# A Comprehensive Survey on Imputation of Missing Data in Internet of Things

DEEPAK ADHIKARI, WEI JIANG, JINYU ZHAN, and ZHIYUAN HE, University of Electronic Science and Technology of China, China
DANDA B. RAWAT, Howard University, USA
UWE AICKELIN and HADI A. KHORSHIDI, Melbourne University, Australia



The Internet of Things (IoT) is enabled by the latest developments in smart sensors, communication technologies, and Internet protocols with broad applications. Collecting data from IoT and generating information from these data become tedious tasks in real-life applications when missing data are encountered in datasets. It is of critical importance to deal with the missing data timely for intelligent decision-making. Hence, this survey attempts to provide a structured and comprehensive overview of the research on the imputation of incomplete data in IoT. The article starts by providing an overview of incomplete data based on the architecture of IoT. Then, it discusses the various strategies to handle the missing data, the assumptions used, the computing platform, and the issues related to them. The article also explores the application of imputation in the area of IoT. We encourage researchers and data analysts to use known imputation techniques and discuss various issues and challenges. Finally, potential future directions regarding the method are suggested. We believe this survey will provide a better understanding of the research of incomplete data and serve as a guide for future research.

CCS Concepts: • **Information systems** → **Data cleaning**; **Computing platforms**; **Data cleaning**; • **Computing methodologies** → **Machine learning approaches**; • **Networks** → **Network architectures;**;

Additional Key Words and Phrases: Imputation of missing data, multiple imputations, machine learning, deep learning, computing platform for incomplete data, Internet of Things

**ACM Reference format:**
Deepak Adhikari, Wei Jiang, Jinyu Zhan, Zhiyuan He, Danda B. Rawat, Uwe Aickelin, and Hadi A. Khorshidi. 2022. A Comprehensive Survey on Imputation of Missing Data in Internet of Things. *ACM Comput. Surv.* 55, 7, Article 133 (December 2022), 38 pages.
https://doi.org/10.1145/3533381

## 1 INTRODUCTION

Industry and academia have directed their attention toward the **Internet of Things (IoT)**, generating tremendous data by embedding billions of physical components to the Internet at an unprecedented rate. The Internet connects actuators and sensors enclosed in physical objects

Authors' addresses: D. Adhikari, W. Jiang (corresponding author), J. Zhan, and Z. He, University of Electronic Science and Technology of China, No. 4, Section2, North Jianshe Road, Chengdu, China; emails: deepakadhikari7@hotmail.com, weijiang@uestc.edu.cn, zhanjy@uestc.edu.cn, allenhzy3@126.com; D. B. Rawat, Howard University, 2300 6th Street, NW, Washington D.C., 20059, USA; emails: db.rawat@ieee.org; U. Aickelin and H. A. Khorshidi, Melbourne University, Melbourne, Victoria, Australia; emails: uwe.aickelin@unimelb.edu.au, hadi.khorshidi@unimelb.edu.au.

0360-0300/2022/12-ART133 $15.00
https://doi.org/10.1145/3533381

allowing them to see, think, hear, execute tasks together, share information, and coordinate decisions [5]. With the development of intelligent sensing devices and IoT in data generation and collection technology, various data are generated and collected from different sources, such as sensors, web, high definition cameras, videos, and surveys, and are transmitted to their destinations. Given the multiple sensing devices in different situations, various types of data have different semantics, shape (space and time), and format flows in the system and get collected [102]. Sensor systems are crucial in modern networked digital infrastructures, such as smart cities, environmental monitoring, industrial automation, autonomous vehicles, intelligent city, and building [44]. Multiple heterogeneous sensors are typically integrated to build a system where nearby sensing devices communicate and transfer data to cloud infrastructure for further analysis. All essential decisions on IoT services and applications rely on the available data generated/collected prior to the decision. Such decision-making processes use predictive models relying upon the available or observed data to increase the reliability, efficiency, profitability, and performance of IoT applications. This helps to create business models, improves the business process, and reduces risk and cost and decision-making through data visualization [78]. However, regardless of the unquestionable benefits of IoT, numerous issues and challenges need to be handled. Among them, unreliable outcomes generated by missing values is one of them. Missing data or value is the absence of the data value in the variable of an observation. No matter how strictly the various intelligent systems are designed to collect data from high-quality sensors or how hard investigators try to prevent them, missingness occurs for various reasons in the IoT domain with respect to health, environmental, and traffic monitoring, specificially in a long duration screening, data collection and transmission failures, machine failure (the failure or dysfunctioning of the sensors providing information), devices running out of battery or power, and boundary specification problems (task of specifying inclusion rules for relations in a network study) [2, 67, 145]. When incomplete data are not handled properly, they result in inaccurate and unreliable analysis during the decision-making procedure.

### 1.1 Essence of Complete Data

Datasets are the source of appropriate information for different types of knowledge, such as classification, pattern, trend, and analysis [159]. However, given the missing values in certain datasets, such models break down, preventing the generation of the required information for intelligent decision-making. The existence of missing observations are frequently encountered in IoT research and studies such as biometric system [37], machine translation [97], intelligent transportation systems [29], Internet of things [44, 102], big data [87, 151], sensor network [161], environmental monitoring [2], Internet of medical things [62, 142], industrial database [43], credit system [68], finance [104], safety [133], physical activity [135], cybersecurity [164], power system [66, 148], and social network [65].

Genuine issues, including insufficient information, and broken data structure, arise when experts and researchers need every piece of information from the data containing the missing data. Ignoring or failure in the handling of missing data causes serious problems. The high risk of obtaining bias results due to the difference between complete and missing data, complications in handling and analyzing the data, reduction in sample size, loss of efficiency, the precision of confidence intervals are harmed, reduction of analytical power, and occurrance of misleading results when the researcher uses the missing data procedures without considering the assumptions [52, 75, 114].

Most decision-making techniques use intelligent systems, including **Machine Learning (ML)** and **Deep Learning (DL)**, to analyze. All of these techniques require complete quality data for better estimation. Due to the recurring nature of missing data even in a well-designed study and highly sensitive intelligent devices, complete and quality data for intelligent analysis and decision-making becomes a crucial unsolved component.

Handling incomplete data is particularly a tedious task, requiring prudent inspection to determine the mechanism, pattern, and nature of the data. An approach used frequently for analysis of incomplete data is imputation. The imputation procedure is based on the basic principle of replacement or substitution [117]. The new reliable data are imputed (reinserted) through the principled imputation techniques, which replace the missing value indirectly. The observed data combine available information with statistical conditions to estimate the missing mechanism and the population parameter.

In IoT applications, sensing devices are true sources of streaming big data, implying that the imputation of incomplete data at the edge might be a powerful tool for dealing with the inevitable bottlenecks in data communication. From the deluge of sensed raw data, the challenges lie in extracting valuable information. In such a scenario, the objective would be to reduce the amount of sensor data transmission (processing and storage) from the Edge network to the data centers.

In comparison to other data-intensive systems, IoT sensor data have a propensity to transition from offline data operations to near-real-time or real-time operations, making a quick illustration of raw data into quality data an essential requirement. However, imputation of incomplete data is needed not only for industrial IoT and medical IoT but also in a variety of mission-critical sectors, including intelligent transportation, smart cities, and several data processing functions. Because of this, several studies have been conducted to determine the optimum way to divide data operation between the Edge system (sensor node) and the Cloud infrastructure. As a result, paradigm-shifting from Cloud to Fog/Edge devices helps in reducing energy and computational processing budgets.

### 1.2 Methodology

This section outlines the methodology that we tracked to accumulate state-of-the-art research to address the imputation of missing data in IoT and computing platforms associated with it.

*Research Scope:* The main objective of this article is to overview, classify, and analyze approaches on the imputation of the missing data and the associated computing platform. Hence, this survey article aims to provide solutions to the following research questions:

- RQ1: What are the sources of missing data in the IoT domain? How are they handled?
- RQ2: What are the computing platforms used in handling the missing data?
- RQ3: What are the challenges and issues related to imputation techniques in IoT, and what are the future perspectives in handling missing data?

To answer the above-mentioned research questions, we extract essential information from multiple databases such as Google Scholar, IEEE, ACM, Elsevier, and Springer databases until June 2021. From the deluge of literature available in the databases that deal with the imputation of the missing data, 485 articles were selected based on various keywords such as "Incomplete data in IoT," "Missing data in IoT," "Imputation of missing data," and "Missing data." After rejecting the genetic algorithm and optimization-based techniques, 279 research articles were selected that belong to the various applications of IoT. Of 279, 111 articles were excluded, because methods or applications were repeated, and the method was not extended or used in other applications. Finally, 168 articles were selected based on the scope of this manuscript.

### 1.3 Related Work

Missing data have been the subject matter of several review articles, surveys, as well as books. A broad review of handling missing data based on different topics in medical data is in Reference [118]. Norazian et al. [98] presented a review on imputation methods and software to handle missing data in time-series datasets. A survey on **Multiple Imputation (MI)** approach has been presented in References [40, 52, 86]. However, most of these surveys and reviews were focused

Table 1. Comparison of the Existing Literature and Our Survey

| | | [40] | [86] | [46] | [80] | [101] | [15] | [98] | [145] | [74] | [110] | This Survey |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Imputation Approach | Deep Learning | | | | | | | | | | | √ |
| | Fuzzy Learning | | | √ | | | | | | | | √ |
| | Machine Learning | | | √ | √ | √ | √ | | √ | √ | √ | √ |
| | Statistical Approach | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ |
| | Hybrid Approach | | | | √ | √ | | | √ | √ | √ | √ |
| Computing Platform | Cloud-based | | | | | | | | | | | √ |
| | Fog-based | | | | | | | | | | | √ |
| | Edge-based | | | | | | | | | | | √ |
| | Hybrid-based | | | | | | | | | | | √ |
| New Perspectives | | | | | | | | | | | | √ |
| Applications | | | | | | | | | | | | √ |

√indicates topics covered in the literature.

on the medical and survey data relying upon statistical methods. A review about the missing data problems in pattern classification based on intelligent techniques are presented in References [46, 80, 101, 113, 145]. Apart from these, there exist various textbook that covers handling of the missing data [42, 48, 70, 75, 117, 144].

Table 1 presents the set of imputation approaches, computing platform, and new perspectives on imputation covered by our survey and the multiple related survey literature.

### 1.4 Our Contribution

This survey article is focused on the imputation techniques implemented in various applications in IoT. Most of the existing surveys on the imputation of missing data either concentrate on a single research domain or a specific application area. Compared with other systems, IoT requires real-time analysis and prediction in many applications, including automation of the system, which helps in timely, efficient, and reliable decision-making. IoT deploys numerous sensors resulting in multiple types of data such as numerical, image, and binary, requiring various approaches to handle the missing data. To address most of the datasets in IoT, this survey focuses on "time-series-based" imputation and "pattern-based" imputation. Osman et al. [101] and Christian et al. [145] are two works that group the imputation approach into multiple categories and discuss techniques under each category. This survey extends these two previous works to address the gap in the literature by substantially broadening the discussion in the domain of IoT and its application. We add two more categories of imputation approaches: fuzzy learning and deep learning. We not only discuss the strategies in each of the categories but also determine the specific assumptions about the nature of imputations made by the approach in that category. These assumptions are crucial to determining whether the techniques in that category can impute missing values to retain the original pattern or structure. We provide a basic imputation approach for each category, then interpret how the various existing methods in that class are variants of the approaches in that group. Additionally, we identify pros and cons for each imputation approach. We also discuss the computing platform that helps to impute the missing data in real time or offline domains. We also present a detailed discussion of the application domains where imputation of the missing data has been used. We discuss different aspects of the imputation approach, issues, and the challenges faced for each domain in IoT applications. This survey is dedicated to discussing various strategies and ideas related to the imputation of incomplete data in the area of IoT. The major goals of this article are as follows:

(1) We attempt to address the gap as presented in Table 1 by providing a structured and comprehensive overview of extensive research on imputation strategies in IoT and its application. Similarly, it also discusses the computing platform that helps imputation in real time or offline domains. In addition, we highlight the sources of missing data based on the architecture of IoT. More specifically, the benefits and drawbacks of the imputation approaches are examined.

(2) We analyze critically and describe the presented state of the art by conducting a thorough discussion. We focus on the current challenges and constraints that come with developing and implementing imputation algorithms in IoT.
(3) We discuss issues and emerging problems related to imputation and highlight new applications for imputation schemes to provide a guide for implementing imputation approaches for different types of data in IoT systems.
(4) We suggest some future research directions to overcome the limitations of imputation techniques and enhance the adoption of imputation techniques in the real-world context.

### 1.5 What Is Next

In the following sections, we consider the background of the incomplete data in IoT. Strategies to handle the incomplete data are provided in Section 3. Techniques based on statistical methods are illustrated in Section 4. In Section 5, intelligent imputation techniques such as machine learning-based imputation and deep learning-based imputation are assessed. In Section 6, we highlight various performance indicators. Section 7 multiple computing platform to address the incomplete data are highlighted. In Section 8, we highlight the applications of the IoT, where imputation strategies have been applied. In the subsequent section, a discussion about the imputation techniques in IoT is discussed in detail. Similarly, Section 10 illustrates various research challenges, open issues, and new perspectives to deal with imputations. Finally, concluding remarks are presented.

## 2 BACKGROUND OF MISSING DATA IN IOT

### 2.1 Nature of Input Data in IoT

One of the most important aspects of handling missing data in IoT is the nature of input data. IoT data possess various characteristics such as noisy, erroneous and uncertain, periodicity, correlation, continuous, and voluminous, which shows information complexity due to multiple reasons like real-time processing, scalability, technical constraints (battery power, sensor aging, computing power, storage), and heterogeneity. It requires specific strategies to deal with such issues and challenges. The data generated by IoT are time-series data, which can be continuous, discrete, categorical, binary, mixed (numeric and categorical), and so on. These data instances can have relationships, e.g., spatial, graph, and sequential data. Spatial data are related to neighboring instances, i.e., data points are related in space. Sequence data contains the data points in the linear order, such as temporal data (both continuous and discrete). Spatial data with temporal components are termed Spatio-temporal data. Analytics of such data usually takes place in three classes: real time, predictive, and descriptive, which require complete data for analysis.

### 2.2 IoT Architecture and Missing Data

IoT architecture consists of sensing/object layer, network layer, and application layer [5, 73]. The sensing/object layer consists of various sensors that generate huge amounts of data, such as motion sensors, environmental sensors. Missing data results from environmental issues, meteorological extremes, routine maintenance, faulty sensors/devices, or attacks in the object layer. The network layer transmits the data using different communication tools such as mobile networks (3/4/5 G), Bluetooth, and Wi-Fi. Reasons for losing data in the network layer due to various attacks, boundary specification problems (a task of specifying inclusion rules for relations in a network study), encryption/decryption problems, or faulty devices. Those transmitted data are stored and processed in the application layer. The collected data are preprocessed (noise detection, imputation) to decide on the available data for various applications. The application layer is responsible for defining and delivering the application and services to the specific user. Incomplete data exist in this layer due

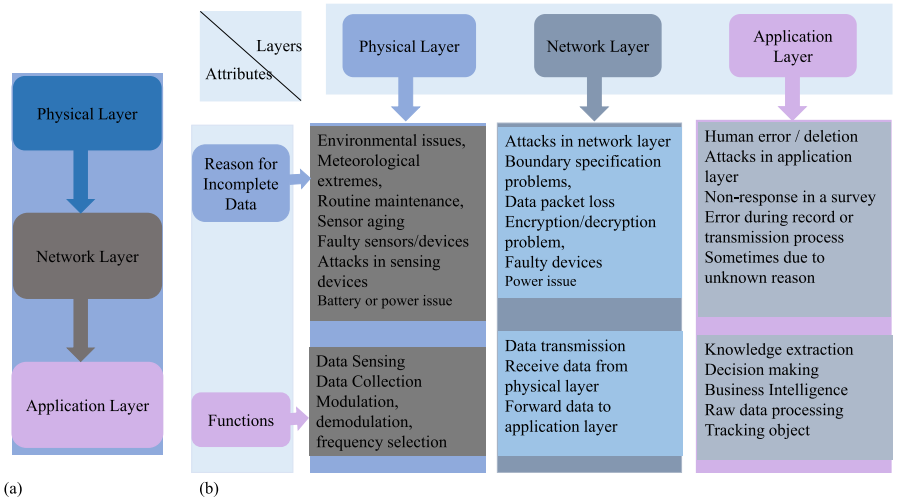


Fig. 1. (a) A Schematics illustrating the architecture of IoT. (b) The sources of incomplete data in each layer of IoT architecture along with the functions.

to human error (user may forget or refuse to use wearable sensors all time, files are lost, or data are not recorded properly), non-response in an online survey (research participants could ignore or forget to respond the questions), the high dimensionality of real problems, sometimes due to unknown reason or deletion, or error during the record or transmission process. With multiple heterogeneous networks, IoT is a complex system.

