IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 September 2023, accepted 5 October 2023, date of publication 10 October 2023, date of current version 18 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3323435

RESEARCH ARTICLE

A New Imputation Technique Based a Multi-Spike Neural Network to Handle Missing Data in the Internet of Things Network (IoT)

NADIA ADNAN SHILTAGH AL-JAMALI[1], (Member, IEEE), IBTESAM R. K. AL-SAEDI[2,4], AHMED R. ZARZOOR[3], AND HONGXIANG LI[4]

[1]Department of Computer Engineering, University of Baghdad, Baghdad 10071, Iraq
[2]Communication Engineering Department, University of Technology (UoT), Baghdad 10066, Iraq
[3]Directorate of Inspection, Ministry of Health, Baghdad 20250, Iraq
[4]Electrical and Computer Engineering Department, University of Louisville, Louisville, KY 40292, USA

Corresponding author: Ibtesam R. K. Al-Saedi (Ibtesam.R.Karhiy@Uotechnology.edu.iq)

This work was supported by the Ministry of Higher Education in Iraq through the research grants project in cooperation with the University of Louisville (UofL), KY, USA.

ABSTRACT Over the past decade, the Internet of Thing (IoT) devices have been deployed in wide-scale several applications to collect vast amount of data from different locations in a time-series manner. However, collected data may be missing or damaged due to several issues such as unreliable communications, faulty sensors and synchronization problem that decrees application accuracy. Therefore, a several imputation-based machine learning approaches have been suggested to handle this problem in IoT application. In this study, a new approach is proposed called impute missing data (IMD) based on multi-Spike Neural Network learning method called IMD-SNN, to increase the reliability of missing value imputation in IoT. The method consists of three phases: Inserting missing data, to evaluate the missing values based on the cumulative distribution function (CDF), the multi SNN phase to estimate missing data according to the timestamp and a performance evaluation phase to evaluate an imputation accuracy via made a comparison with two models: imputation based KNN (I-KNN) and Imputation based (I-MLP) model based on resource usage and imputation accuracy assessment metrics. The implementation results have been shown that IMD-SNN utilizes less energy usage in comparison with (I-MLP) model and I-KNN model and gives highest imputation accuracy in contrast with (I-MLP) model and I-KNN model. Also, the IMD-SNN model utilizes less memory usage and needs execution time less than I-MLP model.

INDEX TERMS The Internet of Things (IoT), spike neural network (SNN), multilayer perceptron (MLP), root mean square error (RMSE).

I. INTRODUCTION

Internet of Things (IoT) makes up of billions of things (i.e., devices), that producing a big important data, which is organized by a large volume, can be utilized in different issues [1]. For example, collected data can be utilized for weather activity monitoring [2] and smart manufacturing system which can collected industrial sounds to identify machine faults and implement corrective maintenance [3]. However, devices may be unsuccessful to convey data due to the networking issues or hardware failure. Thus, the server or gateway may not receive some device measurements, which decreasing system reliability due to missing data. Consequently, it is highly important to handle with missing data in a sensible way, that can assist the system running on missing value of IoT [4]. So, a better approach to guarantee high fineness services, is to impute the incomplete data into the record, so as to complete the system's viewpoint [5]. The core idea of the imputing miss data method is about replacing the missing data with the mean of all non-missing

The associate editor coordinating the review of this manuscript and approving it for publication was Hosam El-Ocla.

data received for this an attribute in the dataset. Moreover, there are other techniques, replaces the missing data with zero values or median of non-missing values. The main disadvantage of this approach it does not implement well in most conditions.

