INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

A Novel Composite DTERM-MDI Model for the Recovery of Missing Data in Internet of Medical Things

P Iris Punitha[1]*, **J G R Sathiaseelan**[2]

1 Assistant Professor, Department of Computer Applications, Bishop Heber College (Affiliated to Bharathidasan University), Tiruchirappalli, 620017, Tamil Nadu, India
2 Associate Professor, Department of Computer Science, Bishop Heber College (Affiliated to Bharathidasan University), Tiruchirappalli, 620017, Tamil Nadu, India

OPEN ACCESS

Received: 14-08-2023
Accepted: 15-09-2023
Published: 27-10-2023

Citation: Punitha PI, Sathiaseelan JGR (2023) A Novel Composite DTERM-MDI Model for the Recovery of Missing Data in Internet of Medical Things. Indian Journal of Science and Technology 16(40): 3479-3490. https://doi.org/10.17485/IJST/v16i40.2064

* **Corresponding author**.

irispunitha.ca@bhc.edu.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Background/Objectives: The Internet of Things (IoT) relies on consistent data delivery, crucial for maintaining service quality. However, challenges like connection issues, external threats, and sensor malfunctions can lead to data insufficiency, affecting IoT applications. Addressing the problem of missing data in the vast streams of IoT-generated data is essential. This paper introduces a novel Composite DTERM Missing Data Imputation (MDI) model for the Internet of Medical Things (IoMT) aimed at robustly recovering missing data. **Methods:** The Composite DTERM-MDI model comprises three phases: Dual Strategy based Missing Data Imputation (DS-MDI), Two Tier Missing Data Imputation (TT-MDI), and Ensemble Regression Model (ERM) for progressive imputation of Missing Not at Random (MNAR) type data. The effectiveness of the Composite DTERM-MDI model is demonstrated using the cStick dataset, achieving significantly improved accuracy, precision, recall, and F-measure compared to the original dataset. Additionally, a comparison of the Composite DTERM-MDI model with two standard imputation techniques, MICE and SICE, is performed using the UCI car dataset. **Findings:** Experimental results showcase the superiority of the Composite DTERM-MDI model-based imputed cStick dataset, with accuracy at 99.12%, precision at 99.98%, recall at 98.53%, and F-measure at 98.89%, outperforming the original cStick dataset. The study highlights the Composite DTERM-MDI model's efficiency and accuracy in addressing missing data challenges in IoMT, which is vital for informed medical decision-making. **Novelty:** Furthermore, a comparison of the Composite DTERM-MDI model with MICE and SICE using the UCI car dataset evaluates accuracy and F-measure across four classification algorithms (PMM, POLYREG, CART, and LDA). The Composite DTERM-MDI model achieves accuracy rates of 91.08%, 81.58%, 97.74%, and 97.74%, along with F-measures of 84.97%, 89.58%, 98.71%, and 99.17% for PMM, POLYREG, CART, and LDA, respectively. This comparison demonstrates the model's performance against established imputation techniques in a different context.

Keywords: Internet of Medical Things (IoMT); Missing data imputation; Missing completely at random (MCAR); Missing at random (MAR); Missing not at random (MNAR); DTERM model

1 Introduction

The Internet of Medical Things (IoMT) is a rapidly evolving and innovative technology that has transformed the healthcare industry in many ways[(1)]. One of the most significant benefits of IoMT is that it allows healthcare professionals to collect, analyze, and transmit health-related data from a range of medical devices, leading to more personalized and efficient patient care. The technology enables doctors and other healthcare professionals to monitor patients remotely and in real-time, providing them with critical insights into their health conditions, medication adherence, and overall wellness[(2)]. This allows healthcare professionals to intervene earlier and more effectively, ultimately leading to better patient outcomes.

Despite the many benefits of IoMT, missing data is a common issue that can arise due to a variety of reasons[(3)]. For example, medical devices may malfunction, resulting in missing data. In addition, data transmission errors can occur, leading to incomplete data sets. Finally, patient non-compliance with data collection protocols can also lead to missing data. The missing data can be classified into three categories: MCAR, MAR, and MNAR, each with distinct characteristics that can impact the precision and reliability of analysis results[(4)]. MCAR is when the missing data is independent of both the observed and unobserved variables, while MAR is when the missing data is dependent on the observed variables but not the unobserved ones. MNAR is when the missing data is dependent on the unobserved variables, and it is the most challenging type to handle.

Missing data in medical datasets can lead to biased or incomplete analysis results, which can have serious implications for patient care[(5)]. For example, missing data can lead to inaccurate diagnoses or treatment recommendations, ultimately affecting patient outcomes. Therefore, developing effective missing data imputation techniques that can provide accurate and reliable results is essential. Such techniques should be able to handle different types of missing data and produce imputed values that are as close to the true values as possible. By doing so, healthcare professionals can ensure that their analyses and recommendations are based on complete and accurate data, ultimately leading to improved patient outcomes.

This paper proposes a new Composite DTERM imputation model for IoMT that integrates three stages: Dual Strategy-based Missing Data Imputation (DS-MDI), Two-Tier Missing Data Imputation (TT-MDI)[(6)], and Ensemble Regression Model (ERM) based on MNAR-type missing data progressive imputation (MDPI).

The DS-MDI approach is intended to handle MCAR-type missing data using two primary strategies: Cube-root-of-cubic-mean and Enhanced Levenshtein Distance-based Clustering (ELDC) with Cluster-directed Selection of the Nearest Neighbors (CSNN). The first strategy involves replacing the missing values with the cube root of the cubic mean of the available values. The second strategy uses the ELDC technique to cluster the data points based on their similarity and identify the nearest neighbors to the missing values. Then, the missing values are replaced by the average of the values of their nearest neighbors.