In IoT architecture, all the layers are responsible for the missing values, no matter how strictly intelligent systems are designed to collect data from high-quality sensors, regardless of how much investigators attempt to prevent the. Figure 1 illustrates the IoT architecture and reasons for the missing value in each layer. Data preprocessing (incomplete data handling) is an essential step that needs to be computed ahead of data analytics to augment data quality accurately and efficiently [144]. Hence, designing a preprocessing tool is essential to ensure a higher possibility of retaining information. High-quality data can only achieve high-quality data mining results.

### 2.3 Incomplete Data Mechanism and Pattern

Based on different assumptions and generating mechanisms, an incomplete data mechanism is categorized into three classes [75, 112, 117] as **Missing completely at Random (MCAR)**, **Missing at Random (MAR)**, and **Not Missing at Random (NMAR)**. The mechanism of the missing data is based on the classification of data. Data are in MCAR if the probability of incomplete data are not dependent on both $X_{obs}$ (observed data) and $X_{mis}$ (missing data) where missingness occurs completely at random in datasets. In simple terms, missingness is independent of input values, where no correlation exists between missing data and observed data. For example, some EEG signals have been missed due to power outages, device failure, or Internet connection problems. In such a case, the probability of an observation being missing depends on itself and the equation reduces mathematically to $Pr\,(R|X_{obs}, X_{mis}) = Pr\,(R)$ [34]. MCAR is a common but strong assumption for datasets. Traditional imputation techniques follow MCAR, and they do not create bias; however, the standard error will be increased due to the reduced sample size. A considerably weaker assumption than MCAR but still strong is MAR, which occurs if the probability of missing data depends only on the observed part but not on the missing part after controlling the observed part

Table 2. Classification of the Various Types of Gaps in the Missing Data

| Types of Missing Gap | Continuous Missing Gap |
|---|---|
| Point Missing Gap | 1 |
| Short Missing Gap | up to 3% |
| High Missing Gap | 3% to 10% |
| Very High Missing Gap | ≥10% |

[117]. The missingness depends on observed input data, where correlation exists between missing and observed data. For example, some particular data values are likely to be missing in the afternoon. The probability of missing value reduces mathematically to $Pr\,(R|X_{obs}, X_{mis}) = Pr\,(R|X_{obs})$. Since the missingness is not dependent on the observed part, i.e., MCAR is a special case of MAR [52]. It is possible to adjust the missing values; hence, MAR and MCAR can be ignorable. NMAR is the case that does not belong to MCAR and MAR. Missingness depends on the observed and unobserved value, i.e., they correlate with the sequence of missingness in NMAR. For example, a man removes wearable devices during shower time. From the three mechanisms mentioned earlier, MCAR or MAR is applicable but not under NMAR.

The result from missing data is governed by the pattern and mechanism of missing data, both of them have a more significant impact on the analysis of data. Univariate, monotone, and arbitrary are the three patterns of missing data [52]. Under the assumption of having $q$ variables, an item on $x_1, x_2, \ldots, x_q$ may have a wholly observed dataset while missing value occurs on item $y$ is a univariate pattern. It also includes the circumstances in which $y$ represents a group of an item that is either entirely missing or observed for each unit. The monotone pattern is that pattern in which the variable $y_j$ is missing for a unit and $y_{j+1}, \ldots, y_q$ is missing as well on respective dataset $y_1, y_2, \ldots, y_q$. The monotone pattern generally occurs in longitudinal datasets. When missing data occurs in a random pattern is called an arbitrary pattern. The arbitrary pattern is more complicated to handle rather than a univariate or monotone pattern.

### 2.4 Missing Amount and Gap in Datasets

Missing amount in the dataset refers to the missing ratio of the missing data and total data in the dataset. The missing gap in the dataset refers to the continuous missing gap in the variable of the dataset. It is one of the critical components that define the success or failure of imputation techniques. The missing gap in IoT dataset can be classified mainly into four categories as presented in Table 2. One single missing gap is a point missing gap. The continuous missing gap of more than 2 continuous gaps to 3% of the data in a variable is termed as a short missing gap. The continuous missing gap from 3% to 10% of the data in a variable is termed as a high missing gap. The continuous missing gap higher than 10% of the data in a variable is a very high missing gap. Point and short missing gap occur frequently. When high and very high missing gaps exist, the trend, pattern, and structure of the data are entirely broken. Recovery of such gaps needs to address patterns, trends, and structure of the data that makes the process more tedious and complicated. It is to be noted that most of the existing research conducts experiments to impute missing data stating that simulated or real data consists of 50% of the missing data. Such missing data are based on the missing rate or amount, lacking the information about the missing gap. Those data usually contain a point missing gap and a short gap only. A significantly small amount of the literature addresses the high missing gap problems; this shows an immense need for more research to handle the ongoing missing gap problems, because most of the existing methods result in a high error and low accuracy. The high missing gap problem with low error and highly accurate imputation results are addressed in References [2, 78].

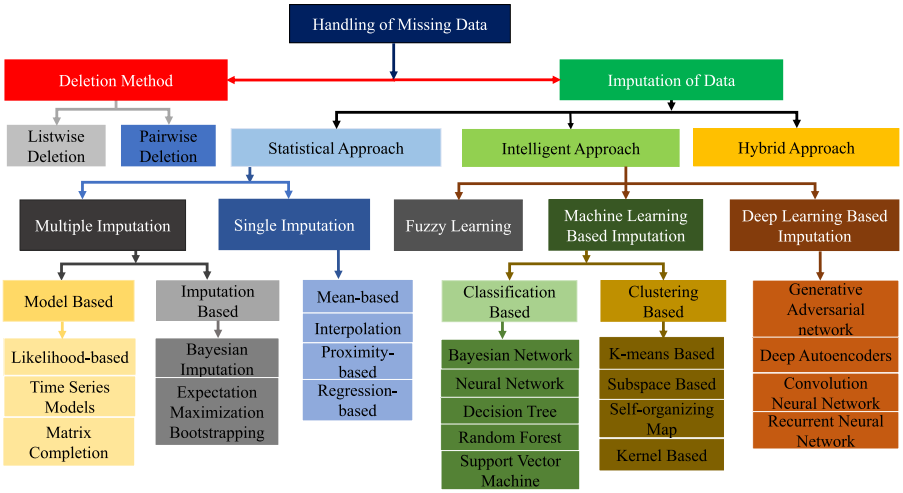


Fig. 2. A Schematics demonstration of the methods to handle the missing data in IoT. Here, imputation techniques are divided into three main categories: statistical, intelligent, and hybrid approach.

## 3 HANDLING OF MISSING DATA IN IOT SYSTEMS

This section provides a brief introduction of the major strategies of imputation in the IoT domain. Moreover, we categorize the existing methods according to the assumptions used, principles, and approaches. Various strategies adopted for the handling of missing data are described in Figure 2.

### 3.1 Deletion Methods

*3.1.1 Listwise Deletion.* Listwise deletion or complete-case analysis, the most commonly used method by data analysts, and the default in data analysis packages remove all the data for a case that has one or more missing values [52]. An "honest" method for handling missing data is compared to an intelligence-based method or a hybrid method, because the standard error predicted is usually precise estimates of true standard errors. It works fine in MCAR due to the sub-sample of cases when complete data are identical to a simple random sample from an original target sample, which rarely happens in reality [130]. However, when data are in MAR, it produces a bias in parameters and estimates, indicates loss of potential information and power in the testing hypothesis, creates wider confidence intervals, and a more massive standard error [75].

*3.1.2 Pairwise Deletion.* Pairwise deletion or available case analysis, where only missing values are discarded is a framework to proceed the imputation process. It does not create pseudo-values to inflate the amount of information or abandons the information [52]. It focuses on the variance–covariance matrix but can produce an inter-correlation matrix that is not positive definite, which prevents further analysis [75]. Based on the available cases, estimates of mean, variance, and covariance or correlation are computed with the caveat that there will be different sample size datasets and standard errors [42, 144]. It is known to be less biased for MCAR or MAR data as appropriate methods such as covariates are included but not suitable for high missing observation.

### 3.2 Imputation Strategy

A standard method to handle missing values is to fill them with estimated (imputed) values. Its primary objective is to maintain the quality of the data by imputing unbiased and trustworthy data. All the sensor-generated data have their own IoT paradigm features and are different from

each other such as numerical, ordinal, and binary. When such data values are missed, they require various methods to handle them. For example, data generated from biometric sensors are based on pattern recognition, whereas temperature sensors are time series. Different techniques lie under the heading of imputation. The imputation approach is further classified as a statistical-based method, an intelligent-based method, and a hybrid method. These methods are illustrated in Figure 2 and discussed in the next section.

## 4 STATISTICAL-BASED IMPUTATION TECHNIQUES

Statistical imputation is classified into two types as **Single Imputation (SI)** and MI [59]. However, Reference [46] classified three statistical methods: SI, MI, and model-based. Based on the definition of MI, a model-based approach is also an MI.

### 4.1 Single Imputation

Single Imputation is the process of filling distinctly one value with some reasonable guess (real value) for each missing one. *Assumption:* Each missing value is assigned one value only; hence, imputation takes place only once. There are various kinds of SI proposed by researchers that are described briefly as follows:

*4.1.1 Mean-based Imputation.* This approach substitutes for all missing values with their corresponding mean from the observed data, which entails inflation of the certainty of the information. This is one of the easiest strategies of imputation and can significantly break the inherent structure of the data. In IoT applications, mean imputation does not preserve the trend, data structure, seasonality and may act like an anomaly, such as conditional anomaly and collective anomaly based on the missing gap. The major drawback of this method is that it leads to inconsistent bias, increases sample size without any information, decreases correlation, and underestimates the true variability of the variables and the sampling error of the estimate. There exist various mean-based imputation that performs competitively in time-series datasets [52]. They are row mean, mean top-bottom, hour mean, mean 6-hour, mean 12-hour, daily mean, last and next mean, and previous year mean [98].

*4.1.2 Regression-based Imputation.* Regression imputation is also termed conditional mean imputation, where an observed item of the data is used to predict the missing item. Regression models build the relationship between the dependent and independent variables (one or more) to predict, relying upon the independent values. It assumes that the imputed values lie on a regression line having non-zero slopes, which implies a correlation of 1 between the missing outcome variable and predictor. When the random residual value is added to the predicted value, it is called stochastic regression [42]. Variability and covariance are better preserved and less underestimated in this solution, but the imputed value lies directly on the regression plane, so there is still some degree of underestimation. Regression imputation is usually implemented when the dataset exhibit a temporal pattern and does not contain noisy data. Better estimates are observed from this method when the correlation is linear; however, accuracy might decrease when predictor and response are not highly correlated. Data deviation from the mean is also preserved through parameter estimation. It also underestimates the variance and overestimates the correlation. From the distribution error, the estimated error can be estimated [52, 75]. Some methods that implement a regression are as follows.

*Autoregressive Moving Average Models:* The **Autoregressive Moving Average Models (ARMA)** combines the basic linear processes, Autoregressive and Moving Average, representing a stationary time series. The model is dependent on past values and random error of the past

time and cannot deal with the issue of trend and seasonal patterns that are non-stationary. The auto-covariance and power spectrum function are decided completely as the estimated parameter of the ARMA model. The accuracy of the ARMA model can be obtained better than the nonlinear optimization of the log-likelihood function by using a reduced-statistics algorithm [23].

*Autoregressive Integrated Moving Average Models:* To solve the problem of non-stationary time series, a new model was proposed, which is the integrated form of the ARMA known as **Autoregressive Integrated Moving Average Models (ARIMA)** [22]. To estimate parameters in the ARIMA, autocorrelation function and partial autocorrelation function are also crucial to compute, which helps to test if the residual is white noise. The ARIMA model is used when the dataset exhibit temporal pattern and contains large datasets. However, alteration in observation and model specification leads the model to be unstable.

*4.1.3 Interpolation.* Linear interpolation fits a gap between the last and next observation based on the mean of values over a similar time interval with an identical fluctuation pattern. The performance of this method is limited though it is one of the best imputation methods when the missing gap is one or two [58]. However, when the missing gap increases, accuracy decreases. Prediction ability is also limited, i.e., lies between last and next forward observed data. There are various types of interpolation, such as cubic, inverse distance weighting, Kriging, optimal interpolation [140, 153]. These methods perform well when there exist a point missing gap and a short missing gap in a sensor-generated dataset.

*4.1.4 Proximity-based Method.* In this procedure, the incomplete values are substituted based on the proximities measured. Methods using different proximities measures are as follows.

*Nearest Neighbor:* Depending on the temporal proximity of the missing values to known observations, the last or next observations of the gap are used to determine all the incomplete values [99]. The method performs well when there is a missing gap of one or two. Prediction is limited so, imputation in high missing gap leads to pattern anomaly, and accuracy decreases [58].

*Hot/Cold Deck:* Hot deck substitutes the missing values from those individuals who have observed matching values and are closer in terms of distance in the current dataset [130]. Thus, non-response bias declines to the extent that there is a collation between the variables defining imputation classes and the tendency to respond and the variables to be imputed [10]. The method is suitable in certain missing patterns. It imputes genuine and reasonable values, invalidates powerful parametric conditions, can include covariate information, and delivers reasonable inferences for linear and nonlinear statistics in the presence of imputation uncertainty. The method is problematic when the pattern of missing case differs; the correlation between other variables and imputed variable could be weak, causing the imputed variable to lose a portion of variance; and estimating standard error can be difficult and not suitable for big data where the number of classification variables may be unmanageable. Cold deck substitutes incomplete values with a model or external information rather than the closest entity in terms of distance or information available in the dataset [98].

### 4.2 Multiple Imputation

In multivariate analysis, MI is one of the most appealing general proposals approaches for dealing with missing data proposed by Rubin [112]. It replaces a missing value by $m \geqslant 2$ with possible values, each with a unique estimate reflecting the uncertainty attached. $m$ estimates are combined to yield a single estimate. Correct estimates of the standard errors and $p$-values can be achieved

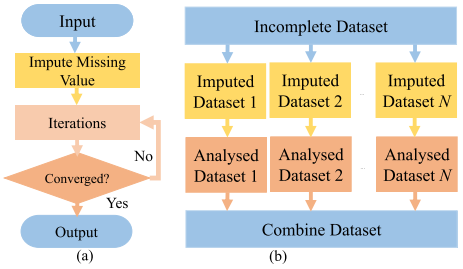


Fig. 3. A Schematics demonstration of the multiple Imputation, (a) represents Model-based Method and (b) represents Imputation-based Method.

from the distribution of the variables from the different estimated $m$ underlying distribution or imputation models [75].

*Assumption:* Imputation takes place multiple times with various unique imputed values. The final estimate is obtained through an optimal solution or by combining multiple values. Relying on the definition of the MI given by Rubin, based on the literature published and the assumption used, we further categorize MI into two classes as model-based approach and imputation-based approach, which are presented in Figure 3.