The missing data can be divided into three types: "Missing Completely At Random" (MCAR), "Missing At Random" (MAR) and "Not Missing At Random" (NMAR) [7], [8]. In MCAR, the missing data occurs randomly and do not produce deviation in the results, that have no linkage with data spotted or data non spotted. MAR points out to pattern where missing data are linked to data spotted that proposes the missing value can be deduced from existing data. While, in NMAR, a missing value is linked to data non spotted. Furthermore, the missing data can be classified into two classes: single imputation missing value and multiple imputation value. In the single imputation missing values approach the incomplete data is replaced with single and unique data such as zero values, median, mode or mean of a given dataset [9]. In the multiple imputation value approach, the incomplete data are replaced with multiple values, that has been predicted based on the deep learning approach [10], [11]. In this context the [12] categorizes the mechanisms to handle incomplete data in two methods: statistical and machine learning (ML). The statistical approaches are the oldest techniques utilized for estimation such as regression model [13], Hidden Markov Model (HMM) [14] and Expectation Maximization (EM) [15]. The ML approach is categorized into two categories: supervised and unsupervised. The supervised learning approach depends on labelled data, to predict missing data. While, the unsupervised learning approach based on unlabeled data in order to elicit patterns for incomplete data [16]. However, the data imputation-based ML gives high accuracy predication of missing. But it requires higher computation time and memory usage in comparison with statistical approaches. Besides, the data imputation based statistical method gives a good model for inference the correlation between the variables.

Therefore, in this study a new technique is proposed based on both (ML techniques and statistical techniques) called imputation a missing value based on multi-Spike Neural Network learning method called IMD-SNN. Based to our Knowledge, this is the first study for data imputation based SNN. The main motivation behind utilizing the SNN, that its optimize the memory usage and required less time for training and testing process in comparison with ML techniques (i.e.; NN, DL, CNN). So, the main contribution of this study: (1) developing a model based multi-SNN learning method to predict incomplete values according to the timestamp of the pervious time for three attributes: atmospheric pressure, humidity and temperature, (2) evaluating the model performance for imputation accuracy based on R2-score and root mean square error (RMSE) and resource usage (execution time and total memory usage). The rest of this paper, maps as follow: section II explores the data imputation techniques, section III, describes the IMD-SNN method, section IV includes the discussions and results. Finally, section V contains the conclusions.

II. RELATED WORKS

In the last decade, a new problem has been the existence of damage/missing values in the big dataset. Besides, the newly arisen term of IoT and its collected data generated with rocket speed, that reasoned more and more missing data due to the low quality and less reliability of IoT devices. Thus, many approaches have been proposed to handle this problem. For instance, in [17] authors utilized K-means clustering method to select the most ideal imputation methods to replace missing values with appropriate value. The method consists of three steps: firstly, divides the missing values into different clusters according to the k means method. Secondly, missing data in each separated cluster is estimates based on the monitoring values with each cluster. In [18] researchers propose a new imputation based on multivariate KNN (wkNNr) method to handle missing values in PM2.5 dataset. They utilized the relationship between AQM stations record to weight between surveillances. Subsequently, they used the reverse of this distance for computing the weighted mean of the KN surveillance. Lastly, the imputation process is based on complete PM2.5 record. Consequently, enabling the use of other assistant data in a forward analysis without considering collinearity. In [19] authors suggested a new imputation technique to manage an online incomplete value by using "Virtual Temporal Neighbors" VTN. In this technique, the VTN utilized the previous characterization of the sensor value-stream so as to estimate the missing data based on regression model.

In [20] researchers used imputation based fuzzy system (FS) to replace missing values with appropriated value in sensor data. The FS is created an oval clustering (i.e., obtained an oval shape for each cluster of missing values) to expect a coordinate axis of oval shape and identical with FS. Reference [21] suggests "Fuzzy-K-Top Machining Value" (FKTM) for missing value imputation. They replace numerical and smart estimates according to the comparable records. Also, used the FS to find the cluster of comparable data and valuation them. Reference [22] proposes an imputation method based on Fuzzy system and entropy measurement, to select the elect features (or patterns) loading missing data assistance the forecast missing data within the election feature utilizing the "Bayesian Ridge Regression" BRR. In their technique the missing data values are manipulated within other patterns in accumulative order, the loaded patterns are combined within BRR equation to predict the missing data for the next elected missing pattern. In [23] study, two supervised approaches (Multilayer perception and deep belief network) are utilized to predict incomplete values in continuous data and compare their performance to the discretization data. Reference [24] used imputation-based auto encoder to predict missing data and peripheralization over missing values in a shared model of common variables and outcomes. In [25] authors utilized

TABLE 1. A summary of existing imputation based machine learning method.