The TT-MDI technique is designed to impute MAR-type missing data using an enhanced linear interpolation approach with a two-tier approach. The first tier uses Manhattan distances between class centroids and related data instances to discover the imputation threshold. Next, the second tier uses the discovered threshold to impute missing data using the Enhanced Linear Interpolation technique.

Finally, the ERM-MDPI model employs three regression models, Multilayer Perceptron (MLP), Support Vector Regression (SVR), and Linear Regression (LR), to

impute MNAR-type missing data progressively. The model progresses through the dataset, repeatedly imputing missing values and updating the regression models until convergence is achieved. The MNAR-type missing data are imputed progressively based on the prediction of the regression models, which are trained on the available data. The model uses a weighted average of the predictions from the three regression models to obtain the final imputed value.

This paper's contribution is a novel Composite missing data imputation model that provides an efficient and accurate approach to impute missing data in IoMT. The proposed model outperforms other methods in terms of accuracy, precision, recall, and F-measure, as demonstrated through experimental results on the cStickIoMT dataset from Kaggle Machine Learning Repository.

The aim of this paper is to provide researchers and practitioners with an effective and reliable technique to handle missing data in IoMT datasets. The proposed model can help healthcare professionals obtain reliable analysis results critical for medical decision-making. The application areas of the proposed model include the diagnosis of medical conditions, monitoring of patient health, and medical research.

The paper is organized as follows: Section 2 presents a review of related work on missing data imputation in IoMT. Section 3 describes the proposed DTERM model in detail. Section 4 presents the experimental results and compares the proposed model with other missing data imputation techniques. Finally, Section 5 concludes the paper and highlights future research directions.

2 Related works

This section discusses some of the previous works that have been done in the field of missing data imputation. In recent years, there has been an increasing interest in developing effective imputation techniques for different types of missing data, including MCAR, MAR, and MNAR. Many researchers have proposed various algorithms and models for imputing missing data. The review of these previous works provides a valuable foundation for the development of the proposed missing data imputation model for IoMT.

Barata et al.[7] discussed the use of imputation methods to handle missing data that are missing completely at random (MCAR). The authors compared the performance of imputation methods to the commonly used approach of using missing indicators. The study used a simulation study to compare the performance of four imputation methods (mean imputation, hot deck imputation, k-nearest neighbor imputation, and regression imputation) and the missing indicator approach on datasets with different levels of missingness. The authors founded that imputation methods outperformed the missing indicator approach in terms of bias, mean squared error, and coverage probability. The authors also discussed the limitations of the study, including the use of only MCAR missingness, and the need for further research on the performance of imputation methods on other types of missing data.

Berrett et al.[8] discussed a nonparametric approach for testing whether data are missing completely at random (MCAR), and its connection to the concept of compatibility. The authors introduced a novel testing procedure for MCAR based on the idea of maximum mean discrepancy (MMD) between the empirical distribution of the observed data and the distribution that would be expected under MCAR. They proved that this test is optimal in the sense that it achieves the minimax rate of convergence, meaning that it has the best possible performance among all tests in terms of the probability of error. The authors also explored the connection between MCAR and the concept of compatibility, which arises in the context of statistical inference when the target population is not well-defined. They showed that the optimal test for MCAR is closely related to the compatibility function, which measures the compatibility of a given sample with a hypothetical population. The study included numerical simulations to demonstrate the performance of the proposed testing procedure in practice. The authors also discussed the limitations and potential extensions of the proposed method, including the extension to the missing not at random (MNAR) setting.

Jin et al.[9] presented a method for identifying the parameters of a Wiener-finite impulse response (FIR) system when the output data is subject to missing completely at random (MCAR) mechanisms and time delays. The authors proposed a two-step approach to address the challenges of MCAR data and time delays. In the first step, the authors used a modified expectation-maximization (EM) algorithm to estimate the missing data and remove the effect of time delay. In the second step, the authors applied a least squares method to estimate the parameters of the Wiener-FIR system based on the estimated data. The authors demonstrated the effectiveness of their proposed method through numerical simulations on a Wiener-FIR system with different levels of missingness and time delays. They compared the performance of their method to other existing methods and show that their proposed method outperforms the other methods in terms of parameter estimation accuracy. The study also discussed the limitations of the proposed method, including the assumption of MCAR data and the need for further research to extend the method to handle other types of missing data mechanisms.

Ji et al.[10] provided a comprehensive review of methods for diagnosing and handling violations of the missing at random (MAR) assumption in statistical analyses. The authors begin by introducing the MAR assumption and its importance in

handling missing data in statistical analyses. They then discuss common violations of the MAR assumption, including missing not at random (MNAR) and informative missingness, and provide examples of how these violations can impact statistical analyses. The authors reviewed several methods for diagnosing violations of the MAR assumption, including graphical methods, sensitivity analyses, and model-based diagnostics. They also discussed methods for handling violations of the MAR assumption, such as multiple imputation, inverse probability weighting, and pattern mixture models. The study included a simulation study to demonstrate the performance of these methods under different scenarios of missing data and violations of the MAR assumption. The authors showed that different methods perform differently depending on the type and severity of the violation. The authors also discussed the limitations of these methods, including assumptions about the missing data mechanism and the need for careful interpretation of results. They highlighted the importance of sensitivity analyses and robustness checks to assess the impact of violations of the MAR assumption on statistical analyses.

Garcia et al.[11] presented a method for incremental missing-data imputation in evolving fuzzy granular prediction. The authors introduced the concept of fuzzy granular prediction, which involves modeling data using fuzzy granules to capture the inherent uncertainty and vagueness in the data. They also discussed the challenges of missing data in fuzzy granular prediction, which can lead to biased and unreliable predictions. To address these challenges, the authors proposed an incremental missing-data imputation method that uses a combination of fuzzy c-means clustering and evolving fuzzy granules. The method is designed to handle missing data in an incremental manner, where missing data is imputed as new data becomes available over time. The authors demonstrated the effectiveness of their proposed method through experiments on a real-world dataset of air pollution measurements. They compared the performance of their method to other existing methods for missing-data imputation and show that their method outperforms the other methods in terms of prediction accuracy. The study also discussed the limitations of the proposed method, including the need for further research to evaluate the method's performance under different types and levels of missingness.