*4.2.1 Model-based Technique.* In the **Model-based Technique (MT)**, a predictive model for each target variable where data assumptions are made through the combined distribution of the variables available in the model [46]. Usually, imputation takes place by the SI procedure or by the random filling of the data. Imputation of the variable takes place once and is iterated further till the value gets converged. Various types of iterative imputation techniques are illustrated in Reference [77]. Figure 3(a) illustrates the schematics of model-based multiple imputation. In this method, each missing value gets two values, the initial imputed value and the final converged value.

There are various types of model-based methods, which are described as follows.

*(i) Maximum Likelihood-based Imputation:* Handling missing data using the theory-based **Maximum Likelihood Imputation (MLI)** approach long has been known [36]. When assumptions are met, an MLI estimates missing data with desirable properties such as asymptotic normality, asymptotic efficiency, and consistency. Asymptotic normality is essential for a normal approximation to calculate $p$-values and confidence intervals. Asymptotic efficiency is essential to reduce standard error, i.e., estimation close to being fully effective. Consistency indicates estimates are approximately unbiased in large samples. The MLI method is suitable when there is a realistic assumption on the distribution of the data. Given the available data points, this procedure may produce the variance–covariance matrix for the variables in the model and then use the variance–covariance matrix to evaluate the regression model [117]. This approach is based on parametric mode, i.e., multivariate Gaussian mixture model, which is applied in both imputations of missing data and fitting the model. Some of the techniques that implement an MLI are as follows.

*(a) Expectation maximization:* An **Expectation maximization (EM)** was proposed by Dempster [36], an iterative algorithm to find maximum likelihood estimates and fit models of missing data problems. It capitalizes on the relationship between the unknown parameter and the missing data. It does not impute the missing data but instead estimates data based on an **expectation step (E-Step)**. The E-step creates a function for the expectation of likelihood using the current estimate of the parameter, and a maximization step computes parameters to maximize the likelihood available on the E-step, and this is repeated several times until the MLI estimates are obtained. The

sequence of parameters converges to the MLI estimates that implicitly average over the distribution of missing value [40]. It can be used to get consistent estimates of the parameter of interest by estimating means, standard deviation, and correlations or equivalently known as means and covariance matrix. It requires a large sample size and data in MAR [52]. An EM estimator is more efficient and unbiased when the missing mechanism is ignorable. The method is complex, convergence is slow, does not compute the derivatives of a log-likelihood function, and does not provide a better estimate of standard error.

*(b) Full Information Maximum Likelihood:* The **Full Information Maximum Likelihood (FIML)** was first examined by Arbuckle using the Confirmatory Factor Analysis model. This estimation uses information from both the complete and incomplete observations in an integrated manner [7]. It maximizes the sum of the log-likelihood function for individual observation, which contains both missing and non-missing observations. Modal parameters are estimated by MLI function [42]. Different types of modeling are used to analyze the FIML estimation, like the regression model and structural equation modeling. The "full information" identifier in the FIML demonstrates more about the incomplete data analysis than the fundamental estimation principle [7]. The reasons that make the FIML popular are testing model fit can be performed easily and directly, and produced estimation result is always the same, can handle any linear model, and gives efficient estimates with correct standard errors. However, the computational cost is high, requires specialized software as a specific model, fitting structural equation model when missing data are higher and observed data are low, and cannot deal with negative binomial regression (count data) or logistic regression (dependent dichotomous variable).

*(ii) Matrix Completion Method:* The matrix completion method uses a matrix method to complete the incomplete datasets. There are various methods that use the matrix completion tool.

*(a) Probabilistic Matrix Factorization:* The **Probabilistic Matrix Factorization (PMF)** decomposes a single matrix into a product of two matrices and can obtain the original matrix by computing the product of two matrices [92]. It is used when there is a numeric dataset. Fekade et al. [44] implemented PMF to recover incomplete data using the similarity measures and $k$-means clustering. The PMF reduces the total number of stored values in a big data array due to the low dimensionality after factorization, making it suitable for large datasets. Scalability in large datasets is achieved by using stochastic inference methods that randomly sub-sample missing matrix entries. Obtaining a convergent solution might be a problem.

*(b) Singular Value Decomposition:* This is a powerful property based on the matrix factorization approach that decomposes a matrix into a product of three simpler matrices [141]. The imputation begins by replacing the missing values with zero and iterates through the **Singular Value Decomposition (SVD)** until convergence using the EM algorithm and computes the SVD to obtain eigenvalues. he obtained eigenvalue is applied to regression to the complete attributes of the instance and the estimated imputation of the missing value. Using the matrix method, one must be cautious that the results are not definite positive, or the last eigenvalue must be negative. Sometimes the methods might fail, especially when the correlation matrix is not positive definite [52]. The SVD is used for many purposes, and imputation is one of them. In **Intelligent Transportation Systems (ITS)**, various methods about tensor decomposition methods have been proposed and implemented. Tensor decomposition is the extension of SVD in higher order [31, 87]. SVD is computationally fast, can handle various sized data, and is used in mixed types of data.

To get a preliminary **Principal Component Analysis (PCA)** model from the available data, it requires a previous estimation of the number of components. Then, the imputed values obtained from a regression loadings substitute for the missing value.

*(c) Principal Component Analysis:* The PCA approach is identical to the projection to the model plane of Philip et al. [93] without contributing error prediction. Estimating the number of components before generating an initial PCA model from the provided data is necessary. Then, the imputed values obtained from regression loadings substitute for the missing value, which calculates from the updated data matrix, and the process is repeated until convergence [93]. It is based on a strong theoretical framework that improves the performance of an algorithm that is implemented in the mixed dataset. The computational complexity is high, sometimes convergence and overfitting is a problem. Instability in the imputations may yield due to the presence of too many voids [50]. This method has been extended using **Probabilistic Principal Component Analysis (PPCA)** is a combination of an EM and PCA, and **Bayesian Principal Component Analysis (BPCA)** is the combination of PCA, Bayesian estimation, and EM [91, 106] and implemented in ML.

*(d) K-Nearest Neighbor:* **K-Nearest Neighbor (KNN)**, the most popularly implemented machine learning–based imputation algorithm [74], identifies the closest $k$ observation based on a distance matrix and computes the weighted average (weighted based on distance) of $k$, for each observation. The incomplete data are substituted by the mean of the corresponding attributes of $k$ neighbors that contain complete data and take the correlational structure of the data [141]. The common distance metrics used are Pearson correlation, Euclidean distance, Mahalanobis, variance minimization, and so on [30]. The value of $k$ will be different based on the problems. Increasing the value of $k$ in the KNN and its extensions such as Weighted KNN), Sequential KNN, Gray KNN augments the time complexity to impute missing value and capture close records but may not considerably enhance the results [80]. Imputation using the KNN is preferred in the presence of numeric dataset. Modified KNN based on temporal and spatial correlation is implemented in Reference [137]. Imputation of the real data, all the missing values can be imputed in all variables with a single call function are its benefits. However, this method is not suitable for large datasets and encounters hardships when there exist many variables.

The KNN has also been extended using the ML technique to impute the incomplete data [39]. In [103] implemented KNN in the spatiotemporal data, compared with other time-series imputation strategies where KNN is the most superior imputation. Reference [32] proposed a new purity-based KNN for financial datasets, and the results achieved better accuracy and stability.

*(iii) Time-series Modelling:* A specific type of analysis that is based on time is time-series analysis. An attractive method to model the data as time series is due to its data-dependent nature on the previous values within the same variable. Imputation can build a model for the data and then use the imputation model to impute missing data. Some of the models are described here.

*(a) Box–Jenkins:* The Box–Jenkins model is a univariate model based on statistical principles and concepts capable of the model-wide spectrum of time-series nature. For time-series forecasting, it is a general framework that prioritizes identifying the interactive approach with a suitable model represented by a linear combination of random variables and past data [22]. The Box–Jenkins is one of the popular choices on time-series modeling among data analysts due to a wide variety of time-series patterns that can be modeled. The model can be verified using many statistical validity tests, used for accurate prediction, and so on. However, the Box–Jenkins methods cannot be applied directly but can be computed with the help of the MLI or Kalman filter algorithm [49].

*(b) Kalman Filter:* The Kalman filter is a recursive, optimum technique for estimating the state of a system by disseminating a probability density function accustomed to a collection of observations at any point in time. It works under the assumption of linearity, Gaussian distribution, and all noises are white. The Kalman filter always gives a unique estimate due to its assumptions and probability density function. It imputes or estimates missing value using the Kalman smoothing

and state-space models. Sometimes overestimates and underestimates the missing value resulting in biased results [20].

*(c) Singular Spectrum Analysis:* The **Singular Spectrum Analysis (SSA)** is also considered as a principal component analysis in time series that evoke information without the knowledge of dynamics affecting the time series from noisy and short time series [120]. Eigenvectors and eigenvalues are computed for the lagged autocorrelation matrix. Then, each PCA of the original time series identified by the SSA is reconstructed. The PCA is computed with scale factor and eigenvector to compensate for the missing values. Improved SSA for imputation is implemented in Reference [126] where lagged correlation matrix is computed as in SSA and PCA are computed directly from the eigenvector and eigenvalues of the lagged correlation matrix. Similarly, the SSA with conditional prophecy for real-time state computation and forecasting is implemented [100]. SSA imputation is deployed for continuous and discrete time-series data where every component of the observed time series can be reconstructed.

*(d) Kernel-based Imputation:* Kernel-based imputation using parameter optimization method is proposed in Reference [105]. This method makes optimal inference on various statistical parameters such as mean, distribution function, and quantile after the imputation of the missing data takes place. This method is believed to be better than the deterministic regression imputation in terms of efficiency. The traditional Kernel function is extended using non-parametric iterative imputation using a mixture kernel for estimating missing values in a mixed-attribute dataset [166]. Kernel-based imputation is used when there is a discrete or continuous time-series dataset. The method appears to be more robust to the violations of distributional assumptions than the existing doubly robust methods. However, sometimes there is an augmented bias in the estimation.

*(e) Ratio-based Imputation:* The **ratio-based imputation (RBI)** approach is based on information fusion techniques to handle high missing gap datasets generated from the sensors [2]. The method undergoes multiple steps such as ratio computation, imputation, and recovery. Imputation takes place by interpolation techniques. The method handled the highest missing gap of 21% with the best results among compared techniques such as **EM Bootstrap (EMB)**, **Multiple Imputation by Chained Equation (MICE)**. The method is improved by using an EM algorithm known as **Iterative RBI (IRBI)** [3]. The IRBI improved the accuracy of the imputed dataset compared to RBI, EMB, MICE. RBI and IRBI are deployed under the numeric temporal dataset. The method performs well in the very high missing gap and when the dataset is highly correlated; however, it works only on numeric data.

*4.2.2 Imputation-based Techniques.* In **imputation-based techniques (IT)**, multiple values are allocated for each missing value based on Rubin's concept [112]. For each missing value, multiple unique values are allocated. At last, using certain rules, the imputed data are combined to get a single estimate. Figure 3(b) illustrates the schematics of imputation-based multiple imputation. Various types of IT are illustrated below:

*(a) Bayesian Approach:* Natural-based solution to treat the missing value as random values by estimating their posterior distribution is offered by the Bayesian paradigm. Multiple Imputation involves the Bayesian paradigm, the fundamental law of probability known as Bayes's Theorem, which offers an alternative model-based solution where missing values are treated as an unknown parameter, drawn randomly through appropriate distribution, and we assume that the missingness mechanism is ignorable. Two imputation methods exist to deal using this approach as follows.

*Markov Chain Monte Carlo:* Multiple Imputation was first proposed by Rubin based on a Bayesian computational algorithm known as **Markov Chain Monte Carlo (MCMC)**, also known

as Data Augmentation [138]. Data are assumed to be multivariate normal distribution and applied to the Bayesian inference that repeats the **imputation step (I-step)** and **posterior step (P-step)**. I-step simulates the missing data value randomly from the available distribution value and iterates the process until simulated error distribution is smaller than the pre-specified criterion or distribution is stationary for mean, variance, and covariance. Imputation occurs if the condition is satisfied. From the imputed datasets, P-step recalculates mean, variance, and covariance matrix from I-step when iterations are not enough, making a random draw from the posterior distributions of this parameter. Finally, these parameter values are combined to estimate standard error, efficiency, and update essentials needed for imputation [144]. These two steps are iterated sufficiently undergoing estimation by repeated conditional substitution and create a stochastic process known as a Markov chain, which stabilizes or converges in distribution [52, 138]. MCMC imputation is used when the assumptions are reasonable. Some of the benefits of using this method include using all the available data, considering the data variability, and missing observations are imputed using estimates as the starting points for augmenting the missing data points. Drawbacks include the assumption of the multivariate normal distribution, computationally demanding, and high iterations requirements.

*Multiple Imputation by Chained Equation:* MICE, also known as fully conditional specification, is an alternative algorithm that is a flexible and semi-parametric alternative that specifies the multivariate normal distribution through a series of conditional densities by which imputation occurs [138, 149]. The initial imputation starts with random draws from the corresponding posterior predictive distribution known as proper imputation [26]. Random draws are made at each step from both the posterior distribution of the parameters. When the missing value is replaced by the imputed value (from posterior predictive distribution), an approach identical to the Gibbs sampler is repeated multiple times before selecting a complete dataset to stabilize the imputed results [149]. MICE is deployed in all types of data. The method is simple and flexible and deals with complexities such as bounds or survey skip patterns, which can handle all types of data. Some demerits include that statistical properties are difficult to establish, computationally much slower, and do not have theoretical justification.

*(b) Expectation-Maximization with Bootstrapping:* EMB is a technique generated by integrating the EM algorithm and the Bayesian classifier using the bootstrap method to make draws from the posterior distribution [54]. Imputation Posterior in MCMC and EM are the two statically appropriate ways of taking those drawn from a posterior distribution. EMB takes advantage of the best features of each technique and integrates them. Bootstrap estimation is usually used when a parameter distribution is supposed to be non-normal, and bootstrap inference with missing data is not clear [121]. Honaker et al. [55] describes the procedure to implement EMB using software named AMELIA II. EMB imputation is implemented when data undergo multivariate normal distribution. The result produced by EMB is slightly better than the consequence performed alone. EMB is highly accurate in building a classification rule. However, it is challenging to implement for attaching standard error, especially when imputing missing data, and the process to obtain valid bootstrap inference remains unanswered.

Due to space limit, the Combining Estimates, Standard Error, and Efficiency in Multiple Imputation is available in the supplementary online material.

## 5 INTELLIGENT-BASED IMPUTATION TECHNIQUES

Intelligent-based imputation techniques are further classified into two divisions as machine learning-based techniques and deep learning-based techniques. The advanced form of machine

Table 3. Some Examples of Classification-based Imputation Using Neural Network

| Neural Network Used | References |
|---|---|
| Multi Layered Perceptron (MLP) | [128, 129] |
| Radial Basis Function (RBF) | [38, 81] |
| Auto-associative Neural Network | [47, 66], |
| Probabilistic Neural Network (PNN) | [96, 132] |

learning is deep learning that uses multiple layers of a neural network to learn and make an intelligent and precise decision on its own. It possesses high computational complexity requiring a large amount of data and a longer processing time. Similarly, due to the existence of many imputation approaches based on deep learning, the difference in computational complexity, and algorithm structure, we classify intelligent techniques into machine learning and deep learning.

### 5.1 Machine Learning–based Imputation

*5.1.1 Classification-based Imputation.* Classification is a technique for learning a model (classifier) from a set of labeled data attributes (training) and then using a learned model to classify a test instance into one of the classes (testing). Using the accessible labeled training data, the training phase learns a classifier. Similarly, using the classifier, the testing phase classifies the test instances as normal or imputed. Classification-based imputation operates under the following assumptions.

*Assumptions:* A classifier can impute the missing value in the given feature space.