Study	Year	Imputation Based ML	Summary
[17]	2022	K-Means Clustering	-Divides the missing values into different clusters according to the k means method. - Missing data in each separated cluster is evaluated based on the monitoring values with each cluster. -The optimal imputation approach is selected according to the valuation of "reverse error score" RES
[18]	2022	Multivariate KNN (wkNNr) to handle missing values	-Uses the reverse of the distance between stations for computing the weighted mean of the KN surveillance. - The imputation process is based on complete PM2.5 record, to enable the use of other assistant data in forward analysis without considering collinearity
[19]	2022	Virtual Temporal Neighbors and linear Regression	- Utilizes the previous characterization of the sensor value stream so as to estimate the missing data based on regression model
[20]	2022	Fuzzy System	- Obtained an oval shape for each cluster of missing values that created by fuzzy system, via expecting a coordinate axis of oval shape and identical with FS
[21]	2022	Fuzzy-K-Top Machining Value	- Replace numerical and smart estimates according to the comparable records. - The Fuzzy System is used to find the cluster of comparable data and valuating them.
[22]	2022	Fuzzy System, entropy measurements	- Selects the elect features (or patterns) loading missing data assistance the forecast missing data within the election feature by utilizing the BRR
[23]	2022	Multilayer perception and deep belief network	- Predicts incomplete values in continuous data and compare their performance to the discretization data
[24]	2022	Auto encoder	- Predicts missing data and peripheralization over missing values in a shared model of common variables and outcomes
[25]	2021	Adaptive multiple imputation of missing value (Clustering)	- Uses the class center and identifies a threshold according to the weighted distance between (the center and monitoring values) for the imputation missing data
[26]	2021	regression model with neural network	- Prophesy and input missing value in IoT gateways in order to attain major autonomy at the network gateway
[27]	2020	Decision Trees (DT) and Fuzzy Clustering Iterates (FCI)	- The FCI method is used to learn new evaluating values from the dataset with a comparable attribute value that specified by DT approach
[28]	2020	Clustering	-Uses the similarity among the monitor data to achieve imputation. This accomplished by separating missing data and grouping them within a cluster according to the similar records in each group to valuation the missing data.
[29]	2023	Clustering	-Utilizes temporal and locative relationship of IoT nodes and share neighbor. -Missing value can be restored with the aid of the share neighbor nodes' information
[30]	2021	deep learning (DL) based imputation approach	-The method sequentially, eliminates bias, seasonality, descent, seasonality and remaining of input time chains data. -The missing values can be predicted from identified information such as bias, slop.
Propose study		SNN	- predicts incomplete values according to their pervious timestamp,

"Adaptive multiple imputation of missing value" AMIC method to impute missing data. In AMIC they used the class center and identifies a threshold according to the weighted distance between (the center and monitoring values) for the imputation missing data. Furthermore, in AMIC the distance can be adjusted the center or nearest neighbors to valuation missing data.

In study [26] researchers, proposes a technique to prophesy and input missing value in IoT gateways in order to attain major autonomy at the network border. They used a regression model with two layers of neural network to impute missing data of the gateways. Suggests [27] a new approach called DIFC that based on merging the characteristics of Decision Trees (DT) and Fuzzy Clustering Iterates (FCI) into reduplicate leaning approach to impute missing data type MCAR. The FCI method is used to learn new evaluating values from the dataset with a comparable attribute value, that specified by DT approach. Authors [28] utilized a new imputation based on data clustering and active election of the appropriate imputation equation for each missing value in the dataset. The main idea of their method is about using the similarity among the monitoring data to achieve imputation. This accomplished by separating missing data and grouping them within a cluster according to the similar records in each group to valuation the missing data.

While, [29] presents a method for recovering missing value in IoT nodes. The method consists of two levels: clustering (CL) and data recover (DR). In the CL, the nodes are grouped according to their temporal and locative relationship and share neighbor, which are extracted. In the DR level, missing value can be restored with the aid of neighbor nodes utilizing the ST "hierarchical long short-term memory" (ST-HLSTM) approach. Reference [30] suggests a new deep learning (DL) based imputation approach which espouses the explicate ability of N-BEATS for imputation job. The method sequentially eliminates bias, seasonality, descent, seasonality and remaining of input time chains data. Thus, the missing values can be predicted from identified information such as bias, slop. Besides, impute missing data which happens in multiple places simultaneously is accomplished by using temporal-locative information. In this study an SNN based imputation approach is utilized handle missing values in IoT. To our knowledge, this is the first research used SNN to impute missing. To summarize all imputation-based ML approaches in comparison with this study, see table 1.

