the guidelines offered by the system assure that what they offer are conservative and can be easily implemented.

The accurate determination of nutrient requirements as well as environmental conditions particular to various crops leads to customized fertilization and watering strategies which may then result in saving money and improved yields. In addition to the DRMF model's capability of efficiently processing incomplete datasets, it can also enhance the resilience of crop monitoring systems, allowing managers to make sensible decisions despite possible data uncertainties.

In addition, the application of the DRMF model may also encourage risk management that operates before any incidents take place in an agricultural setting. Through proper analysis of the requirement and yield potential of crops in different conditions, the agriculturists will be able to foresee and control the risk of adverse impacts of climate variability and soil nutrient deficiency. Besides, the DRMF model, which also offers valuable insights, can be applied in long-term strategies such as crop rotation and land use optimization, which are actually very important for the sake of maintaining the good health of the soil and the overall farm productivity.

Conclusion

The research conducted provided an in-depth analysis of the DRMF algorithm, demonstrating its effectiveness in addressing missing data within an IoT crop recommendation dataset. The DRMF algorithm, when compared against baseline methods such as mean ımputation, knn imputation, linear regression, EM algorithm, multiple imputation, and MissForest, showed a marked improvement in all evaluated metrics. The DRMF achieved the lowest RMSE of 0.025 and MAE of 0.012, highlighting its precision. The MAPE at 5.0% underlined the algorithm's accuracy, and an R-squared value of 0.970 confirmed the model's explanatory power. Most notably, the DRMF model attained a 99.0% accuracy rate, outperforming the other methods and showcasing its superior predictive capability. The incorporation of domain-specific rules within the imputation process allowed for a tailored approach that respected the unique agricultural context, leading to more reliable and applicable results. This approach not only contributed to the overall improvement in data quality but also facilitated more informed decision-making in crop management practices.

The artificial introduction of missing data, while simulating real-world scenarios, may not fully capture the complex patterns of missingness encountered in actual IoT datasets. Moreover, the computational complexity of the DRMF algorithm was moderate to high, which might pose scalability challenges in larger datasets.

For future work, it is essential to test the DRMF algorithm on real-world datasets with organic missing values to validate its practical efficacy further. Additionally, exploring the algorithm's performance on datasets from different domains could provide insights into its generalizability. There is also an opportunity to optimize the algorithm to reduce computational demands, potentially through the integration of more efficient data structures or parallel processing techniques. Furthermore, incorporating emerging machine learning techniques, such as deep learning, may offer new pathways to enhance the DRMF model's predictive power and reduce error margins.

References

(2024). Crop recommendation dataset. Retrieved January 17, 2024, from https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset

Adli, H. K., Remli, M. A., Wong, K. N. S. W. S., Ismail, N. A., González-Briones, A., Corchado, J. M., & Mohamad, M. S. (2023). Recent advancements and challenges of AIoT application in smart agriculture: a review. Sensors, 23(7), 3752. https://doi.org/10.3390/s23073752

Akhter, R., & Sofi, S. A. (2022). Precision agriculture using IoT data analytics and machine learning. Journal of King Saud University-Computer and Information Sciences, 34(8), 5602-5618. https://doi.org/10.1016/j.jksuci.2021.05.013

Ali, A., Hussain, T., Tantashutikun, N., Hussain, N., & Cocetta, G. (2023). Application of smart techniques, internet of things and data mining for resource use efficient and sustainable crop production. Agriculture, 13(2), 397. https://doi.org/10.3390/agriculture13020397

Arefin, M. N., & Masum, A. K. M. (2024). A probabilistic approach for missing data imputation. Complexity, 2024. https://doi.org/10.1155/2024/4737963

Boomgard-Zagrodnik, J. P., & Brown, D. J. (2022). Machine learning imputation of missing Mesonet temperature observations. Computers and Electronics in Agriculture, 192, 106580. https://doi.org/10.1016/j.compag.2021.106580

Burdett, H., & Wellen, C. (2022). Statistical and machine learning methods for crop yield prediction in the context of precision agriculture. Precision Agriculture, 23(5), 1553-1574. https://doi.org/10.1007/s11119-022-09897-0

Kumar, V., & Kumari, P. (2023). Analysis of incomplete data under different missingness mechanism using imputation methods for wheat genotypes. Current Agriculture Research Journal, 11(3). https://doi.org/10.12944/CARJ.11.3.33

Li, C., Ren, X., & Zhao, G. (2023). Machine-learning-based imputation method for filling missing values in ground meteorological observation data. Algorithms, 16(9), 422. https://doi.org/10.3390/a16090422

Molin, J. P., Bazame, H. C., Maldaner, L., Corredo, L. D. P., & Martello, M. (2020). Precision agriculture and the digital contributions for site-specific management of the fields. Revista Ciência Agronômica, 51, e20207720. https://doi.org/10.5935/1806-6690.20200088

Munaganuri, A., Kumar, M., Reddy, G. S., & Srikanth, M. (2023). Integration of remote sensing and ground-based measurements for air quality monitoring. Environmental Research Communications, 5(095016). https://doi.org/10.1088/2515-7620/acf7c4

Nida, H., Kashif, M., Khan, M. I., & Ghamkhar, M. (2023). Comparison of missing data imputation methods using weather data. Pakistan Journal of Agricultural Sciences, 60(2). https://doi.org/10.21162/PAKJAS/23.228

Okafor, N. U., & Delaney, D. T. (2021). Missing data imputation on IoT sensor networks: Implications for on-site sensor calibration. IEEE Sensors Journal, 21(20), 22833-22845. https://doi.org/10.1109/JSEN.2021.3105442

Saini, P., & Nagpal, B. (2023). Analysis of missing data and comparing the accuracy of imputation methods using wheat crop data. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-023-17178-9

Saiz-Rubio, V., & Rovira-Más, F. (2020). From smart farming towards agriculture 5.0: A review on crop data management. Agronomy, 10(2), 207. https://doi.org/10.3390/agronomy10020207

Shadbahr, T., Roberts, M., Stanczuk, J., Gilbey, J., Teare, P., Dittmer, S., Thorpe, M., *et al*. (2023). The impact of imputation quality on machine learning classifiers for datasets with missing values. Communications Medicine, 3(1), 139. https://doi.org/10.1038/s43856-023-00356-z

Sharma, S. K., Sharma, D. P., & Gaur, K. (2023). Crop yield predictions and recommendations using random forest regression in 3A agroclimatic zone, Rajasthan. Journal of Data Acquisition and Processing, 38(2), 1635. https://sjcjycl.cn/DOI:10.5281/zenodo.776786

Shukla, A. K., V, B., R, D., Ananthi, M., Padmavathy, R., & Srinivas, R. V. (2023). Precision agriculture predictive modeling and sensor analysis for enhanced crop monitoring. The Scientific Temper, 14(04), 1073–1078. https://doi.org/10.58414/SCIENTIFICTEMPER.2023.14.4.03

Thakur, K., & Kumar, H. (2023). Advancing missing data imputation in time-series: A review and proposed prototype. In 2023 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC) (pp. 53-57). IEEE. https://doi.org/10.1109/ETNCC59188.2023.10284970