*(a) Neural Network Based:* A neural network is a probabilistic model for processing data performed by the biological nervous system, such as the human brain. Neural networks are known to be capable and highly efficient in solving complex tasks related to forecasting and modeling of the experts and intelligent systems [19]. The neural network models are based on the classification of the missing data and have been applied under multi-class and one-class settings. Various models under the neural network are used to impute missing data that belongs to various types of data. Neural networks imputation is used when there is a nonlinear relationship between variables, and the missing mechanism cannot be determined. Neural networks can be modeled for complex patterns without prior information. However, unrealistic results can be achieved while performing numerous patterns containing noise. A variety of combinations can result in necessitating numerous neurons. Table 3 illustrates the several neural networks techniques and references that have been implemented in the handling of the sensed data.

*(b) Bayesian Network Based:* A Bayesian network-based method is ubiquitous for modeling and reasoning under uncertainty in classification and prediction problems. It is a probabilistic graphical model representing a joint distribution of random variables [116]. The testing phase is fast; however, each test instance might be essential to be compared again to the pre-computed model. Computational complexity is high, making it computationally slow. Roosevelt et al. [116] presented a strategy for revising a Bayesian network structure in the presence of missing data. A node is associated in a directed acyclic graph where an edge connecting a pair of nodes stands for a direct relationship at each random variable. The relationship between the variables is represented in a human-readable way, taking local probability distribution and conditional independence into consideration. Chen et al. [152] implemented a joint model to impute the missing data by integrating the Bayesian inferences with crowd-sourcing. This helps to improve the performance and accuracy of the imputation algorithm, and reduces cost. Similarly, Ma et al. [86] presented a review of the imputation using Bayesian methods.

*(c) Support Vector Machine (SVM):* Relying upon statistical learning, a powerful ML techniques for regression or classification is developed known as SVM. SVM is based on SVM regression to prophesy the missing condition attributes values. The procedure begins with selecting the no missing attributes value where the condition attributes (input attributes) are set, with incomplete data, decision attributes (output or classes), and decision attributes as the condition attributes by contraries. Then, decision attributes are predicted using SVM regression [56]. SVM imputation can deal with continuous and discrete numerical types of data. It can effectively handle non-linear problems and complex missing data patterns by lowering the generalization bound error, ensuring best-case performance [33]. When over-fitting occurs, a remedy to diminish the number of errors on the training set that underperform on data is identified. However, a kernel function must be chosen, and cross-validation must be used to estimate a small number of parameters. SVM was extended using least square SVM [146] and is comparatively faster than the SVM.

*(d) Decision Tree:* The main objective of **Decision Tree (DT)** is to generate a model that can predict the value of a target variable depending upon the various input variables. Decision Tree splits the dataset into leaves, where mutually exclusive records exists in each leaf. Imputation takes place within the leaves. The classification tree and regression tree are two types of DT. Commonly used DT are **Iterative Dichotomiser 3 (ID3)**, C4.5, **Classification And Regression Tree (CART)** [108].

C4.5 is the extended form of ID3 that allows the features of the discrete attributes. Pruning is done by the rule's prediction that takes place by the removal of the rules precondition. CART is similar to C4.5; however, it does not compute rule sets and supports regression (numerical target variables). CART and C4.5 are used for imputing numerical and categorical attributes, respectively. C4.5 is extended using the EM algorithm known as EMI [119]. EMI is extended using DMI and SiMI [109]. The extended methods showed clear superiority of the techniques compared to the existing ones. Decision Trees are deployed in the presence of mixed types of data. Non-linear variables are handled efficiently, and training time is reduced, because the system is simple and can automatically handle missing values. Sometimes DTs are unstable (in addition to data points), unsuitable for large datasets due to high variance and high risk of overfitting.

*(e) Random Forest:* **Random Forest (RF)**, an extension of the classification and regression tree, integrates the tree predictors where each tree is dependent on the random vector sampled independently in the same distribution of all trees in the forest. The RF-based algorithm for imputation is MissForest [134]. This procedure accurately predicts the individual incomplete value instead of random draws from a distribution, leading to biased parameter estimates in statistical models. Random Forest imputation is used especially for mixed types of data. It has a robust predictive power and does not rely on specific distributional assumptions (such as regression model) and can accommodate interactions and nonlinearities [123]. To reduce the risk of overfitting, RF uses bootstrap aggregation of multiple regression trees, and more accurate predictions occur after the combination of predictions from the various trees. Apart from this, various algorithms exit under RF, such as proximity imputation, on-the-fly imputation, and imputation utilizing multivariate unsupervised and supervised imputation [139]. Computational time is high, making it not suitable for high datasets and lacks interpretability. Kokla et al. [64] compared mean, SVD, PPCA, BPCA, RF, and KNN. The results showed that the RF performed better in all the experimental analyses. Similar results were illustrated in Reference [123].

*5.1.2 Clustering-based Imputation.* Clustering-based imputation techniques are based on the following assumption:

*Assumption:* Partition of the data into several clusters based on similarity patterns and minimize the intra-cluster dissimilarity.

The clustering algorithm categorizes data elements into various classes known as clusters. Cluster analysis includes analytical decisions, selection of a number of clusters, which algorithm and metrics to use. Xavier et al. [14] designed a framework to integrate multiple imputations to cluster analysis. Based on the recent literature, the clustering method adopts two different frameworks: single-view and multi-view clustering. The classic clustering method is single-view clustering, in which data are shown using all available features. Similarly, multi-view clustering is the new technique where data are depicted based on the multiple subsets of features [153, 163]. Recently, temporal, spatial, global, and local views are being considered in the multi-view approach. Xiuwen et al. [154] used the temporal correlation and spatial correlation-based multi-view learning method to impute time-series data. The main goal of the clustering technique is to classify the dataset into clusters according to the similarity of the objects by minimizing the intra-cluster dissimilarity.

- Kernel-based methods: Using the kernel matrix, incomplete datasets are preprocessed, and final clustering is achieved by exploiting the multiple kernel learning [124]. It is also known as partitioning or prototype. Here, non-linear structures have been identified, and the dataset is suitable for real-world use. In complex types of data, the fast testing phase can be adopted. Specifying the number of clusters and initialization of random assignment of observation has a big impact on the outcome of the cluster.
- Subspace-based approach: The transformation metric or matrix factorization integrated with the regularization to project missing views into a shared latent clustering subspace [155].

*(a) k-means Clustering:* In this technique, intra-cluster dissimilarity is computed by using the distances between the objects in a cluster is measured along with the mean value (centroid) of the cluster [71]. Every data item in a cluster contains the membership function that defines the degree to which the data object belongs. Only the complete attributes are taken into consideration during the membership updating process. Each missing data item replaces non-reference attributes based on the information about membership degrees and the cluster mean (centroid) value. It is to be noted that cluster mean is not assigned as the data items of the concrete cluster. After the cluster gets converged, all the non-reference attributes for each incomplete object based on cluster information are imputed. The same cluster data objects are treated similarly to KNN. *k*-means clustering is extended using a fuzzy approach known as fuzzy *k*-means clustering [88], where missing values are estimated using a weighted average estimated from cluster centroid and membership degree. *k*-means clustering imputation can be deployed on mixed types of data. The algorithm is flexible, easy to implement, efficient, and more accurate. However, it is dependent on the initial value and needs to choose the value of k manually.

*(b) Self-Organizing Map:* **Self-Organizing Map (SOM)** defines a mapping from the input space of a low-dimensional space, i.e., 2D, and enables the feature space dimension approximation into a projected 2D space by preserving the topological properties. Hence, it is useful in visualizing the dimensional view of high-dimensional data through the nonlinear projection using the neighborhood function [45]. SOM is a competitive learning method where the training algorithm is iterative, used in finding the imputation classes [115]. The weighting vectors are initially set randomly; however, they converge until the end of the training process to achieve a stable value. SOM is capable of modeling the nonlinearities of a system considering relationships among pertinent variables in the vector-profile of the data record [89]. As a clustering method, it is a basic competitive topological network preserving during the mapping of input space to the clusters from the input space. The network parameter and structure can change using the adaptive self-organizing structure through which samples of inherent laws and important qualities may be automatically sought. SOM is a

good tool used in the prediction of a variety of datasets, including temporal data, as it preserves the crucial relationship of the observed data elements. In the presence of large datasets, the convergence rate is robust, and the computational benefit is significant. However, there exists a limited capacity for prediction, i.e., prediction is within the range of the observed value.

*5.1.3 Fuzzy Approach. Assumption:* Imputation techniques based on Fuzzy approach.

The fuzzy approach was developed by Zadeh [158] based on probability theory to handle uncertainty, impreciseness, and incertitude tangled in the decision-making process. Various imputation strategies based on fuzzy approach exist in the literature [8, 16, 88, 94, 95, 122]. Fuzzy approaches have been used for both the classification [8, 61] and clustering [88, 94] imputation techniques. A fuzzy rule-based classifier was proposed in Reference [16] where each principle can be broken down into a particular single-dimensional member functions analogous to fuzzy sets. Mehran and Richard [8] used a fuzzy rough method to impute missing values without using iteration methods or optimization methods based on classifier techniques. Fuzzy clustering is appropriate when an instance does not belong to any class as a missing value. The imputation process undergoes modularity and explainability. Execution rules can be carried out in parallel. However, defining rules is a time-consuming process, making it computationally expensive.

Usually, Euclidean distance is deployed as the similarity function in fuzzy c-means [18]. Razavi and Saif [111] implemented fuzzy neighborhood-based clustering techniques. Nikfalazar et al. [95], Ming et al. [88] used $k$-means to initialize the fuzzy model. The global $k$-means algorithm is to compute the number of clusters present and to improve the computation. This method was further improved by using grey system theory in Reference [122]. The use of grey system theory increased the performance of Fuzzy c-Means Clustering. Sanaz et al. [94] used iterative methods on **fuzzy $k$-means (IFC)**, which improved the imputation results. Similarly, the integration of IFC and DT leads to a DFIC [95], resulting in robust and better performance compared to other techniques.

## 5.2 Deep Learning–based Imputation

Deep learning techniques allow computational models consisting of several processing layers for learning data representations with numerous abstract levels [69].

*Assumption:* Methods that use deep learning to impute the missing data.

Various types of imputation techniques implemented in deep learning are as follows.

*5.2.1 Deep Autoencoders.* Deep autoencoder networks are such networks that memorizes input as output in the output layer, predicting new input values as the outputs when presented with new inputs [70]. Miranda et al. [90] implemented an autoencoder with backpropagation to impute missing data, where each autoencoder consisted of a single hidden layer, and linear activation function was available in input neurons. The use of the evolutionary particle swarm optimization approach helped in recomposing missing values quickly and efficiently. Contrary to the autoencoder, variation autoencoder was proposed in Reference [63]. Variation autoencoder is a generative model that encourages generalizing the features by reconstructing samples and aggregating them, forcing the latent space to be continuous. Here, a stochastic gradient optimization approach was used. Fitting an approximation inference model to a dataset containing continuous latent variables per datapoint can make posterior inference more efficient. Similarly, Abiri et al. [1] proposed denoising autoencoder, which can reconstruct data by stochastically corrupting it. The method can handle multiple types of data, including mixed types of data in low computational time, which implemented a stacked denoising autoencoder. Linchao et al. [72] proposed MultiModal Deep Learning model based on stacked autoencoder, a deep network created by stacking many layers of autoencoders, to impute the spatial and temporal traffic data. The distance between the latent features of

two heterogeneous datasets is reduced in the total loss function. The method accurately captured temporal and spatial dependencies. Autoencoder imputation is robust and compact in the final encoding layer, which can be applied for all types of data and real-valued data. The computational complexity and cost are high. Sometimes, the method is slower on continuous and binary data during learning time due to decoder backpropagation. Overfitting can occur due to a lack of data on using the overparameterized model.

*5.2.2 Recurrent Neural Network.* The **Recurrent Neural Network (RNN)** is based on the objective to use the sequential information, which links between the unit to form a graph with directed edges along the sequence [70]. Strong prediction performances with the ability to capture variable-length observations and long-term temporal dependencies are the significant attraction properties. Various RNN such as **Gated Recurrent Unit (GRU)** [27], **Long Short Term Memory (LSTM)** [127] has been implemented in the prediction and imputation of the sensed data. To accomplish better results by utilizing the missing patterns and preserving the durable temporal dependencies, GRU effectively incorporates two representations of missing patterns, namely masking and time interval. Back-propagation is used in jointly training all model elements. Xingjian et al. [127] implemented **Convolution LSTM (ConvLSTM)** to predict the missing values in spatiotemporal sequences. ConvLSTM utilizes and creates a trainable end-to-end model for prediction. The results prove the proposed method could preserve the spatiotemporal dependencies with precise results even in high missing gaps. Recently, transferred long short-term memory-based iterative estimation is proposed to impute sensed data by Jun et al. [84] based on the LSTM technique. The method takes benefits from deep learning, transfer learning, and an iterative process. The method learns from the observed data, transfers the information on the missing value, and uses that information on the imputation of the missing data where the iterative method estimates and imputes missing value. The method can secure the long-term moving trend by estimating the values based on observed data. RNN imputation is a versatile method that can be combined with other techniques to produce precise results. It can be deployed in a sequential dataset, used to model a set of records (for example, a time series) so that each pattern is thought to be dependent on the preceding ones. However, due to higher computational complexity, training the RNN is tedious.

*5.2.3 Convolution Neural Network.* The **Convolution Neural Network (CNN)** is designed to process data that are generated in the form of multiple arrays that is specially designed for image analysis [69]. Zhuang et al. [167] used CNN to impute missing data in intelligent transportation systems where traffic data were converted into an image and analyzed. The basic strategy is to integrate the features of the encoder-decoder pipeline and the loss function of **Generative Adversarial Network (GAN)**. Wang et al. [147] investigated to address the incomplete and multisource structure of medical data using CNN and designed a technique to integrate feature clustering to enable the matrix-based representation and CNN for feature extraction and fusion to explicitly exploit data structure from multisource. Similarly, Zhang et al. [160] used the unified spatial-temporal-spectral CNN technique to reconstruct missing data in remote sensing images. The model learns through deep CNN by acquiring a nonlinear end-to-end mapping among incomplete data and intact data with auxiliary data. The model was able to handle multisource data (spectral, spatial, and temporal) containing missing values. Generally, CNN imputation is used in Spatio-temporal image datasets.

*5.2.4 Generative Adversarial Network.* GAN was introduced by Goodfellow et al. [51] as a class of generative models that consists of discriminator and generator. The synthetic data are generated by a generator that is compared by discriminator against the real data. It works by mimicking the distribution of the real data, which is able to generate "real" samples from a random "noise"

learning from the latent distribution of the dataset. GANs provide us with more options for modeling data distribution. Jinsung et al. [156] introduced Generative Adversarial Imputation Nets to impute missing values using the GAN framework. The objective of the generator and discriminator is to precisely impute missing data and discriminate between observed and imputed data, respectively. Classification loss is minimized through training of discriminator, and misclassification rate of the discriminator is maximized by training generator. An adversarial technique is used to train these two networks. However, this method was focused on non-sequential datasets and could not process temporal data. Luo et al. [83] implemented GAN in multivariate temporal datasets by integrating multi-channel CNN and deep convolution GAN. The method imputes missing values robustly, and more accuracy was obtained even when the missing rate increased. GAN imputations are implemented in non-sequential datasets, where the model learns in detail using an adversarial approach from observed data producing highly accurate results. However, it is harder to train due to high computational complexity, and some approaches could not process the temporal data.

### 5.3 Hybrid Imputation

*Assumptions:* Integration of two or more different algorithms.

The most recent development in imputation technology is a hybrid system. Hybrid algorithms are a combination of two or more algorithms. Hybridization is used to overcome deficiencies of a particular algorithm to exploit the benefits of multiple strategies while overcoming their shortcomings to handle in all situations. The hybrid models are designed to trade off efficiency for improved accuracy by strengthening the reduced modeling. Such methods are selected and tested so that the results are better than the methods performed independently [28, 31].