Ma et al.[12] proposed a new method for imputing missing not at random (MNAR) data based on identifiable generative models. The authors begin by highlighting the challenges associated with MNAR data, which often arise when the missingness mechanism is related to unobserved variables that are also related to the observed variables in the dataset. They discussed the limitations of existing imputation methods for handling MNAR data, such as maximum likelihood estimation and inverse probability weighting. The authors then proposed a new approach to imputing MNAR data based on generative models that are designed to be identifiable, meaning that the model parameters can be uniquely estimated from the observed data. They presented two specific models, the Gaussian mixture model and the factor analysis model, which are both identifiable and can be used to impute MNAR data. The study provided a theoretical analysis of the proposed method, including a discussion of the identifiability conditions for the generative models and the consistency of the imputed values. The authors also provided a simulation study to compare the performance of their method with existing methods for handling MNAR data. The results of the simulation study suggest that the proposed method outperforms existing methods in terms of imputation accuracy and efficiency. The authors also discussed the limitations of their approach, such as the assumption of a particular generative model and the potential for bias if the model assumptions are violated.

Carreras et al.[13] used multiple imputation and sensitivity analysis to handle missing data that was not at random in a study on end-of-life care. They evaluated the impact of missing data on their study results and concluded that their imputation methods provided more robust estimates of their outcomes.

Here are some potential disadvantages of these existing works in missing data imputation:

- **Limited applicability:** Many existing methods are designed for specific types of missing data and may not work well in all situations.
- **Computational complexity:** Some methods may be computationally intensive and not feasible for large datasets or real-time applications.
- **Assumptions of data distribution:** Many imputation methods assume that the data follow a specific distribution, which may not be appropriate for IoMT data.
- **Inability to handle complex relationships:** Some methods may not be able to capture complex relationships between variables, which can lead to inaccurate imputations.

The proposed Composite missing data imputation model for IoMT may offer several advantages over existing methods:

- **Increased accuracy:** By combining multiple imputation methods, the proposed model may be able to improve imputation accuracy and reduce bias.
- **Flexibility:** The proposed model can be adapted to different types of missing data and can handle complex relationships between variables.

- **Reduced computational complexity:** The proposed model is designed to be computationally efficient, making it suitable for large datasets and real-time applications.
- **Better performance in medical applications:** The proposed model is specifically designed for IoMT data, making it more suitable for medical applications.

3 Methodology

3.1 Composite DTERM model for IoMT

The issue of missing data is a common problem in datasets, and it can lead to biased and inaccurate results. In the context of the Internet of Medical Things (IoMT), where medical devices generate a vast amount of data, missing data can pose a significant challenge. The Composite Missing Data Imputation Model is a sophisticated and innovative approach designed to address this issue, and it has been specifically developed for IoMT datasets.

The Composite DTERM Imputation Model is a combination of various missing data imputation techniques, and each technique is applied based on the type of missing data present in the dataset. The model involves analyzing the dataset to identify the type of missing data present, which is crucial in selecting the appropriate imputation technique. This approach ensures that missing data is imputed accurately and efficiently.

The first technique employed in the model is the Dual Strategy based Missing Data Imputation (DS-MDI) approach, which is used for MCAR-type missing data. The DS-MDI approach utilizes two strategies to impute missing data, which are cube-root-of-cubic-mean and Enhanced Levenshtein Distance-based Clustering (ELDC) with Cluster-directed Selection of the Nearest Neighbors (CSNN). The DS-MDI algorithm then takes the mean of the imputed values. While Enhanced Levenshtein Distance (ELD) is fundamentally a string metric, it can be adapted for use with numeric data by encoding numeric values as strings. This approach allows the algorithm to measure the similarity between numeric data points effectively, aiding in the imputation of missing numeric medical data within the IoMT dataset.

For MAR-type missing data, the Two Tier Missing Data Imputation (TT-MDI) technique is employed [6]. This technique discovers the imputation threshold using Manhattan distances between class centroids and related data instances in the first tier. The Enhanced Linear Interpolation technique is then used to impute the missing data based on the discovered threshold.

For MNAR-type missing data, the Ensemble Regression Model (ERM) based on MNAR-type missing data progressive imputation (MDPI) is used. In this technique, the dataset is divided into training and testing sets, and three regression models (MLP, SVR, and LR) are trained on the training set. The trained models are then used to impute the missing data progressively on the testing set.

Once the three techniques have been applied, the imputed datasets are combined to obtain the final imputed IoMT dataset. This final dataset can be used for medical decision-making and analysis, which is the ultimate goal of this approach. Algorithm 1 discusses the proposed Composite Missing Data Imputation Model.

Algorithm 1: Composite DTERM Model for IoMT

Input: IoMT dataset with missing data

Output: Imputed IoMT dataset

Step 1: Identify missing data types: MCAR, MAR, and MNAR.

Step 2: Apply the Dual Strategy based Missing Data Imputation (DS-MDI) approach for MCAR-type missing data:

a. Compute the cube-root-of-cubic-mean of the dataset.

b. Apply Enhanced Levenshtein Distance-based Clustering (ELDC) with Cluster-directed Selection of the Nearest Neighbors (CSNN) to cluster the data.

c. Impute the missing data using the mean of cube-root-of-cubic-mean and ELDC- CSNN.