III. METHOD

In this study a new technique based on cumulative distribution function (CDF) data and SNN (IMD-SNN) to impute missing data in dataset ''MonitorAr'' utilizing four attributes Humidity, Temperature, Timestamp and Atmospheric pressure. The IMD-SNN is consisting of three phases: Inserting of missing data, impute based SNN and performance evaluation, see figure 1. The dataset [33] has been used in this study, where it includes meteorological data that gauged every hour, of a station in Rio de Janeiro, positioned in the section near of the São Cristóvão. The station contains a complete data record, some records with incomplete values for some attributes, and gaps among timestamps, i.e., hours in case no meteorological data received. To overcome this problem, preprocessed process on dataset has been accomplished by adding empty records when a timestamp missing. So, each day (24 hours, where each record for one hour), and year has 8760 or 8784 records for a leap year).

A. INSERTING MISSING DATA PHASE

The desirable incomplete data percentage is elected by using uniform distribution (CDF) approach for three input attributes atmospheric pressure, humidity and temperature. The percentage in the dataset of this study, 5%, 10%, 25% and 50% For example, if the dataset has 100% records, 20% of incomplete data denotes each record has twenty missing values. So, 50% percentage utilized in this step. Where, data is divided into two datasets: training dataset (2012 to 2015) and testing dataset (2016 to 2019). Also, missing data can be happened as separately or in bursts. In each burst, the burst volume (BV) coincides with how many successive incomplete data occur for one attribute. Since a separated loss is in burst of single volume. Figure 2, where demonstrates the CDF for each BV of the main test dataset and the plurality of missing values is small. Consequently, thither are bursts with greater than 100 successive missing data. Figure 2, demonstrates the CDF of the BV of one pattern of the test after deleting to attain 50% missing value of the dataset where, BV size 1000 represents most of the missing data.

The BV of characteristics is so close to each other. The prospect of having a separated losses is about 79% loss bigger than the main test dataset prospect (0.59). This was expected because the data is selected to remove by utilizing CDF. Also, in this phase the missing value is replaced with the mean value of three previous values of three attributes: Temperature, Atmospheric Pressure and Humidity for the

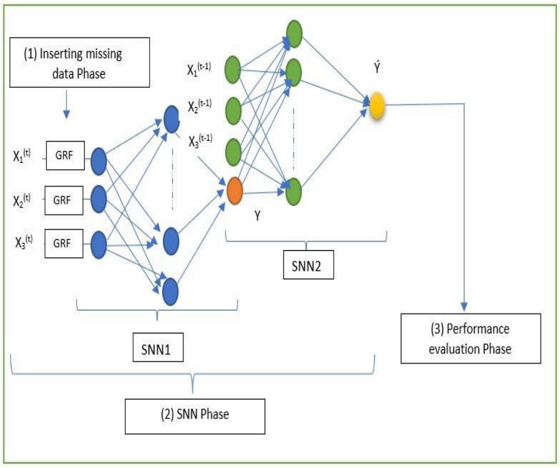

FIGURE 1. Illustrates IMD-SNN method three phases.

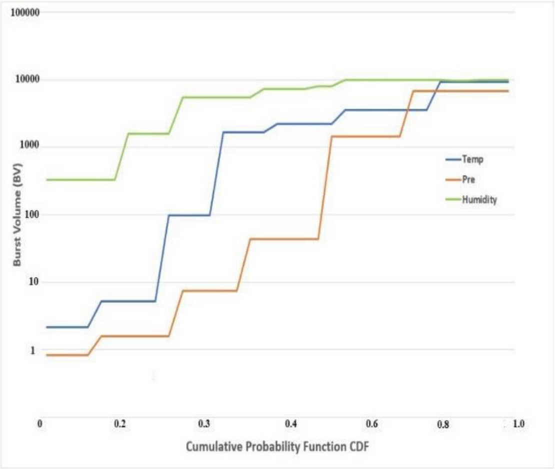

FIGURE 2. Main test dataset.