Such integration occurs between SI and MT or SI and MT and classification-based or SI and MT and clustering or IT and classification or IT and clustering or classification and clustering or among deep learning techniques. Michail et al. [28] implemented a hybrid method by combining MICE and KNN. The method performed better than performing independently. A hybrid method, fuzzy c-means, combines support vector regression, and a genetic algorithm was implemented, and the result performed well [11]. Zhongrong et al. [162] implemented a spatiotemporal hybrid imputation method using SOM, fruit fly optimization, and least square SVM. The results demonstrated that the hybrid method was robust and accurate compared to other methods of performing independently. Liu et al. [76] implemented multiple kernel clustering, integration of the kernel method, and the clustering method. The algorithm is claimed to achieve superior performance, verifying the effectiveness, and the improvement is significant with the increase in missing ratio. Junninen et al. [59] compared SI, MI, ML, and hybrid imputation techniques in air quality datasets. The authors concluded that better performance is possible by the hybridization of the multivariate methods.

The Bagging algorithm [9] is designed with the objective to improve accuracy by utilizing block bootstrap techniques and marked point processes. Moving, non-overlapping, and circular block bootstrap techniques along with integer-valued sequences and amplitude modulated series are considered, such as Linear and Stineman interpolations, weighted moving average, and Kalman filters. Datta et al. [35] deployed an integration of both the KNN and the penalized dissimilarity measure with a feature weighting approach to address the problem of the incomplete value, which can be directly applied to datasets containing the missing value without any pre-processing. Imputation based on ensemble approach is deployed in Reference [45] where SOM and KNN were used as a classifier. The integration of powerful modeling ability of deep learning network (LSTM) and flexible transferability of transfer learning (bidirectional imputation) is proposed in Reference [85].

Due to space limit, the comparison of imputation methods highlighting pros, cons, and when to be used in context of IoT is available in supplementary online material Table 4.

## 6 PERFORMANCE INDICATOR

Various performance indicators were designed to delineate the goodness of imputation. Direct evaluation and Classification accuracy are two common techniques used for imputation.

### 6.1 Direct Evaluation

Evaluation of imputation techniques is done based on the predicted and corresponding observed value using the performance indicator in direct evaluation. Some of the commonly implemented approaches are as follows: The coefficient of correlation ($\sigma$) is the most popular indicator of agreement on evaluating imputation values. It explains the variability in imputed data and the amount related to the observed values. It is computed by using Equation (1),

$$\sigma = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\left(P_i - \overline{P}\right)\left(O_i - \overline{O}\right)}{\sigma_P \sigma_O}\right\}, \tag{1}$$

the coefficient of determination ($\sigma^2$) denotes the square of $\sigma$. Value of $\sigma^2$ and $\sigma$ are circumscribed to a range between 0 and 1, values closer to 1 resembles a better fit. The magnitude of the difference between imputed and observed values may not be associated with the values from $\sigma$. Index of Agreement(d2) represents the measure of relative error between observed and imputed data values. The range of d2 lies between 0 and 1, that indicates a lack of agreement and perfect agreement. d2 is analyzed using Equation (2),

$$d2 = 1 - \left[\frac{\sum_{i=1}^{N}\left(P_i - O_i\right)^2}{\sum_{i=1}^{N}\left(\mid P_i - \overline{P} \mid\right) + \left(\mid O_i - \overline{O} \mid\right)}\right]. \tag{2}$$

**Root Mean Square Error (RMSE)** shows the mean error of the model through the difference between observed and imputed concentration. Equation (3) helps to compute RMSE,

$$RMSE = \frac{1}{N}\left[\sum_{i=1}^{N}\left(P_i - O_i\right)\right]^2. \tag{3}$$

**Mean Absolute Error (MAE)** is mean the difference between observed and imputed data points, which is calculated using Equation (4). It makes the comparison of the more sensitive measure of residual error as RMSE,

$$MAE = \frac{1}{N}\sum_{i=1}^{N} \mid P_i - O_i \mid, \tag{4}$$

where N represents the number of imputation, $O_i$ and $P_i$ are the observed and imputed data points, $\overline{O}$ and $\overline{P}$ the average of observed and imputed data, and $\sigma_O$ and $\sigma_P$ are the standard deviation of observed and imputed data.

### 6.2 Classification Accuracy

Classification accuracy is one of the performance indicators to evaluate the imputation quality by some selected classifiers trained by the imputed dataset. The classifiers train and test the classification performance on the imputed dataset, i.e., data without missing value. The results with higher classification accuracy indicate better imputation results. Various classifiers used for classification in the imputed dataset are KNN, ANN, clustering, SVM, MLP, and so on [74].

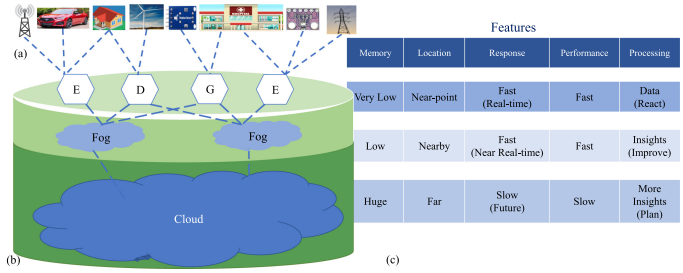


Fig. 4. Illustration of (a) data generation and (b) data collection along with the computing platform in IoT, and (c) highlights the different features along with the comparison of three computing platforms.

### 6.3 Computational Complexity

The computational complexity defines the number of resources required to run the algorithm. The cost associated with the attributes increases along with the augmentation in computational complexity [165]. Every algorithm is efficient in its scenarios; however, reliability is a tremendous concern due to the accuracy of prediction of the missing value can be low in various circumstances, such as time, cost [165].

## 7 COMPUTING PLATFORM

Recently, the majority of the imputation techniques are based on intelligent techniques. However, it also presents severe challenges in terms of scalability, data processing, and computing resources, although these approaches have contributed to developing imputation technology. To comprehend the current issues, it is necessary to describe and examine available strategies for implementing imputation systems. Figure 4 illustrate the various computing platforms along with the features. Multiple computing platforms are highlighted below:

### 7.1 Cloud-based Technique

The cloud computing platforms provide the computational and storage resources using remote servers, where end-users deploying imputation solutions need to connect to them over the Internet to run algorithms to handle missing data. The flexibility of the cloud infrastructures allows providers to dynamically alter storage and compute capacity to meet end-users demands. The cloud techniques are inefficient for real-time applications due to bandwidth, latency, and communication expenses. Aside from this, misconfigured network traffic hampers the cloud-centric technique because of the enormous volume of data. Durham et al. [41] implemented imputation using cloud-based tensor decomposition. Similarly, Bu et al. [25] deployed a feature selection and cluster analysis-based incomplete high-dimensional data imputation algorithm based on cloud. The results show that the suggested approach for imputing high-dimensional data achieves improved imputation accuracy and requires much less time than other algorithms.

### 7.2 Fog-based Technique

Fog computing platforms are distributed computational models that perform data storage, computing, pre-processing, and analysis in a layer between the cloud and the sensing devices. In this regard, the computing ability of imputation is achieved near sensing devices, and clouds for the generation and handling of data [150]. Balasubramanian and Meyyappan [13] implemented Fog-based imputation where the method provided efficient computational performances with higher accuracy in a short time.

### 7.3 Edge-based Technique

Edge computing refers to the decentralized computational infrastructure where data storage and computing resources lie near or close to the base station or end-user. Edge computing helps in data processing at the sensor allowing real-time response, which improves output while also speeding up data processing and reducing bandwidth usage. Recently, there have been numerous attempts to design intelligent techniques that perform imputation at the sensor nodes [6]. Guastella et al. [53] designed real-time imputation using edge-based technology. The model shows augmentation in latency and execution time.

### 7.4 Hybrid Technique

The hybrid approach integrates several architectural models into one. A hybrid platform combines edge, fog, and cloud, or any two of these layers. The information is processed in the source/sensor, while the remaining crucial work is done in the cloud, fog, or mixed solution. When minimal computation costs are required, the algorithms could be implemented on edge and/or in fog; conversely, when high computational costs are required, they could be implemented in the cloud. Imputation using a hybrid platform of cloud and fog is deployed in Reference [125].

## 8 SCOPE/APPLICATION OF IMPUTATION TECHNIQUES IN IOT

This section discusses a variety of missing data applications in IoT. We examine each application area based on the concept of missing data, the structure of data, challenges involved, and the existing approach implemented.

### 8.1 Medical IoT

**Internet of Medical Things (IoMT)** refers to collecting medical equipment, devices, and applications that can communicate in real time or online using computer networks. Missing data in the IoMT domains usually deal with patients' records. IoMT depends on the data collected from such data acquisition tools, sensors, and communication mediums. Medical data consist of categorical, spatial, and temporal aspects and contains various features such as patient age, a test of the blood sugar level, monitoring of the electrocardiogram, heartbeat. Due to various reasons such as human problems (attackers), hardware problems (network device and senors problem) causes serious problems such as missing important information for continuation of those applications. Hence, the strategy to recover the lost information is mandatory in ensuring IoMT systems perform accurately and correctly, providing quality of service to the patient. The most tedious aspects of imputing the missing data in IoMT are computational cost and quality data imputation. Various methods have been implemented to impute the missing data in the medical area [62, 142].

### 8.2 Intelligent Transportation Systems

The integration of communication and computation to control and monitor the transportation network is an ITS. ITS consists of four major components, vehicle subsystem, station subsystem, monitoring system, and security subsystem. All these consist of multiple sensors and actuators. The data generated and collected in ITS belong to spatiotemporal features and provide real-time services such as travel time prediction, monitoring, traveler information to enhance safety, mobility, and efficiency. Furthermore, ITS data are distinctive because of the drastic changes in traffic conditions, sudden speed drops, and time and space interval correlations with neighbors. The "missingness" problem resulted from various factors, including failure of hardware/software, communication issues, undermine the spatiotemporal data resulting from limiting the accuracy. Due to the explicit characteristics, the imputation approach must be chosen cautiously to provide

accurate and reliable data for ITS services. The most tedious aspects of imputing the missing data in ITS are handling large datasets in real-time computational cost and quality data imputation. Various approaches are implemented to impute missing data in ITS [21, 29, 31, 137].

### 8.3 Environmental Monitoring Network

**Environmental Monitoring Network (EMN)** uses various sensing devices to monitor and control various aspects such as air pollution, pressure control system, water quality, and distribution. Epidemiological studies, especially the health effects of water and environmental air pollutants, are no more exception to having complete data. They possess missing data due to data corruption, power outage, equipment failure, sampling error, human error, and so on. Data collected by EMN sensors belong to spatiotemporal data. Various challenges of missing environmental data include significant bias in the system and efficiency degradation, making analysis more complicated. A wide range of research has been conducted to impute the missing data in the environmental and water datasets [2, 59, 101, 107].

### 8.4 Energy Management System

A computer system that automates the monitoring and controlling of associated electromechanical infrastructure consuming a huge amount of energy. Data collected from various sensing devices such as sensors, electricity meters, thermometers are delivered to **Energy Management System (EMS)** to monitor and control multiple electromechanical devices such as electric power transmission, air conditioners, and related energy-saving devices. Environmental factors, low quality or unstable network connections, unreliable communication infrastructure hinders the data collection procedures resulting in incomplete data in EMS. The most tedious aspects of imputing the missing data in EMS are handling large datasets in real time to address the computational cost, complexity, and quality data imputation. Various approaches have been used to recover the lost information [44, 66, 79, 148].

### 8.5 Education

Efficiency of the education system depends on the IoT where data can be collected through Massive Open Online Course using mobile devices, and intelligent analytical techniques can be used to interpret and predict learners' achievement and progress. Data from educational institutions are being used for a wide range of analyses, including improving the educational system and supporting economic decisions. However, due to the structure of the data collection procedure, similar data frequently possesses non-negligible shares of incomplete values, invalidating the above operations. Hence, incomplete information needs to be reconstructed optimally by imputing data of similar nature. Bruni et al. [24] designed an imputation method based on the integration of average and linear regression using a donor for partial numerical sequence reconstruction. The optimal solutions were searched for preserving the global data attributes.

### 8.6 Others

The industries have undergone significant changes to increase productivity that follow four elements: transportation, sensing, processing, and communication. Industrial automation is only possible when IoT is deployed in industrial design to monitor and control the production machines' functionality, operation, and productivity. For efficiency and quality, the decision-making procedure [78] used various approaches to impute missing data in the industrial sector.

In recent years, sports analytics have been emerging swiftly. This plays a vital role in conducting a competitive advantage for players and a team. Analytics and predictions in sports are also essential to track the players' performance, behavior, and so on. Various activity recognition using

intelligent techniques and IoT are investigated in multiple sports [60]. Imputation in physical activity datasets (accelerometer generated datasets) is implemented by Stephens et al. [135].

The presence of the missing data deprives IoT undertake smart decisions and analysis. All the IoT domains and applications suffer from the missing data problem. However, very few applications have only addressed the imputation of the missing data, though the analysis has been performed in all applications. This shows that IoT lacks quality decision and analysis in multiple applications.

## 9 DISCUSSION

Due to multiple reasons such as aging sensor, meteorological extremities, network attacks, device failure, and human error, all the sensor-generated datasets contain missing data. Hence, before undergoing analysis or taking a decision, it is crucial to address the missing data issue. However, most literature in the IoT domain does not acknowledge the missing data before performing analysis. This shows that either the missing data have been deleted or do not contain the missing value. The latter case is almost impossible, which strongly supports the assumption of missing data. Because various factors influence the imputation results, which makes the selection of the appropriate techniques more challenging. Such factors include detection, mechanism, pattern, and distribution of the data, various features present in the dataset, the assumption used, imputation methods, analysis model, number of the missing gap and values, sample size and software used. Due to this, various research persists involving commonly used imputation, such as mean imputation and deletion methods, not testing the assumptions assumed, not describing the approach used in dealing with missing data, ignoring sensitivity tests [74]. Deploying imputation algorithm in the computing platform, Edge/Fog is used for real-time and near real-time imputation, suitable for point and short missing gap problems due to dependency in the historical dataset. The high missing gap of the accumulated data is better estimated through a cloud-based platform. Hence, it is essential to perform imputation on a historical dataset prior to real-time-based imputation for the accurate and reliable prediction.

Generally, deletion approaches are used when there is less than 5% of the missing data, under the assumption that loss of power and biases are probable to be inconsequential. However, it is strongly recommended to implement imputation to preserve data structure and save information. For example, in time-series analysis, the sequential structure is broken, creating complications such as determining the seasonal pattern, because deletion can result in unethical selection bias and poor performance. The only solution to address missing data is imputation, which requires special attention as the imputation approach alters the data distribution, it has an impact on the optimization problem's solution. Similarly, imputations would be insufficient [17] based on the following:

- When the prediction models are able to de-impute using certainty leading to consistent/compatible predictions.
- Missing data in training and testing data should be imputed with the same approach.
- Missingness should be represented as a different class for discrete or categorical variables, and mean imputation can produce compatible predictions for continuous variables.
- The missing mechanism should be investigated and while dealing with various patterns of missing data in testing and training, also called "distributional shifts," remains an open research question.

### 9.1 Selection of Appropriate Method in IoT

There are two datasets for performing imputations: publicly available datasets and real-world datasets. Based on the features available, IoT datasets possesses numerical, categorical or mixed

characteristics. The performance of the imputation algorithm differs based on the dataset used, missing gaps, and the missingness mechanism. In IoT, all the missing mechanisms (MCAR, MAR, and MNAR) can occur depending upon the scenario. Most of the literature considers only one mechanism to perform the simulation, which is not enough to comprehend the performance of the specific approach. Feature selection before imputation can help achieve better imputation results, whereas after imputation help classifiers perform better compared to early feature selection. There is no exact rule to claim the practical missing rates to be implemented, however, imputations of missing data on high missing rate with different missing gaps (60%), very/high missing gap (e.g., continuous missing gap of 21%), a broad range of missing rate (e.g., 5% to 60%), a mixture of high missing gap and high missing rate, is considered more practical in IoT. Such simulations require complex imputation methods to address the missing gap. Most of the data generated by IoT belong to time series, and imputation can be performed both by time-series and pattern-based approaches. It is to be noted that the pattern-based methods can handle time-series datasets; however, sometimes it might lead to a decrease in the accuracy of the predicted models. The decision to use a time-series or pattern-based method in time-series dataset depends on the missing gap and missing ratio present in the dataset. When the missing gap is low, time-series imputation techniques might handle the case efficiently. If the missing gap and missing ratio are high, then the pattern-based method performs better. Similarly, in the presence of a very high missing gap, integration of time-series and pattern-based method gives better results. However, one should be cautious in using the imputation model, because the wrong imputation leads to anomalies in the data.