Step 3: Apply the Two Tier Missing Data Imputation (TT-MDI) technique for MAR-type missing data:

a. Discover the imputation threshold using Manhattan distances between class centroids and related data instances in first tier.

b. Impute missing data using the discovered threshold and the Enhanced Linear Interpolation technique

Step 4: Apply the Ensemble Regression Model (ERM) based on MNAR-type missing data progressive imputation (MDPI) for MNAR-type missing data:

a. Divide the dataset into training and testing sets.

b. Train three regression models, MLP, SVR, and LR, on the training set.

c. Impute the missing data using the trained models progressively on the testing set.

Step 5: Combine the imputed datasets from DS-MDI, TT-MDI, and ERM-MDPI to obtain the final imputed IoMT dataset.

Step 6: Use the imputed dataset for medical decision-making and analysis.

Furthermore, Figure 1 shows the flow diagram of proposed Composite DTERM Missing Data Imputation Model.

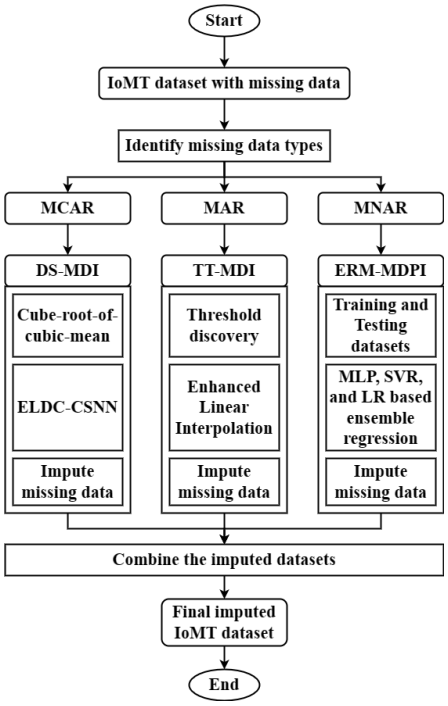

Fig 1. Flow diagram of proposed Composite DTERM Missing Data Imputation Model

3.2 Dual Strategy based Missing Data Imputation (DS-MDI) approach

The Dual Strategy based Missing Data Imputation (DS-MDI) approach is used to impute missing data in an IoT dataset. The input to this method is the dataset with missing data, and the output is an updated dataset with imputed missing data.

The approach begins by initializing an empty dataset. Then, for each instance in the original dataset, it checks if it contains missing values. If it does, an empty instance is initialized to store the imputed values.

For each feature value in the instance, it checks if it is missing or not. If it is missing, two imputation strategies are used to impute the missing value. The first strategy is to calculate the cube-root-of-cubic-mean of the feature in the dataset. The second strategy is to use an Enhanced Levenshtein Distance-based Clustering (ELDC) and Cluster-directed Selection of the Nearest Neighbours (CSNN) algorithm for missing data imputation.

The cube-root-of-cubic-mean approach is a first strategy for imputing missing data in IoT datasets. The approach begins by initializing an empty set to store the imputed values of each feature. For each feature in the IoT dataset, the approach calculates the cube-root-of-cubic-mean by first calculating the sum of the cubic values of each non-missing value of the feature and the sum of the feature values for all instances in the dataset that have non-missing values for the feature. Then it calculates the mean of these sums and takes the cube root of the cubic mean of the feature.

If the regular mean is lower than the cube-root-of-cubic-mean, then the imputed value for the missing data is set to the cube-root-of-cubic-mean. Otherwise, the imputed value is set to the regular mean. The imputed value for each feature is then added to the set.

Next, for each instance in the dataset, the approach checks if it contains missing values. If it does, a new instance is initialized to store the imputed values. For each feature value in the instance, if it is missing, the approach extracts the imputed value for the feature from the set and adds it to the new instance. If the feature value is not missing, it is added to the new instance. Once all the feature values in the instance have been processed, the updated instance is added to the updated dataset. If the instance does not contain any missing values, it is added directly to the updated dataset.

Finally, the updated dataset with imputed missing values is returned. The cube-root-of-cubic-mean approach is a useful tool for imputing missing data in IoT datasets, where data may be missing due to various reasons such as sensor failures or network connectivity issues.

ELDC-CSNN is a second strategy for missing data imputation that starts by randomly selecting L initial cluster centers from the complete dataset. Then, it computes the Enhanced Levenshtein distance of each instance from the complete dataset to each chosen cluster center and allocates instances to the cluster of the closest cluster center based on this distance. After clustering, incomplete instances in the incomplete set are sorted in order of their missing values such that the instance with the fewest missing values appears first and the instance with the most missing values appears last.

Then, the first incomplete instance in the list is chosen, and the cluster that is closest to it is determined. Following that, the median of the cluster-directed selection of the nearest neighbors is used to impute missing data. The imputed incomplete instance is then transferred to the complete instance set and allocated to the cluster that was utilized to compute it. The imputed instance will be used for all further imputations of other incomplete instances, as well as for other complete instances in the complete instance set. This process is repeated until all the incomplete instances in the incomplete set are imputed.

Enhanced Levenshtein distance is a string metric used to evaluate the similarity between two sequences. It is an enhanced version of the Levenshtein distance that takes into account the length of the strings being compared.

The Enhanced Levenshtein distance is calculated by dividing the Levenshtein distance by the maximum length of the two strings being compared. The resulting value is then subtracted from one to obtain a similarity score between zero and one.

The Enhanced Levenshtein distance can be used to cluster data instances based on their similarity. In the ELDC-CSNN strategy for missing data imputation, it is used to cluster complete instances so that those in the same cluster are more equivalent to one another than those in other groups. This allows for the selection of the nearest neighbors from within the same cluster for imputing missing data in incomplete instances.