''datasetMonitorAr: Cityhall of Rio de Janeiro-MonitorAr dataset'' [33] to reduce the network load, since there is only a record with predicted values.

B. SNN PHASE

In this study, the SNN consists of two phases (SNN1, SNN2), each phase composed of: an input layer, one hidden layer and the output layer, in order to reduce the execution time and alleviates memory usage. In SNN1, three input attributes (Humidity (X^t_1), Temperature(X^t_2) and Atmospheric Pressure (X^t_3)) used as input values in the input layer, where (t) represents a timestamp of one day. For training the selected patterns value that obtained for the input layer a ''Gaussian Receptive Fields'' GRF algorithm [31], is utilized to encode information into firing times for the input layer by utilizing equation 1. Where, input value between (minimum data value (Vmin) and maximum data value Vmax) with σ is centered by utilizing equation 2. The spike timing is arranged between (0 to T), T value is computed by using equation 3. The σ is allocated by the passing points of the V with corresponding

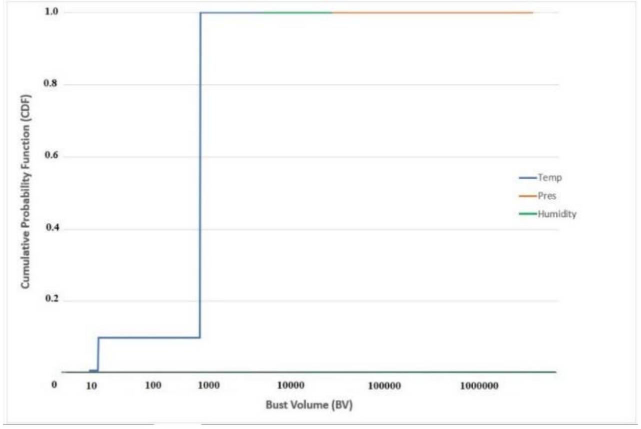

FIGURE 3. CDF and BV after attaining 50% of the dataset.

Gaussian summits: the i-th input receives a spike at T- $a_i\left(V\right)$. So, when $a_i\left(V\right) > 0.01$ and no spikes, then the nearest value of v to the σ will be taken, while Backpropagation method is used in the hidden layer to update weight so as to alleviate error, see equation 4. Where, W_i, represents a new weight and (b) represent the learning rate (the minimum value) for the error function [32]. The output layer gives the predicted missing value (Y), see figure 1.

$$a_i\left(V\right) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{(x-\mu)^2}{\sigma^2}\right) \tag{1}$$

$$\mu_i = V_{min} + \left(V_{max} - V_{min}\right)\cdot\frac{i}{n-1}$$

$$fori = 0ton - 1 \tag{2}$$

$$T = \max\left(a_i\left(V\right)\right) \tag{3}$$

$$W_i = W_i - b(\frac{\partial Error}{\partial Wi}) \tag{4}$$

In SNN2, the input layer used four attributes (X_1^{t-1}, X_2^{t-1}, X_3^{t-1}, Y), where (t-1) represents the pervious timestamp (i.e., calculated for the period of the selfsame timestamp of the previous day) and Y represents the predicted output value of SNN1. Also, Backpropagation method is used in the hidden layer to update weight so as to alleviate error. While, SNN2 output layer gives the final predicted value (Ý) for each input attribute (i.e.; $\acute{Y}_1$, $\acute{Y}_2$and $\acute{Y}_3$ final predicted output value for atmospheric pressure, humidity and temperature respectively)