There exist numerous imputation techniques that are classified as statistical and intelligent learning-based approaches. The previous section highlights various approaches regarding those imputations. However, to date, there is no comparison between those techniques based on multiple mechanisms containing various missing gaps and missing rates in multiple domain datasets. Similarly, multiple performance indicators such as direct evaluation, classification accuracy, and computational complexities are used to validate the performance of the imputation approach. However, most studies only perform one or two indicators to evaluate the imputation techniques, which lacks the understanding to provide a better imputation strategy. Such limitations prohibit determining the best appropriate techniques for domain-based imputation techniques.

*9.1.1 Selection of Statistical Techniques.* The adoption of SI is computationally simple, fast, and inexpensive and requires less memory than MI. In the meantime, it is problematic, because it creates substantial bias, underestimates variance, uncertainty, standard error, a correlation between variables gets affected, analysis becomes more sensitive, and so on. In the presence of point and short missing gap, SI provides better results compared to MI [58, 59, 99]. However, when the gap increases, accuracy decreases. Authors claim that MI performs better than SI [58, 59]. Multiple Imputation analyses many imputation samples to take the uncertainty of imputed values into account, properly amends the major disadvantages, and retains the advantages of SI through the addition of error based on variation in estimating a parameter across imputation known as "between imputation error" [130]. Thus, MI is computationally complex and expensive (computational cost scales linearly as a function of the number of imputations to be performed) and is considered the advanced version of SI with the uncertainty of the imputed values taken into account.

Selection between MT and IT is tedious as both are based on the likelihood function. Model-based Technique is relatively faster and computationally simple compared with IT [40]. Despite the various benefits of MT, IT is still attractive and widespread over the MT-based methods due to computational flexibilities, such as many models can fit/ estimate the data after imputation. Fitting models after imputation does not require repeated actions. Auxiliary variables can easily be included in the IT methods than in MT-based methods. On handling the categorical variables,

IT methods perform better than MT-based methods. The comparison between various MT- and IT-based techniques shows that MLI techniques are superior [77, 91]. Similar results were demonstrated in References [7, 23]. Yet, some research results show that IT and MLI-based methods are equivalent [40, 50]. The MLI-based method is a better choice when the data analysts are clear about the parameter to be estimated as they do not need to introduce randomness on data. Nonetheless, MT is also a better choice if the data analysts are clear about the relationship between the data and the imputation model.

*9.1.2 Selection of Intelligent Techniques.* In the specific case of classification, learning from the data containing missing values becomes more crucial. Due to the presence of missing data, most classification algorithms cannot work directly. Intelligent imputation techniques are the most sophisticated procedures, consisting of the generation of the predictive models to estimate the value to replace the missing value. It is performed by using the training and testing of the datasets through modeling the incomplete data estimation based on the available information in the datasets. It is essential to train a classifier utilizing the imputed training set. Missing values in either testing or training set or both sets affect the prediction accuracy of the learned classifier. Handling of missing data using intelligent techniques, two scenarios get distinguished:

- Complete training datasets, missing value in test data.
- Both test and training datasets contain a missing value.

In the first case, training datasets are complete; thus, no assumptions are made in the training datasets. Thus, it reflects that either missing values are excluded from the datasets or the missing value has been imputed using some estimations. This is often essential to enable training models. In the second case, the testing and training datasets are gathered and processed similarly, and the classifier is trained considering incomplete input vectors.

Intelligence-based techniques are more computationally complex and expensive than statistical (or statistics)-based approaches [146]. The cost of intelligent techniques increases with the training and testing of datasets and the formation of clusters or hidden networks [146]. In intelligent-based techniques, deep learning techniques possess higher computational complexity than the simpler machine learning-based approaches. Novel techniques adopt more complex procedures that require high computational effort; yet, the efficiency and quality of the imputation increases. Though intelligent techniques are computationally complex, they are the first choice because of the efficient results. However, one should be aware about the missing gaps and mechanism because on the presence of point and 3/4 missing gap, methods such as interpolation or nearest neighbor performs better than others.

## 10 RESEARCH CHALLENGES, OPEN ISSUES, AND NEW PERSPECTIVES

After highlighting numerous imputation strategies and discussions, it is important to discuss various issues, challenges, and new perspectives to improve the imputation technology in IoT domains. In IoT systems, the imputation of missing data opens up a wide range of options, presenting practical implementation, challenges, and issues. Integration of Cloud-Fog-Edge environment with intelligent techniques generates huge potentials to intrigue the researcher and analyst. It also demands computational power, which contrasts sharply with the significant hardware and software limits of IoT sensors.

### 10.1 Research Challenges and Open Issues

*10.1.1 Essence of Real-life Datasets.* There lacks public standard and real-world datasets essential for the imputation approach to assess the efficacy and reliability of new technologies.

*10.1.2 Robust Imputation Techniques.* Deployment of intelligent techniques in the hardware/sensor nodes require high performance computing services. Thus, more robust, efficient, and reliable algorithms based on various computing platforms for handling the incomplete dataset is of crucial requirement.

*10.1.3 Power Efficiency.* Sensing devices are operated through battery or are dependent on energy-harvesting, creating a limitation of power. Similarly, it is tedious to obtain the optimal trade-off between energy consumption and algorithmic complexity in edge-enabled devices. Imputation in real time requires high energy consumption in sensing devices due to where power drainage and memory limitations. Thus, efficient energy remains a hugely unsolved issue.

*10.1.4 Accuracy and Reliability.* There exist multiple strategies for imputing the missing data; however, reliability and accuracy of the imputed data remain as challenges. As the missing gap and amount increase, imputed data become unreliable, and, consequently, the accuracy and reliability of the data degrade. Sometimes accuracy is obtained through high computational cost and processing time. Intelligent techniques and multiple imputations help to achieve accuracy and reliability.

*10.1.5 Architectural Design.* Traditional imputation techniques are not designed to operate in the sensor system, especially because of their restricted connectivity, memory, and processing ability. The majority of existing architectures are incapable of handling real-time imputation. Real-time design components must integrate application and analytics to present a new approach to a working environment that meets the accuracy and quick response needs. Such architectural difficulties will be addressed by combining various data technologies with intelligent learning strategies that are flexible to adapt to architectural changes, such as switching from edge to cloud or vice versa, which is a tedious task. Managing resources dynamically is crucial, however, diversity and heterogeneity of frameworks and limitations of the edge components lack application of IoT and computing services offer considerable open challenges and issues.

*10.1.6 Computational Cost and Complexity reduction.* The computational complexity defines the number of resources required to run the algorithm. The cost associated with the attributes increases along with the augmentation in computational complexity [165]. Novel techniques adopt more complex and computational efforts for improving accuracy and efficiency by addressing various issues and challenges. Implementing Cloud/Fog/Edge will alleviate the computational complexity and cost by adding parallel and distributed processing.

*10.1.7 Real-time vs. Offline Imputation.* Efficient and reliable imputation on real-time and streaming data poses an emerging essence due to the enhancement of complex sensor-based systems and IoT. The advancement of software, hardware, and computing resources allows real-time applications to handle streaming data and demanding constraints. Shifting from conventional offline data processing and analytic to real-time methods is a result of this circumstance. Real-time processing has restricted computational resources as well as time limits for producing a solution. In addition, the input data must be examined upon its arrival for further real-time processing. Real-time techniques can be adapted based on historical measurements (short-term memory), windowing, frequent model updating, and so on. Offline processing, however, is involved in evaluating the entire dataset, where all the necessary data has been collected and is readily available. Because there exist no time or computational limits (assuming the computation takes place in a capable environment like the Cloud), this technique usually enables high-complexity approaches to be used.

Algorithms, whether online or offline, must be used according to system requirements. When dealing with IoT data, near-real-time or real-time approaches are better. When additional data are available and an immediate response is not required, offline algorithms enable the execution of

more complex jobs on vigorous resources. On creating and implementing these algorithms, trade-offs among processing time, computational cost and energy, response time, and performance must be considered.

*10.1.8 Sensitivity Analysis.* The sensitivity analysis study determines how the uncertainty in the output model might be attributed to various sources of uncertainty in the input model. Supplementary conditions on the reasons for incomplete data are made during analysis, though these assumptions cannot be validated for correctness. MAR approach is based on a pattern mixture model, the mixture of distribution of the missing response, and distribution of the observed response [157]. Though the emphasis is on IT, the standard error can be corrected using a single imputed dataset for a limited set of variables such as correlation coefficients, means, and proportion. A method for computing correct standard error is known as the Jackknife method is proposed in Reference [12].

*10.1.9 Proper Imputation Model.* The models can be classified into an ignorable or non-ignorable models. The missing data mechanism is ignorable if (a) data are MAR and (b) parameters guiding the missing data process are not related to the parameter to be estimated. There is no direct proof in the data to address the accuracy of any such assumptions, a good reason to consider several models and explore resultant sensitivity wherever possible. Treating MAR and ignorability as equivalent also work, but that could be done better by modeling the missing data mechanism. When data are in NMAR, it is said the mechanism of missing data is non-ignorable. In this situation, to get a reasonable estimate of the parameter of interest, we model the missing data mechanism. The imputation model should contain two assets: First, the imputation model should contain useful variables where chances of increasing the variance of estimates or leading to non-convergence increases while including too many variables. Generally, three kinds of variables are included in the imputation model [117]: (a) variables having theoretical importance, (b) variables with missing data mechanism, and (c) variable correlating with a variable having missing data. Second, the imputation model should be universal enough to preserve the consequences of concerns in the data structure. However, it is not necessary to have a fundamental scientific theory on the imputation model [7].

When the imputation model is more restrictive than the analysis model [40], one of the two consequences might occur. The first consequence is that when additional restrictions are true, the results are valid, but the conclusion is conservative, failing to reject the false null hypothesis. The second consequence is when one or more restrictions are not reasonable. The results are invalid, restricting the relationship between a variable and another variable in the imputation model. Therefore, the consequences of any interaction that associate at least three variables will be biased toward zero.

## 10.2 New Perspectives

Recently, many researchers have focused on implementing imputation strategies in various IoT applications for real-time monitoring and automation of the system, which helps improve the quality of the data essential for better decision-making. In this subsection, we highlight some of the new perspectives in the imputation of the incomplete data.

*10.2.1 Edge-based Computing.* Incomplete data in IoT applications require real-time monitoring with fast response time and quick processing. The promising solution for handling incomplete data lies in edge computing and the integration of intelligent techniques. Using edge computing, incomplete data handling takes place at the sensor node, and the decisions are transmitted to the server. This helps to reduce the data transmission, and power consumption is low, however,

effective optimization strategies are essential to sustain the longevity of the edge devices. Ivan et al. [82] implemented a framework for edge management to recover the missing values using limited resources for accurate decision-making.

*10.2.2 Visualization.* Visualization of handling missing data helps analysts draw decisions regarding the quality of the data and the effective reasoning essential for drawing conclusions. Processed and analyzed data provides clear insights on using the visualizations approach of the data, whereas incomplete data produces the biggest challenges. In designing incomplete data visualization, multiple visualization strategies ranging from simple line graphs, bar charts to 2D and 3D are used. Visualizing the incomplete and imputed data also helps an analyst reduce the biasness, strengthen the data quality, perceived confidence, and accuracy. Song and Szafir [131] implemented visualization technique. The obtained results show that visualization of the imputed data points achieved the highest data quality and confidence in preserving the continuity of the data.

*10.2.3 Deep Reinforcement Learning.* **Reinforcement Learning (RL)** is an emerging topic in ML where agents discover appropriate actions to maximize a reward in a given environment. Intelligence to RL agents is integrated with the DL to enhance the ability to optimize, making it proficient in solving complex computational tasks, such as the complex pattern of the missing data, known as **Deep Reinforcement Learning (DRL)** [143]. Huang et al. [57] deployed DRL in the handling of incomplete data proving its efficiency and reliability. DRL is a promising technique showing opportunities to handle incomplete data efficiently and effectively.

*10.2.4 Deep Ensemble Learning.* Ensemble strategy is the fusion of multiple trained models integrating predictions to enhance the performance of the single model. These meta-algorithms are promising on decreasing bias (boosting), or variance (bagging), or augmenting predictions (stacking). Generally, ensembles methods are used when there exists multiple missing data assumptions Integration of ensemble and DL forms deep ensemble learning. Sun et al. [136] designed an online framework using an ensemble-learning strategy for the imputation of data in an offshore wind farm. The method performed efficiently, effectively using less computational time.

*10.2.5 Imputation Using Explainable Artificial Intelligence.* Explainable models are deployed in AI to assure accountability, trustworthiness, and transparency. The quality of the data and the algorithm implemented determine the accuracy and reliability of the explanations. The inevitability of the missing data that exist in real-world datasets affects the quality of the explanations. Ahmed et al. [4] explores various cases on dealing with the explainable AI on the handling of missing dataset.

## 11 CONCLUSION

This article investigates various reasons for missing data in IoT architecture, classifying the various imputation techniques and their advantages and shortcomings. Numerous applications of IoT where imputation algorithms can be implemented were highlighted. Multiple computing platforms based on the architectural model for handling missing data were discussed in brief. To ensure the precise decision-making of intelligent systems, it is essential to deal with the missing data timely. Before applying any imputation techniques, one should make the best effort to know the reason for missing and prevent minimum in number by designing and implementing intelligent tools during data collection. Similarly, different open issues and challenges related to the imputation techniques and selection for appropriate imputation methods in the IoT area were also addressed. Finally, some future research directions to overcome the limitations of imputation techniques and enhance the adoption of imputation techniques in the real-world context based on low energy consumptions, scalability, easy deployment, and decentralization were also suggested.

## REFERENCES


[1] Najmeh Abiri, Björn Linse, Patrik Edén, and Mattias Ohlsson. 2019. Establishing strong imputation performance of a denoising autoencoder in a wide range of missing data problems. *Neurocomputing* 365 (2019), 137–146.

[2] Deepak Adhikari, Wei Jiang, and Jinyu Zhan. 2021. Imputation using information fusion technique for sensor generated incomplete data with high missing gap. *Microprocess. Microsyst.* (2021), 103636.

[3] Deepak Adhikari, Wei Jiang, and Jinyu Zhan. 2021. Iterative imputation using ratio-based imputation for high missing gap. In *Proceedings of the International Conference on Intelligent Technology and Embedded Systems (ICITES'21)*. 1–6.

[4] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. 2019. The Challenge of Imputation in Explainable Artificial Intelligence Models. *CoRR* abs/1907.12669 (2019). arXiv:1907.12669.

[5] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash. 2015. Internet of Things: A survey on enabling technologies, protocols, and applications. *IEEE Commun. Surv. Tutor.* 17, 4 (2015), 2347–2376.

[6] Md Golam Rabiul Alam, Mohammad Mehedi Hassan, Md. ZIa Uddin, Ahmad Almogren, and Giancarlo Fortino. 2019. Autonomic computation offloading in mobile edge for IoT applications. *Fut. Gener. Comput. Syst.* 90 (2019), 149–157.

[7] Paul D. Allison. 2012. Handling missing data by maximum likelihood. In *SAS Global Forum*, Vol. 2012. Statistical Horizons Haverford, PA, 1038–21.

[8] Mehran Amiri and Richard Jensen. 2016. Missing data imputation using fuzzy-rough methods. *Neurocomputing* 205 (2016), 152–164.

[9] Agung Andiojaya and Haydar Demirhan. 2019. A bagging algorithm for the imputation of missing values in time series. *Expert Syst. Appl.* 129 (2019), 10–26.