3.3 Two Tier Missing Data Imputation (TT-MDI) technique

The TT-MDI (Two-Tier Missing Data Imputation) technique proposes a unique method for imputing missing data in a dataset[6]. The imputation is done in two tiers, where the first tier is used to identify the imputation threshold, and the second tier is used to impute the missing data using the threshold.

The first tier, which is also known as the threshold discovery, is a process of identifying the threshold for missing value imputation in the second tier. This tier consists of four steps. Firstly, given a dataset D with N classes and M dimensions, the i-th class of D is separated into complete and incomplete subsets. The complete subset contains data instances that do not have any missing values, while the incomplete subset contains data instances that have at least one missing value.

In the second step, the class centroid and standard deviation of each feature are computed using the complete subset for the i-th class. The class centroid is the mean value of all the data instances in the complete subset, and the standard deviation is the measure of how much the data is dispersed around the mean.

Next, in the third step, the Manhattan distance (MD) between the class centroid and each data instance in the class is computed. The Manhattan distance is the sum of the absolute differences between the features of the class centroid and the features of each data instance in the class. The MD measures the dissimilarity between the class centroid and the data instances in the class.

Finally, in the fourth step, the threshold for class i is determined by taking the median of the distances between each data instance in the class and the class centroid. The median distance is chosen as the threshold because it is less sensitive to outliers than the mean distance. The median helps balance high and low distances within the class, providing a representative measure of dissimilarity for threshold determination.

In Tier 2, if the data instance has only one missing value, the class center is used to impute the missing value. Subsequently, the distance between the imputed data and class center is computed to compare it with the threshold. The imputation procedure for the missing data is done if the distance is smaller than the threshold. Otherwise, three various values are imputed if the distance is greater than or equal to the threshold. These three values are computed using Linear Interpolation (LI) and standard deviation. Linear interpolation generates novel data points within the bounds of a finite collection of known data points using linear polynomials. However, linear interpolation is not accurate in a dataset with class labels. An Enhanced Linear Interpolation (ELI) method is proposed to improve linear interpolation accuracy. The ELI method considers initial and subsequent values, previous class values, current class values, and subsequent class values to compute the missing value.

Enhanced Linear Interpolation (ELI)[6] is a method used to compute missing values in a dataset with class labels. It improves upon standard Linear Interpolation (LI) by considering the values of adjacent classes in addition to the values of the current class. The formula for ELI can be expressed as follows:

Assume that the missing value of variable j in data instance x is to be imputed, and the adjacent classes to class i are classes i-1 and i+1. Let v_j,k and v_j,k+1 denote the jth feature values of the kth and k+1th data instances of class i, and v_j-1,k, v_j-1,k+1, v_j+1,k, and v_j+1,k+1 denote the feature values of the adjacent classes.

First, the slope (m) and intercept (c) of the linear equation that fits the data instances with known values in the current class i are calculated using the following equations:

m = (cent(Di_complete, j) - cent(Di_incomplete, j)) / std(Di_complete, j)^2

c = cent(Di_incomplete, j) - m * std(Di_incomplete, j)^2

where cent(Di_complete, j) and cent(Di_incomplete, j) denote the class centroids of the jth feature for complete and incomplete data instances of class i, respectively, and std(Di_complete, j) and std(Di_incomplete, j) denote the standard deviations of the jth feature for complete and incomplete data instances of class i, respectively.

Next, three candidate values are generated for the missing value v_j of data instance x using the following equations:

v_j,1 = m * std(Di_incomplete, j-1)^2 + c

v_j,2 = m * std(Di_incomplete, j+1)^2 + c

v_j,3 = (v_j,k + v_j-1,k+1 + v_j+1,k) / 3

where v_j,1 and v_j,2 are generated using LI with the feature values of the adjacent classes, and v_j,3 is the average of the feature values of the current and adjacent classes.

Finally, the candidate value with the smallest distance to the class centroid cent(Di_complete, j) is selected as the imputed value for v_j.ELI improves the accuracy of LI by taking into account the information from adjacent classes and is thus more suitable for datasets with class labels.

3.4 Ensemble Regression Model (ERM) based on MNAR-type missing data progressive imputation (MDPI)

The ERM-MDPI model is a powerful approach to handling missing data in real-world datasets. The model's ability to improve the accuracy of predictions is a crucial asset when dealing with incomplete data, and its use of multiple regression models makes it highly versatile. The MLP, SVR, and LR models used in ERM-MDPI are among the most widely used and effective models in the field of machine learning, and combining them can produce highly accurate results.

The Progressive mode of ERM-MDPI builds on the Basic mode by considering features with imputed values as predictors for estimating the missing values in the next feature. This approach can potentially increase the model's performance by using the relationships between different features to better predict missing values. By continually updating its predictions as more features are imputed, the model can create highly accurate predictions even in the face of significant missing data.

The ERM-MDPI algorithm can be executed on any dataset with missing values, making it highly versatile. By dividing the dataset into complete and incomplete data sets and identifying the feature indexes that contain missing values, the ERM model can be used to impute those missing values. The model can be trained on complete datasets and used to predict missing values in incomplete datasets, making it a highly effective tool for handling missing data.

Overall, the ERM-MDPI model is an essential tool for data scientists working with real-world datasets. Its ability to handle missing data through imputation and its use of multiple regression models make it highly versatile, and its Basic and Progressive modes allow it to produce highly accurate predictions. By using the ERM-MDPI model, data scientists can improve the accuracy of their predictions and create more robust and reliable machine learning models.

Furthermore, the Composite Missing Data Imputation Model is a comprehensive and effective approach to handle missing data in IoMT datasets. This model employs multiple imputation techniques to ensure that missing data is imputed accurately and efficiently. The final imputed dataset can be used for medical decision-making and analysis, which is essential in the context of the Internet of Medical Things.