C. PERFORMANCE EVALUATION PHASE

In this phase, the R2-score and root mean square errors (RMSE) used to measure the imputation accuracy of the IMD-SNN and two other imputation models: imputation based KNN (I-KNN) [34] and imputation based multilayer perceptron (I-MLP) (i.e., more than one neural network (NN)) [7], [35]. The I-KNN algorithm imputation missing value is performed by dividing a dataset into k distinguished clusters, in the first stage. Thus, this technique results in generating membership values to all the points, which gather with a particular cluster or gather according to centroid. In the next step, all the missing cases are evaluated by utilizing the membership gauge of other points which fill with a border of the same group. While, in I-MLP method, each layer consists of neural network NN, used to predict the pattern's measure. Each NN consists of an input layer, one hidden layer and the output layer. The input layer used to input attribute's value to the hidden layer, which in turn handle values to output layer. The input layer used to input attribute's value to the hidden layer, which in turn handle values to output layer. The output layer gives the predict missing value that utilized as input value with other attributes to the next neural network. However, this process increases the accuracy of missing value predication, but it exhausts the network resources (battery life, time computation). Therefore, in this study, only two NNs (NN1, NN2) utilized, to reduce network resource usage and increasing the prediction accuracy of incomplete data for each attribute in the dataset. The imputation performance for the three methods is assessed by using two statistical methods using R^2-score [36] and root mean square (RMSE) [37]. The R^2 is the coefficient of determination value that calculated by utilizing equation 5. Where, R^2 value could be (0 for prediction of missing values, 1 means a perfect prediction of missing value, between 0 and 1 means partially predicated of missing value), Y the real value, Ý predicts value, $\bar{Y}$ is the mean of all records, n number of records. Also, RMSE used to gauge the mean difference between imputed values predicted via a model and the actual missing values the RMSE is calculated by using equation 6, where $\sum_{i=1}^{n}\left(Y - \acute{Y}\right)^2$ is the total of the errors

$$R^2\left(Y, \acute{Y}\right) = 1 - \frac{\sum_{i=1}^{n}\left(Y - \acute{Y}\right)^2}{\sum_{i=1}^{n}\left(Y - \bar{Y}\right)^2} \tag{5}$$

$$RMSE\left(Y, \acute{Y}\right) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y - \acute{Y}\right)^2} \tag{6}$$

IV. RESULTS AND DISCUSSION

In this study, the IMD-SNN approach has been implemented on laptop type Lenovo (CPU speed 2.5 GHz Intel Core i7, RAM 6GB and operation system MS Window 11. Three scenarios (I-MLP, I-KNN and IMD-SNN) have been conducted in order to evaluate the three models. The three scenarios are conducted by utilizing python (version 3.8) language libraries: NumPy and scikit-learn to implement I-KNN and SNNTroch to apply (IMD-SNN), Scikit-learn 0.24.1 to implement I -MLP. For the first scenario, threshold voltage Vth of input layer node (20 mV), threshold voltage Vth of hidden/output layer node (65 mV), batch size 100, see table 2. While, for I-MLP, number of input neurons 100, hidden layer size 80, activation function (Relu), see table 3. Besides, all the missing value is replaced with the mean value of previous values according to the timestamp for each of Temperature (C^o), Atmospheric Pressure (UR) and Humidity (mbar) in the MonitorAr dataset as initial step so as to reduce the network load, since there is only a record with predicted

TABLE 2. Parameters details for IMD-SNN.

Parameter	Value
max_depth for SNN	2
learning rate	0.001
batch size	100
Threshold voltage V_{th} of input layer node	20 mV
Threshold voltage V_{th} of hidden/output layer node	65 mV
Membrane resistance (all nodes)	1 MΩ
Membrane time constant (all nodes)	20 ms

TABLE 3. Parameters details for I-MLP.

Parameter	Value
Input neuron	100
Hidden layer size	80
Activation function	Relu
Epochs	120/20
Batch size	100
Optimizer	Adam
max_iter	500
Dropout rate	0.9

values. The imputation accuracy for three scenarios have been measured by using R^2-score and root mean square (RMSE). Also, utilized Python libraries (time and tracemalloc) in order to calculate the execution time and the allocated memory, respectively, in each scenario.