[10] Rebecca R. Andridge and Roderick J. A. Little. 2010. A review of hot deck imputation for survey non-response. *Int. Stat. Rev.* 78, 1 (2010), 40–64.

[11] Ibrahim Berkan Aydilek and Ahmet Arslan. 2013. A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Inf. Sci.* 233 (2013), 25–35.

[12] Melissa J. Azur, Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. 2011. Multiple imputation by chained equations: What is it and how does it work? *Int. J. Methods Psychiatr. Res.* 20, 1 (2011), 40–49.

[13] S. Balasubramanian and T. Meyyappan. 2019. Enhancing the computational intelligence of smart fog gateway with boundary-constrained dynamic time warping based imputation and data reduction. In *Proceedings of the 3rd International Conference on Imaging, Signal Processing and Communication (ICISPC'19)*. 15–23. https://doi.org/10.1109/ICISPC.2019.8935698

[14] Xavier Basagaña, Jose Barrera-Gómez, Marta Benet, Josep M. Antó, and Judith Garcia-Aymerich. 2013. A framework for multiple imputation in cluster analysis. *Am. J. Epidemiol.* 177, 7 (2013), 718–725.

[15] Mohamed-Aymen Ben Aissia, Fateh Chebana, and Taha B. M. J. Ouarda. 2017. Multivariate missing data in hydrology–Review and applications. *Adv. Water Resour.* 110 (2017), 299–309.

[16] Michael R. Berthold and Klaus-Peter Huber. 1998. Missing Values and learning of fuzzy rules. *Int. J. Uncertain. Fuzz. Knowl.-Bas. Syst.* 6, 2 (April 1998), 171–178.

[17] Dimitris Bertsimas, Arthur Delarue, and Jean Pauphilet. 2021. Prediction with missing data. arXiv:2104.03158 [stat.ML].

[18] James C. Bezdek. 2013. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Springer Science & Business Media.

[19] Christopher M. Bishop et al. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press.

[20] Y. Boiko, C. Lin, I. Kiringa, and T. Yeap. 2019. Navigational data imputation with GPS pinning in compositional Kalman filter for IoT systems. In *Proceedings of the IEEE International Symposium on Robotic and Sensors Environments (ROSE'19)*. 1–7.

[21] Guillem Boquet, Antoni Morell, Javier Serrano, and Jose Lopez Vicario. 2020. A variational autoencoder solution for road traffic forecasting systems: Missing data imputation, dimension reduction, model selection and anomaly detection. *Transport. Res. C: Emerg. Technol.* 115 (2020), 102622.

[22] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. 2015. *Time Series Analysis: Forecasting and Control*. John Wiley &. Sons.

[23] P. M. T. Broersen and R. Bos. 2006. Time-series analysis if data are randomly missing. *IEEE Trans. Instrum. Meas.* 55, 1 (2006), 79–84.

[24] Renato Bruni, Cinzia Daraio, and Davide Aureli. 2021. Imputation techniques for the reconstruction of missing interconnected data from higher Educational Institutions. *Knowl.-Bas. Syst.* 212 (2021), 106512.

[25] Fanyu Bu, Zhikui Chen, Qingchen Zhang, and Laurence T. Yang. 2016. Incomplete high-dimensional data imputation algorithm using feature selection and clustering analysis on cloud. *J. Supercomput.* 72, 8 (2016), 2977–2990.

[26] S. van Buuren and Karin Groothuis-Oudshoorn. 2010. Mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* (2010), 1–68.

[27] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Sci. Rep.* 8, 1 (2018), 1–12.
[28] Michail Cheliotis, Christos Gkerekos, Iraklis Lazakis, and Gerasimos Theotokatos. 2019. A novel data condition and performance hybrid imputation method for energy efficient operations of marine systems. *Ocean Eng.* 188 (2019), 106220.
[29] C. Chen, S. Jiao, S. Zhang, W. Liu, L. Feng, and Y. Wang. 2018. TripImputor: Real-time imputing taxi trip purpose leveraging multi-sourced urban data. *IEEE Trans. Intell. Transport. Syst.* 19, 10 (2018), 3292–3304.
[30] Jiahua Chen and Jun Shao. 2000. Nearest neighbor imputation for survey data. *J. Offic. Stat.* 16, 2 (2000), 113–131.
[31] Xinyu Chen, Zhaocheng He, and Lijun Sun. 2019. A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation. *Transport. Res. C: Emerg. Technol.* 98 (2019), 73–84.
[32] Ching-Hsue Cheng, Chia-Pang Chan, and Yu-Jheng Sheu. 2019. A novel purity-based k nearest neighbors imputation method and its application in financial distress prediction. *Eng. Appl. Artif. Intell.* 81 (2019), 283–299.
[33] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Mach. Learn.* 20 (1995), 273–297.
[34] MIT Critical Data and M. Komorowski. 2016. *Secondary Analysis of Electronic Health Records*. Springer.
[35] Shounak Datta, Debaleena Misra, and Swagatam Das. 2016. A feature weighted penalty based dissimilarity measure for k-nearest neighbor classification with missing features. *Pattern Recogn. Lett.* 80 (2016), 231–237.
[36] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. Ser. B (Methodol.)* 39, 1 (1977), 1–22.
[37] Yaohui Ding and Arun Ross. 2012. A comparison of imputation methods for handling missing scores in biometric fusion. *Pattern Recogn.* 45, 3 (2012), 919–933.
[38] Zengyu Ding, Gang Mei, Salvatore Cuomo, Yixuan Li, and Nengxiong Xu. 2020. Comparison of estimating missing values in IoT time series data using different interpolation algorithms. *Int. J. Parallel Program.* 48 (2020), 534–548.
[39] J. K. Dixon. 1979. Pattern recognition with partly missing data. *IEEE Trans. Syst. Man Cybernet.* 9, 10 (1979), 617–621.
[40] Yiran Dong and Chao-Ying Joanne Peng. 2013. Principled missing data methods for researchers. *SpringerPlus* 2, 1 (2013), 222.
[41] Timothy J. Durham, Maxwell W. Libbrecht, J. Jeffry Howbert, Jeff Bilmes, and William Stafford Noble. 2018. PREDICTD parallel epigenomics data imputation with cloud-based tensor decomposition. *Nat. Commun.* 9, 1 (2018), 1–15.
[42] Craig K. Enders. 2010. *Applied Missing Data Analysis*. Guilford Press.
[43] A. Farhangfar, L. A. Kurgan, and W. Pedrycz. 2007. A novel framework for imputation of missing values in databases. *IEEE Trans. Syst. Man. Cybernet. A Syst. Hum.* 37, 5 (2007), 692–709.
[44] B. Fekade, T. Maksymyuk, M. Kyryk, and M. Jo. 2018. Probabilistic recovery of incomplete sensed data in IoT. *IEEE IoT J.* 5, 4 (2018), 2282–2292.
[45] Zhun ga Liu, Quan Pan, Jean Dezert, and Arnaud Martin. 2016. Adaptive imputation of missing values for incomplete pattern classification. *Pattern Recogn.* 52 (2016), 85–95.
[46] Pedro J. García-Laencina, José-Luis Sancho-Gómez, and Aníbal R. Figueiras-Vidal. 2010. Pattern classification with missing data: A review. *Neur. Comput. Appl.* 19, 2 (2010), 263–282.
[47] Chandan Gautam and Vadlamani Ravi. 2015. Counter propagation auto-associative neural network based data imputation. *Inf. Sci.* 325 (2015), 288–299.
[48] Andrew Gelman and Jennifer Hill. 2006. Chapter 25 on missing data imputation. In *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.
[49] Victor Gomez, Agustin Maravall, and Danie Pena. 1999. Missing observations in ARIMA models: Skipping approach versus additive outlier approach. *J. Econometr.* 88, 2 (1999), 341–363.
[50] M. P. Gómez-Carracedo, J. M. Andrade, P. López-Mahía, S. Muniategui, and D. Prada. 2014. A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. *Chemometr. Intell. Lab. Syst.* 134 (2014), 23–33.
[51] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2672–2680.
[52] John W. Graham. 2009. Missing data analysis: Making it work in the real world. *Annu. Rev. Psychol.* 60 (2009), 549–576.
[53] Davide Andrea Guastella, Guilhem Marcillaud, and Cesare Valenti. 2021. Edge-based missing data imputation in large-scale environments. *Information* 12, 5 (2021).
[54] James Honaker and Gary King. 2010. What to do about missing values in time-series cross-section data. *Am. J. Pol. Sci.* 54, 2 (2010), 561–581.
[55] James Honaker, Gary King, Matthew Blackwell, et al. 2011. Amelia II: A program for missing data. *J. Stat. Softw.* 45, 7 (2011), 1–47.
[56] Feng Honghai, Chen Guoshun, Yin Cheng, Yang Bingru, and Chen Yumei. 2005. A SVM regression based approach to filling in missing values. In *Proceedings of the International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 581–587.

[57] Xiaoshui Huang, Fujin Zhu, Lois Holloway, and Ali Haidar. 2020. Causal discovery from incomplete data using an encoder and reinforcement learning. *CoRR* abs/2006.05554 (2020).
[58] W. L. Junger and A. [Ponce de Leon]. 2015. Imputation of missing data in time series for air pollutants. *Atmos. Environ.* 102 (2015), 96–104.
[59] Heikki Junninen, Harri Niska, Kari Tuppurainen, Juhani Ruuskanen, and Mikko Kolehmainen. 2004. Methods for imputation of missing values in air quality data sets. *Atmos. Environ.* 38, 18 (2004), 2895–2907.
[60] Thomas Kautz, Benjamin H. Groh, Julius Hannink, Ulf Jensen, Holger Strubberg, and Bjoern M. Eskofier. 2017. Activity recognition in beach volleyball using a deep convolutional neural network. *Data Min. Knowl. Discov.* (2017).
[61] J. M. Keller, M. R. Gray, and J. A. Givens. 1985. A fuzzy K-nearest neighbor algorithm. *IEEE Trans. Syst. Man Cybernet.* SMC-15, 4 (1985), 580–585.
[62] H. A. Khorshidi, M. Kirley, and U. Aickelin. 2020. Machine learning with incomplete datasets using multi-objective optimization models. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'20)*. 1–8.
[63] Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational Bayes. arXiv:1312.6114 [stat.ML].
[64] Marietta Kokla, Jyrki Virtanen, Marjukka Kolehmainen, Jussi Paananen, and Kati Hanhineva. 2019. Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: A comparative study. *BMC Bioinform.* 20, 1 (2019), 1–11.
[65] Gueorgi Kossinets. 2006. Effects of missing data in social networks. *Soc. Netw.* 28, 3 (2006), 247–268.
[66] J. Krstulovic, V. Miranda, A. J. A. Simões Costa, and J. Pereira. 2013. Towards an auto-associative topology state estimator. *IEEE Trans. Power Syst.* 28, 3 (2013), 3311–3318.
[67] İbrahim Kök and Suat Özdemir. 2021. DeepMDP: A novel deep-learning-based missing data prediction protocol for IoT. *IEEE IoT J.* 8, 1 (2021), 232–243. https://doi.org/10.1109/JIOT.2020.3003922
[68] Qiujun Lan, Xuqing Xu, Haojie Ma, and Gang Li. 2020. Multivariable data imputation for the analysis of incomplete credit data. *Expert Syst. Appl.* 141 (2020), 112926.
[69] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 13 (2015), 436–444.
[70] Collins Achepsah Leke and Tshilidzi Marwala. 2019. *Deep Learning and Missing Data in Engineering Systems*. Springer.
[71] Dan Li, Jitender Deogun, William Spaulding, and Bill Shuart. 2004. Towards missing data imputation: A study of fuzzy k-means clustering method. In *Proceedings of the International Conference on Rough Sets and Current Trends in Computing*. Springer, 573–579.
[72] Linchao Li, Bowen Du, Yonggang Wang, Lingqiao Qin, and Huachun Tan. 2020. Estimation of missing values in heterogeneous traffic data: Application of multimodal deep learning model. *Knowl.-Bas. Syst.* 194 (2020), 105592.
[73] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao. 2017. A survey on Internet of Things: Architecture, enabling technologies, security and privacy, and applications. *IEEE IoT J.* 4, 5 (2017), 1125–1142.
[74] Wei-Chao Lin and Chih-Fong Tsai. 2020. Missing value imputation: A review and analysis of the literature (2006–2017). *Artif. Intell. Rev.* 53, 2 (2020), 1487–1509.
[75] Roderick J. A. Little and Donald B. Rubin. 2019. *Statistical Analysis with Missing Data*. Vol. 793. John Wiley &. Sons.
[76] X. Liu, X. Zhu, M. Li, L. Wang, E. Zhu, T. Liu, M. Kloft, D. Shen, J. Yin, and W. Gao. 2020. Multiple kernel *k*k-means with incomplete kernels. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 5 (2020), 1191–1204.
[77] Yushan Liu and Steven D. Brown. 2013. Comparison of five iterative imputation methods for multivariate classification. *Chemometr. Intell. Lab. Syst.* 120 (2013), 106–115.
[78] Yuehua Liu, Tharam Dillon, Wenjin Yu, Wenny Rahayu, and Fahed Mostafa. 2020. Missing value imputation for industrial IoT sensor data with large gaps. *IEEE Internet of Things Journal* 7, 8 (2020), 6855–6867.
[79] C. Lu and Y. Mei. 2018. An imputation method for missing data based on an extreme learning machine auto-encoder. *IEEE Access* 6 (2018), 52930–52935.
[80] Julián Luengo, Salvador García, and Francisco Herrera. 2012. On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowl. Inf. Syst.* 32, 1 (2012), 77–108.
[81] Julián Luengo, Salvador García, and Francisco Herrera. 2010. A study on the use of imputation methods for experimentation with radial basis function network classifiers handling missing attribute values: The good synergy between RBFNs and EventCovering method. *Neural Netw.* 23, 3 (2010), 406–418.
[82] Ivan Lujic, Vincenzo De Maio, and Ivona Brandic. 2020. Resilient edge data management framework. *IEEE Trans. Serv. Comput.* 13, 4 (2020), 663–674. https://doi.org/10.1109/TSC.2019.2962016
[83] Yonghong Luo, Xiangrui Cai, Ying ZHANG, Jun Xu, and Yuan xiaojie. 2018. Multivariate time series imputation with generative adversarial networks. In *Advances in Neural Information Processing Systems*, Vol. 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, Inc., 1596–1607.
[84] Jun Ma, Jack C. P. Cheng, Yuexiong Ding, Changqing Lin, Feifeng Jiang, Mingzhu Wang, and Chong Zhai. 2020. Transfer learning for long-interval consecutive missing values imputation without external features in air pollution time series. *Advanced Engineering Informatics* 44 (2020), 101092.

[85] Jun Ma, Jack C. P. Cheng, Feifeng Jiang, Weiwei Chen, Mingzhu Wang, and Chong Zhai. 2020. A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data. *Energy Build.* 216 (2020), 109941.

[86] Zhihua Ma and Guanghui Chen. 2018. Bayesian methods for dealing with missing data problems. *J. Kor. Stat. Soc.* 47, 3 (2018), 297–313.

[87] M. Mardani, G. Mateos, and G. B. Giannakis. 2015. Subspace learning and imputation for streaming big data matrices and tensors. *IEEE Trans. Sign. Process.* 63, 10 (2015), 2663–2677.

[88] Lim Kian Ming, Loo Chu Kiong, and Lim Way Soong. 2011. Autonomous and deterministic supervised fuzzy clustering with data imputation capabilities. *Appl. Soft Comput.* 11, 1 (2011), 1117–1125.

[89] Ho MingKang and Fadhilah Yusof. 2012. Application of self-organizing map (SOM) in missing daily rainfall data in Malaysia. *Int. J. Comput. Appl.* 48, 5 (June 2012), 23–28.