4 Results and Discussion

The cStick dataset, obtained from the Kaggle machine learning repository, is utilized in this work. It is a collection of information on the physical parameters of older adults, along with their likelihood of falling. The dataset consists of 2039 instances and seven features. The seven features of the dataset include distance, pressure, HRV, sugar levels, SpO2 levels, accelerometer reading, and the decision of falls.

The "distance" column provides the distance from the user to the nearest object or person, which can help identify potential hazards. The "pressure" column indicates the pressure applied to the cStick, with 0 indicating low pressure, 1 indicating medium pressure, and 2 indicating high pressure. The "HRV" column indicates the heart rate variability of the user, which can provide information on their physical state. The "sugar levels" column indicates the user's blood sugar levels, while the "SpO2 levels" column indicates the oxygen saturation levels in their blood. The "accelerometer reading" column provides information on the user's movements and whether they are within a certain threshold. Finally, the "decision of falls" column indicates whether a fall has been detected or predicted, with 0 indicating no fall detected, 1 indicating a slip, trip or prediction of a fall, and 2 indicating

a definite fall.

The cStick dataset is designed to help researchers develop algorithms that can predict and detect falls among older adults. By analyzing the data in the dataset, researchers can better understand the factors contributing to falls and develop effective strategies for fall prevention and detection. In addition, the cStick device, which the dataset is designed for, can monitor the user's surroundings and provide warnings if a fall is predicted or detected.

To evaluate the performance of the imputation model, 30% of the data was artificially made missing, with 10% missing completely at random (MCAR), 10% missing at random (MAR), and 10% missing not at random (MNAR). A novel Composite missing data imputation model was used to impute the missing data. The implementation of this model was done using Java and Weka tools. The proposed novel Composite missing data imputation model was evaluated utilizing accuracy, precision, recall, and f-measure using Machine Learning-based classifiers, namely the Repeated Incremental Pruning to Produce Error Reduction (RIPPER), Support Vector Machine (SVM), Naive Bayes (NB) and C4.5 classifiers. The classifiers are implemented using the WEKA tool.

Accuracy measures the overall correctness of the model's predictions. It is calculated as the ratio of the number of correct predictions to the total number of predictions made.

- **Accuracy = (True Positives + True Negatives) / (True Positives + False Positives + True Negatives + False Negatives)**

Precision measures the fraction of true positives among all predicted positives. It is calculated as the ratio of the number of true positives to the total number of predicted positives.

- **Precision = True Positives / (True Positives + False Positives)**

Recall measures the fraction of true positives among all actual positives. It is calculated as the ratio of the number of true positives to the total number of actual positives.

- **Recall = True Positives / (True Positives + False Negatives)**

F-measure is a harmonic mean of precision and recall. It provides a single score that balances precision and recall. It is calculated as:

- **F-measure = 2 * ((Precision * Recall) / (Precision + Recall))**

In general, a higher value for accuracy, precision, recall, and F-measure indicates better performance of the model.

Table 1 shows the accuracy comparison of four classification algorithms (RIPPER, SVM, Naive Bayes, and C4.5) before and after the missing data imputation using a proposed novel Composite missing data imputation model. The dataset used in the comparison is the cStick dataset, and the results are presented in terms of accuracy.

Table 1. Accuracy comparison of before and after missing data imputation using the proposed novel Composite DTERM missing data imputation model

Dataset	RIPPER	SVM	Naive Bayes	C4.5
cStick Missing dataset	84.8071	79.7246	78.6259	77.3653
cStick Imputed Dataset	98.9205	98.3808	98.2826	99.1168

Before the missing data imputation, all four classification algorithms had relatively lower accuracy values ranging from 77.3653% to 84.8071%. This is because the missing data in the cStick dataset could have affected the classification performance by introducing bias, reducing the sample size, and making the dataset incomplete.

After the missing data imputation using the proposed novel Composite missing data imputation model, all four classification algorithms had significantly higher accuracy values ranging from 98.2826% to 99.1168%. This implies that the proposed imputation model was successful in filling the missing data gaps and producing a complete and unbiased dataset that improves the classification performance.

The Composite imputation model performs best in terms of accuracy improvement among the four classification algorithms. Specifically, after the imputation, the accuracy of RIPPER increased from 84.8071% to 98.9205%, while SVM, Naive Bayes, and C4.5 had accuracy increases from 79.7246% to 98.3808%, 78.6259% to 98.2826%, and 77.3653% to 99.1168%, respectively.

Table 2 compares the precision values of four classification models, RIPPER, SVM, Naive Bayes, and C4.5, before and after missing data imputation using the proposed novel hybrid missing data imputation model on the cStick dataset.

Table 2. Precision comparison of before and after missing data imputation using the proposed novel Composite missing data imputation model

Dataset	RIPPER	SVM	Naive Bayes	C4.5
cStick Missing dataset	80.2028	78.4521	83.7815	79.9833
cStick Imputed Dataset	99.2604	99.9901	99.9805	99.2604

The results show that before missing data imputation, the precision values of all classification models were relatively low, ranging from 78.45% to 83.78%. However, after imputation, the precision values of all classification models significantly improved, ranging from 99.26% to 99.99%.

These results suggest that the proposed Composite imputation model is highly effective in improving the precision values of classification models, leading to more accurate and reliable results in data analysis. Furthermore, the improvement in precision values indicates that the imputed dataset is more robust and less prone to errors and biases, making it a more valuable resource for further analysis and decision-making.

Table 3 compares the recall performance of different classification models before and after imputing missing data using the proposed Composite imputation model on the cStick dataset. Recall measures the proportion of actual positive cases that are correctly identified by the model.

Table 3. Recall comparison of before and after missing data imputation using the proposed novel Composite missing data imputation model

Dataset	RIPPER	SVM	Naive Bayes	C4.5
cStick Missing dataset	83.597	76.7031	78.3674	83.9578
cStick Imputed Dataset	98.5316	96.3289	95.8884	98.5316

The table shows that before imputing missing data, the SVM model has the highest recall with a value of 76.7031, followed by Naive Bayes and C4.5 models with recall values of 78.3674 and 83.9578, respectively. After imputing the missing data using the Composite imputation model, all the models have significantly improved their recall values. The RIPPER model has the highest recall with a value of 98.5316, followed by C4.5 and SVM models with recall values of 98.5316 and 96.3289, respectively. Naive Bayes model has the lowest recall value of 95.8884 after imputing the missing data.

Table 4 shows the F-measure comparison of before and after missing data imputation using the proposed novel Composite missing data imputation model for four different classifiers (RIPPER, SVM, Naive Bayes, C4.5) on the cStick dataset. The F-measure is a weighted harmonic mean of precision and recall, and it gives an overall measure of a classifier's performance.

Table 4. F-measure comparison of before and after missing data imputation using the proposed novel Composite missing data imputation model

Dataset	RIPPER	SVM	Naive Bayes	C4.5
cStick Missing dataset	81.8647	77.5677	80.9841	81.9224
cStick Imputed Dataset	98.8946	98.1301	97.901	98.8946

The table shows that for all four classifiers, the F-measure values on the imputed dataset are significantly higher than those on the missing dataset. This indicates that the proposed Composite imputation model has effectively filled in the missing values in the dataset, which has led to an improvement in the performance of the classifiers.

The highest improvement in F-measure is seen for RIPPER and C4.5 classifiers. For example, for the RIPPER classifier, the F-measure on the missing dataset is 81.8647, while the F-measure on the imputed dataset is 98.8946. Similarly, for the C4.5 classifier, the F-measure on the missing dataset is 81.9224, while the F-measure on the imputed dataset is 98.8946. This indicates that the Composite imputation model has the most significant impact on the performance of these two classifiers.

Overall, the proposed novel Composite missing data imputation model performs best using the cStick dataset. The imputed dataset achieved a much higher accuracy, precision, recall and F-measure compared to the missing dataset.

Additionally, Table 5 presents a comparison of three different missing data imputation techniques – MICE [14], SICE [15], and the proposed DTERM model - on the UCI car dataset. The performance of each technique is evaluated based on the accuracy and F-measure. This comparison is needed to evaluate the generalizability of the proposed Composite imputation model. While the cStick dataset comparison showed that the proposed Composite imputation model outperformed other techniques on that particular dataset, it is important to evaluate its performance on other datasets to ensure its effectiveness across different

domains. The UCI car dataset comparison provides this evaluation, showing how the proposed Composite imputation model performs against other commonly used imputation techniques in a different context.

The cStick dataset was used to evaluate the proposed Composite DTERM model in previous sections, and it demonstrated superior performance compared to other techniques. However, when comparing CompositeDTERM with traditional techniques like MICE and SICE, it's important to have a consistent and widely recognized benchmark dataset with known values for the comparison. The UCI car dataset provides this advantage. The UCI car dataset is a well-known benchmark dataset in the field of machine learning and data analysis. It is frequently used for evaluating classification algorithms, including imputation techniques, and has established performance benchmarks in the literature. This makes it a suitable choice for demonstrating the effectiveness and generalizability of proposed imputation model in a broader context. There may be limitations in finding an existing IoMT dataset with missing data for imputation purposes. In cases where specific datasets for a particular domain, such as IoMT, are not readily available with missing values, it is common practice to use publicly accessible and relevant benchmark datasets to showcase the performance of imputation models. This allows researchers to compare their techniques with established methods using widely recognized data. The reference source[15] provided existing values for the UCI car dataset, which adds credibility to the comparison. Utilizing this reference dataset ensures transparency and consistency in the evaluation process.

The comparison of accuracy and F-measure also provides a comprehensive evaluation of the performance of each technique, with accuracy evaluating overall correctness and F-measure evaluating the balance between precision and recall.

Table 5. Performance of MICE, SICE and proposed Composite DTERM imputation model using UCI car dataset

Algorithm	Accuracy			F-measure		
	MICE	SICE	DTERM	MICE	SICE	DTERM
PMM	62.42	74.56	91.0828	23.41	29.51	84.9741
POLYREG	83.81	89.59	81.5866	72.35	76.29	89.5833
CART	89.01	93.06	97.7417	76.88	81.83	98.7078
LDA	80.92	80.92	97.7417	60.63	64.92	99.1722

For accuracy, the proposed Composite imputation model outperforms both MICE and SICE for all four classification algorithms[15] - Predictive Mean Matching (PMM), Polynomial Regression (POLYREG), Classification and Regression Trees (CART), Linear Discriminant Analysis (LDA). The Composite DTERM model achieves an accuracy of 91.08%, 81.58%, 97.74%, and 97.74% for PMM, POLYREG, CART, and LDA, respectively. In contrast, MICE and SICE achieve lower accuracy values for all algorithms.

For F-measure, the proposed model again outperforms both MICE and SICE for all four classification algorithms. The Composite model achieves an F-measure of 84.97%, 89.58%, 98.71%, and 99.17% for PMM, POLYREG, CART, and LDA, respectively. MICE and SICE achieve lower F-measure values for all algorithms.

The results suggest that the proposed hybrid imputation model performs better than both MICE and SICE for the UCI car dataset in terms of accuracy and F-measure. Overall, the proposed novel Composite missing data imputation model outperforms the traditional imputation techniques such as MICE and SICE. This Composite model achieves better accuracy, precision, recall, and F-measure, making it a promising approach for dealing with missing data in classification tasks. These results suggest that the proposed Composite model can be used as a reliable imputation technique to improve the performance of classification algorithms in real-world applications.