[90] V. Miranda, J. Krstulovic, H. Keko, C. Moreira, and J. Pereira. 2012. Reconstructing missing data in state estimation with autoencoders. *IEEE Trans. Power Syst.* 27, 2 (2012), 604–611.

[91] Juan Javier Miró, Vicente Caselles, and María José Estrela. 2017. Multiple imputation of rainfall missing data in the Iberian Mediterranean context. *Atmos. Res.* 197 (2017), 313–330.

[92] Andriy Mnih and Russ R. Salakhutdinov. 2008. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*. 1257–1264.

[93] Philip R. C. Nelson, Paul A. Taylor, and John F. MacGregor. 1996. Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemometr. Intell. Lab. Syst.* 35, 1 (1996), 45–65.

[94] S. Nikfalazar, C. H. Yeh, S. Bedingfield, and H. A. Khorshidi. 2017. A new iterative fuzzy clustering algorithm for multiple imputation of missing data. In *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'17)*.

[95] S. Nikfalazar, C. H. Yeh, S. Bedingfield, and H. A. Khorshidi. 2020. Missing data imputation using decision trees and fuzzy clustering with iterative learning. *Knowl. Inf. Syst.* 62, 6 (2020), 1–19.

[96] Kancherla Jonah Nishanth and Vadlamani Ravi. 2016. Probabilistic neural network based categorical data imputation. *Neurocomputing* 218 (2016), 17–25.

[97] Y. Nishimura, K. Sudoh, G. Neubig, and S. Nakamura. 2020. Multi-source neural machine translation with missing data. *IEEE/ACM Trans. Aud. Speech Lang. Process.* 28 (2020), 569–580. https://doi.org/10.1109/TASLP.2019.2959224

[98] Mohamed Noor Norazian, Ahmad Shukri, Prof Yahaya, Nor Azam, Prof Ramli, Noor Faizah Fitri, Md Yusof, and Abdullah Mohd Mustafa Al Bakri. 2013. Roles of imputation methods for filling the missing values: A review. *Adv. Environ. Biol.* 7 (1 2013), 3861–3869.

[99] Mohamed Noor Norazian, Yahaya Ahmad Shukri, Ramli Nor Azam, and Abdullah Mohd Mustafa Al Bakri. 2008. Estimation of missing values in air pollution data using single imputation techniques. *Science Asia* 34, 3 (2008), 341–345.

[100] H. Reed Ogrosky, Samuel N. Stechmann, Nan Chen, and Andrew J. Majda. 2019. Singular spectrum analysis with conditional predictions for real-time state estimation and forecasting. *Geophys. Res. Lett.* 46, 3 (2019), 1851–1860.

[101] M. S. Osman, A. M. Abu-Mahfouz, and P. R. Page. 2018. A survey on data imputation techniques: Water distribution system as a use case. *IEEE Access* 6 (2018), 63279–63291.

[102] A. Otgonbayar, Z. Pervez, and K. Dahal. 2020. $X - BAND$: Expiration band for anonymizing varied data streams. *IEEE IoT J.* 7, 2 (2020), 1438–1450.

[103] Jendrik Poloczek, Nils André Treiber, and Oliver Kramer. 2014. KNN regression as geo-imputation method for spatio-temporal wind data. In *International Joint Conference SOCO'14-CISIS'14-ICEUTE'14*. Springer, 185–193.

[104] A. Purwar and S. K. Singh. 2014. Empirical evaluation of algorithms to impute missing values for financial dataset. In *Proceedings of the International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*. 652–656.

[105] Yongsong Qin, Shichao Zhang, Xiaofeng Zhu, Jilian Zhang, and Chengqi Zhang. 2009. POP algorithm: Kernel-based imputation to treat missing values in knowledge discovery from databases. *Expert Syst. Appl.* 36, 2, Part 2 (2009), 2794–2804.

[106] L. Qu, L. Li, Y. Zhang, and J. Hu. 2009. PPCA-based missing data imputation for traffic flow volume: A systematical approach. *IEEE Trans. Intell. Transport. Syst.* 10, 3 (2009), 512–522.

[107] María Elisa Quinteros, Siyao Lu, Carola Blazquez, Juan Pablo Cárdenas-R, Ximena Ossa, Juana-María Delgado-Saborit, Roy M. Harrison, and Pablo Ruiz-Rudolph. 2019. Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in Temuco, Chile. *Atmos. Environ.* 200 (2019), 40–49.

[108] Geaur Rahman and Zahidul Islam. 2011. A decision tree-based missing value imputation technique for data pre-processing. In *Proceedings of the 9th Australasian Data Mining Conference-Volume 121*. 41–50.

[109] Md Geaur Rahman and Md Zahidul Islam. 2013. Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques. *Knowl.-Bas. Syst.* 53 (2013), 51–65.

[110] Wajeeha Rashid and Manoj Kumar Gupta. 2021. A perspective of missing value imputation approaches. In *Advances in Computational Intelligence and Communication Technology*. Springer, 307–315.
[111] R. Razavi-Far and M. Saif. 2016. Imputation of missing data using fuzzy neighborhood density-based clustering. In *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'16)*. 1834–1841.
[112] Donald B. Rubin. 2004. *Multiple Imputation for Nonresponse in Surveys*. Vol. 81. John Wiley &. Sons.
[113] Maytal Saar-Tsechansky and Foster Provost. 2007. Handling missing values when applying classification models. *J. Mach. Learn. Res.* 8 (December 2007), 1623–1657.
[114] Z. Sahri, R. Yusof, and J. Watada. 2014. FINNIM: Iterative imputation of missing values in dissolved gas analysis dataset. *IEEE Trans. Industr. Inform.* 10, 4 (2014), 2093–2102.
[115] Tariq Samad and Steven A. Harp. 1992. Self-organization with partial data. *Network* 3, 2 (1992), 205–212.
[116] Roosevelt Sardinha, Aline Paes, and Gerson Zaverucha. 2018. Revising the structure of Bayesian network classifiers in the presence of missing data. *Inf. Sci.* 439-440 (2018), 108–124.
[117] Joseph L. Schafer. 1997. *Analysis of Incomplete Multivariate Data*. CRC Press.
[118] Joseph L. Schafer and John W. Graham. 2002. Missing data: Our view of the state of the art. *Psychol. Methods* 7, 2 (2002), 147.
[119] Tapio Schneider. 2001. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *J. Clim.* 14, 5 (3 2001), 853–871.
[120] David H. Schoellhamer. 2001. Singular spectrum analysis for time series with missing data. *Geophys. Res. Lett.* 28, 16 (2001), 3187–3190.
[121] Michael Schomaker and Christian Heumann. 2018. Bootstrap inference when using multiple imputation. *Stat. Med.* 37, 14 (2018), 2252–2266.
[122] Amir Masoud Sefidian and Negin Daneshpour. 2019. Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model. *Expert Syst. Appl.* 115 (2019), 68–94.
[123] Anoop D. Shah, Jonathan W. Bartlett, James Carpenter, Owen Nicholas, and Harry Hemingway. 2014. Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *Am. J. Epidemiol.* 179, 6 (2014), 764–774.
[124] W. Shao, X. Shi, and P. S. Yu. 2013. Clustering on multiple incomplete datasets via collective kernel learning. In *Proceedings of the IEEE 13th International Conference on Data Mining*. 1181–1186.
[125] Mohamed Abu Sharkh and Mohamed Kalil. 2018. A quest for optimizing the data processing decision for cloud-fog hybrid environments. In *Proceedings of the IEEE International Conference on Communications Workshops (ICC Workshops'18)*. 1–6. https://doi.org/10.1109/ICCW.2018.8403743
[126] Y. Shen, F. Peng, and B. Li. 2015. Improved singular spectrum analysis for time series with missing data. *Nonlin. Process. Geophys.* 22, 4 (2015), 371–376.
[127] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). Curran Associates, Inc., 802–810.
[128] Esther-Lydia Silva-Ramírez, Rafael Pino-Mejías, and Manuel López-Coello. 2015. Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns. *Appl. Soft Comput.* 29 (2015), 65–74.
[129] Esther-Lydia Silva-Ramírez, Rafael Pino-Mejías, Manuel López-Coello, and María-Dolores Cubiles de-la Vega. 2011. Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Netw.* 24, 1 (2011), 121–129.
[130] Marina Soley-Bori. 2013. Dealing with missing data: Key assumptions and methods for applied analysis. Technical Report. Boston University. 20 pages.
[131] H. Song and D. A. Szafir. 2019. Wherés my data? Evaluating visualizations with missing data. *IEEE Trans. Vis. Comput. Graph.* 25, 1 (2019), 914–924. https://doi.org/10.1109/TVCG.2018.2864914
[132] Donald F. Specht. 1990. Probabilistic neural networks. *Neural Netw.* 3, 1 (1990), 109–118.
[133] Reinaldo Squillante, Diolino J. [Santos Fo], Newton Maruyama, Fabrício Junqueira, Lucas A. Moscato, Francisco Y. Nakamoto, Paulo E. Miyagi, and Jun Okamoto. 2018. Modeling accident scenarios from databases with missing data: A probabilistic approach for safety-related systems design. *Safe. Sci.* 104 (2018), 119–134.
[134] Daniel J. Stekhoven and Peter Bühlmann. 2011. MissForest-non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 1 (10 2011), 112–118.
[135] Samantha Stephens, Joseph Beyene, Mark S. Tremblay, Guy Faulkner, Eleanor Pullnayegum, and Brian M. Feldman. 2018. Strategies for dealing with missing accelerometer data. *Rheum. Dis. Clin. North Am.* 44, 2 (2018), 317–326.
[136] Chuan Sun, Yueyi Chen, and Cheng Cheng. 2021. Imputation of missing data from offshore wind farms using spatio-temporal correlation and feature correlation. *Energy* 229 (2021), 120777.

[137] S. Tak, S. Woo, and H. Yeo. 2016. Data-driven imputation method for traffic data in sectional units of road links. *IEEE Trans. Intell. Transport. Syst.* 17, 6 (2016), 1762–1771.
[138] Masayoshi Takahashi. 2017. Statistical inference in missing data by MCMC and non-MCMC multiple imputation algorithms: Assessing the effects of between-imputation iterations. *Data Sci. J.* 16 (2017).
[139] Fei Tang and Hemant Ishwaran. 2017. Random forest missing data algorithms. *Stat. Anal. Data Min.* 10, 6 (2017), 363–377.
[140] Ramesh S. V. Teegavarapu. 2020. Precipitation imputation with probability space-based weighting methods. *J. Hydrol.* 581 (2020), 124447.
[141] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics* 17, 6 (6 2001), 520–525.
[142] H. Turabieh, M. Mafarja, and S. Mirjalili. 2019. Dynamic adaptive network-based fuzzy inference system (D-ANFIS) for the imputation of missing data for Internet of medical things applications. *IEEE IoT J.* 6, 6 (2019), 9316–9325.
[143] Aashma Uprety and Danda B. Rawat. 2021. Reinforcement learning for IoT security: A comprehensive survey. *IEEE IoT J.* 8, 11 (2021), 8693–8706. https://doi.org/10.1109/JIOT.2020.3040957
[144] Stef Van Buuren. 2018. *Flexible Imputation of Missing Data*. CRC Press.
[145] Christian Velasco-Gallego and Iraklis Lazakis. 2020. Real-time data-driven missing data imputation for short-term sensor data of marine systems. A comparative study. *Ocean Eng.* 218 (2020), 108261.
[146] G. Wang, J. Lu, K. Choi, and G. Zhang. 2020. A transfer-based additive LS-SVM classifier for handling missing data. *IEEE Trans. Cybernet.* 50, 2 (2020), 739–752.
[147] Haolin Wang, Xuhai Tan, Zhilin Huang, Bo Pan, and Jie Tian. 2020. Mining incomplete clinical data for the early assessment of Kawasaki disease based on feature clustering and convolutional neural networks. *Artif. Intell. Med.* 105 (2020), 101859.
[148] Ming-Chang Wang, Chih-Fong Tsai, and Wei-Chao Lin. 2021. Towards missing electric power data imputation for energy management systems. *Expert Syst. Appl.* 174 (2021), 114743.
[149] Ian R. White, Patrick Royston, and Angela M. Wood. 2011. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* 30, 4 (2011), 377–399.
[150] Shengjie Xu, Yi Qian, and Rose Qingyang Hu. 2019. Data-driven network intelligence for anomaly detection. *IEEE Netw.* 33, 3 (2019), 88–95.
[151] X. Xu, Y. Lei, and Z. Li. 2020. An incorrect data detection method for big data cleaning of machinery condition monitoring. *IEEE Trans. Industr. Electr.* 67, 3 (2020), 2326–2336.
[152] Chen Ye, Hongzhi Wang, Wenbo Lu, and Jianzhong Li. 2020. Effective Bayesian-network-based missing value imputation enhanced by crowdsourcing. *Knowl.-Bas. Syst.* 190 (2020), 105199.
[153] Xiuwen Yi, Yu Zheng, Junbo Zhang, and Tianrui Li. 2016. ST-MVL: Filling missing values in geo-sensory time series data. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*.
[154] Xiuwen Yi, Yu Zheng, Junbo Zhang, and Tianrui Li. 2016. ST-MVL: Filling missing values in geo-sensory time series data. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, 2704–2710.
[155] Qiyue Yin, Shu Wu, and Liang Wang. 2017. Unified subspace learning for incomplete and unlabeled multi-view data. *Pattern Recogn.* 67 (2017), 313–327.
[156] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. 2018. Gain: Missing data imputation using generative adversarial nets. arXiv:1806.02920.
[157] Yang Yuan. 2014. Sensitivity analysis in multiple imputation for missing data. In *Proceedings of the SAS Global Forum 2014*.
[158] Lotfi A. Zadeh. 1965. Fuzzy sets. *Inf. Contr.* 8, 3 (1965), 338–353.
[159] Junlin Zhang, Samuel Oluwarotimi Williams, and Haoxiang Wang. 2018. Intelligent computing system based on pattern recognition and data mining algorithms. *Sust. Comput. Inform. Syst.* 20 (2018), 192–202.
[160] Q. Zhang, Q. Yuan, C. Zeng, X. Li, and Y. Wei. 2018. Missing data reconstruction in remote sensing image with a unified spatial-temporal-spectral deep convolutional neural network. *IEEE Trans. Geosci. Remote Sens.* 56, 8 (2018), 4274–4288.
[161] Y. Zhang, P. J. Thorburn, W. Xiang, and P. Fitch. 2019. SSIM-A deep learning approach for recovering missing time series sensor data. *IEEE IoT J.* 6, 4 (2019), 6618–6628.
[162] Zhongrong Zhang, Xuan Yang, Hao Li, Weide Li, Haowen Yan, and Fei Shi. 2017. Application of a novel hybrid method for spatiotemporal data imputation: A case study of the Minqin County groundwater level. *J. Hydrol.* 553 (2017), 384–397.
[163] Liang Zhao, Zhikui Chen, Yi Yang, Z. [Jane Wang], and Victor C. M. Leung. 2018. Incomplete multi-view clustering via deep semantic mapping. *Neurocomputing* 275 (2018), 1053–1062.

[164] L. Zhao, Z. Chen, Z. Yang, Y. Hu, and M. S. Obaidat. 2018. Local similarity imputation based on fast clustering for incomplete data in cyber-physical systems. *IEEE Syst. J.* 12, 2 (2018), 1610–1620.
[165] X. Zhu and X. Wu. 2005. Cost-constrained data acquisition for intelligent data preparation. *IEEE Trans. Knowl. Data Eng.* 17, 11 (2005), 1542–1556.
[166] X. Zhu, S. Zhang, Z. Jin, Z. Zhang, and Z. Xu. 2011. Missing value estimation for mixed-attribute data sets. *IEEE Trans. Knowl. Data Eng.* 23, 1 (2011), 110–121.
[167] Y. Zhuang, R. Ke, and Y. Wang. 2019. Innovative method for traffic data imputation based on convolutional neural network. *IET Intell. Transport Syst.* 13, 4 (2019), 605–613.

Received 2 July 2021; revised 30 November 2021; accepted 23 April 2022