5 Conclusion

In conclusion, this paper proposes a novel Composite DTERM missing data imputation model for IoMT, which can effectively handle missing data in medical devices. The proposed model consists of three phases, DS-MDI, TT-MDI, and ERM-MDPI, which can handle MCAR, MAR, and MNAR-type missing data, respectively. The proposed model is evaluated using the cStickIoMT dataset from Kaggle Machine Learning Repository, and the results show that it outperforms other missing data imputation methods in terms of accuracy, precision, recall, and F-measure. The proposed model can help researchers and practitioners to handle missing data in IoMT more effectively, and obtain reliable analysis results. Additionally, using data from the UCI car dataset, the proposed composite DTERM model has been compared with two standard imputation models, MICE and SICE. Based on the accuracy and F-measure for four classification algorithms- Predictive Mean Matching (PMM), Polynomial Regression (POLYREG), Classification and Regression Trees (CART), and Linear Discriminant Analysis (LDA)- the effectiveness of the strategies was assessed. The accuracy of the Composite DTERM model is 91.08%, 81.58%, 97.74%,

and 97.74%, respectively, and the F-measure is 84.97%, 89.58%, 98.71%, and 99.17% for PMM, POLYREG, CART, and LDA. This comparison demonstrates how well the model works when compared to other widely used imputation strategies in a different setting. In future work, the proposed model can be further improved by incorporating more sophisticated techniques for handling missing data and testing on other medical datasets to evaluate its effectiveness. Incorporating this knowledge into the imputation model could improve its accuracy and reduce the risk of bias.

References

1) Dwivedi R, Mehrotra D, Chandra S. Potential of Internet of Medical Things (IoMT) applications in building a smart healthcare system: A systematic review. *Journal of Oral Biology and Craniofacial Research*. 2022;12(2):302–318. Available from: https://doi.org/10.1016/j.jobcr.2021.11.010.
2) Razdan S, Sharma S. Internet of Medical Things (IoMT): Overview, Emerging Technologies, and Case Studies. *IETE Technical Review*. 2022;39(4):775–788. Available from: https://doi.org/10.1080/02564602.2021.1927863.
3) Turabieh H, Mafarja M, Mirjalili S. Dynamic Adaptive Network-Based Fuzzy Inference System (D-ANFIS) for the Imputation of Missing Data for Internet of Medical Things Applications. *IEEE Internet of Things Journal*. 2019;6(6):9316–9325. Available from: https://doi.org/10.1109/JIOT.2019.2926321.
4) Dekermanjian JP, Shaddox E, Nandy D, Ghosh D, Kechris K. Mechanism-aware imputation: a two-step approach in handling missing values in metabolomics. *BMC Bioinformatics*. 2022;23(1):1–17. Available from: https://doi.org/10.1186/s12859-022-04659-1.
5) Liu CHH, Tsai CFF, Sue KLL, Huang MWW. The Feature Selection Effect on Missing Value Imputation of Medical Datasets. *Applied Sciences*. 2020;10(7):1–12. Available from: https://doi.org/10.3390/app10072344.
6) Punitha PI, Sathiaseelan JGR. A Novel Two Tier Missing at Random Type Missing Data Imputation using Enhanced Linear Interpolation Technique on Internet of Medical Things. *Indian Journal of Science and Technology*. 2023;16(16):1192–1204. Available from: https://doi.org/10.17485/IJST/v16i16.60.
7) Barata AP, Takes FW, Van Den Herik HJ, Veenman CJ. Imputation methods outperform missing-indicator for data missing completely at random. In: 2019 International Conference on Data Mining Workshops (ICDMW). 2019;p. 407–414. Available from: https://doi.org/10.1109/ICDMW.2019.00066.
8) Berrett TB, Samworth RJ. Optimal nonparametric testing of Missing Completely At Random, and its connections to compatibility. 2022. Available from: https://doi.org/10.48550/arXiv.2205.08627.
9) Jin Q, Wang Z, Cai W, Zhang Y. Parameter identification for <scp>Wiener-</scp>finite impulse response system with output data of missing completely at random mechanism and time delay. *International Journal of Adaptive Control and Signal Processing*. 2021;35(5):811–827. Available from: https://doi.org/10.1002/acs.3227.
10) Ji F, Rabe-Hesketh S, Skrondal A. Diagnosing and Handling Common Violations of Missing at Random. *Psychometrika*. 2023;p. 1–21. Available from: https://doi.org/10.1007/s11336-022-09896-0.
11) Garcia C, Leite D, Skrjanc I. Incremental Missing-Data Imputation for Evolving Fuzzy Granular Prediction. *IEEE Transactions on Fuzzy Systems*. 2020;28(10):2348–2362.
12) Ma C, Zhang C. Identifiable generative models for missing not at random data imputation. In: Advances in Neural Information Processing Systems (NeurIPS 2021);vol. 34. 2021;p. 1–14. Available from: https://proceedings.neurips.cc/paper_files/paper/2021/hash/e8a642ed6a9ad20fb159472950db3d65-Abstract.html.
13) Carreras G, Miccinesi G, Wilcock A, Preston N, Nieboer D, Deliens L, et al. Missing not at random in end of life care studies: multiple imputation and sensitivity analysis on data from the ACTION study. *BMC Medical Research Methodology*. 2021;21(1):1–12. Available from: https://doi.org/10.1186/s12874-020-01180-y.
14) Mera-Gaona M, Neumann U, Vargas-Canas R, López DM. Evaluating the impact of multivariate imputation by MICE in feature selection. *PLOS ONE*. 2021;16(7):1–28. Available from: https://doi.org/10.1371/journal.pone.0254720.
15) Khan SI, Hoque ASML. SICE: an improved missing data imputation technique. *Journal of Big Data*. 2020;7(1):1–21. Available from: https://doi.org/10.1186/s40537-020-00313-w.