For the data imputation accuracy, it has found that R^2 scores of the imputation approach, when changing the missing value percentage (5%, 10%, 25%, 50%) for the humidity, temperature and atmospheric pressure. For atmospheric pressure, see table 6, the SNN1, SNN2 achieves high R2 score in comparison with NN1, NN2 and I-KNN because it has low changed for nears moments of time, see figure 4. For humidity, see table 5, Where, temperature score values varying is more significant than atmospheric pressure, see table 4 among SN1, SN2, NN1, NN2 and KNN, see Figure 5 and 6. While for RSME values, the SNN1, SNN2 have achieved better for three attributes than NN1, NN2 and I-KNN see figure 7, 8, 9. Besides, the RSME values, see table 7, 8 and 9 for NN1, NN2, SNN1 and SNN2 are near to each other for all the attributes and missing data percentages. Thus, these results mean adding more input data does not aid NN2 or SNN2 model. Furthermore, Figure 2, 3 demonstrates that missing value percentage has a minimum effect on the performance of the approaches due to, the high values of the RMSE and R^2 values. However, this conduct was expected because the missing value percentage injection traced a uniform distribution. Therefore, there are yet a lot of speared losses, see figure 3 where, the probability of separated loss is bigger than burst loss.

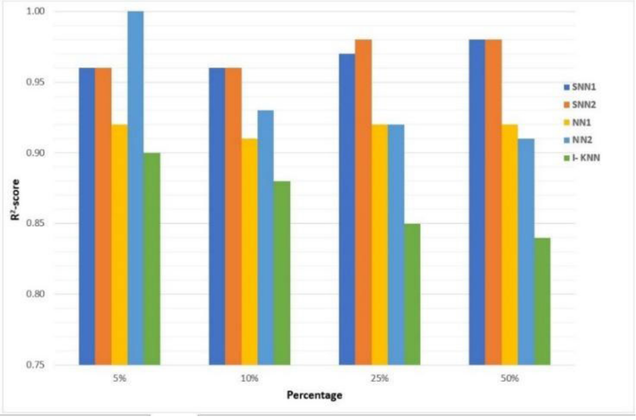

FIGURE 4. R^2 score for atmospheric pressure attribute.

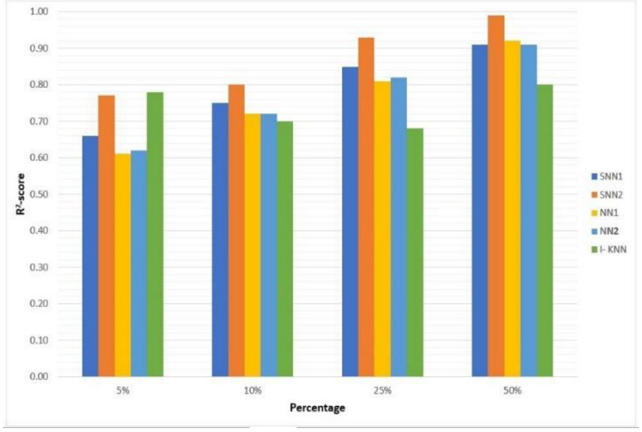

FIGURE 5. R^2 score for humidity attribute.

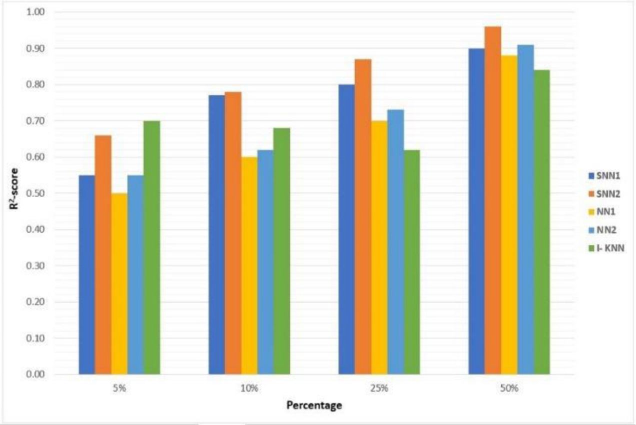

FIGURE 6. R^2 score for temperature attribute.

For the resource computation: memory utilization and execution time both are assessed in the training and testing layer. Where, average of 50 records for each missing value percentage within a confidence interval 95%. Also, only the IMD-SNN and I-MLP needs training. Therefore, attributes are trained in order to measure memory utilization and execution time. While, in testing layer the measurement of both metrics has been preferment for three methods

TABLE 4. R^2score details for atmospheric pressure.

	SNN1	SNN2	NN1	NN2	I- KNN
5%	0.96	0.96	0.92	9.4	0.9
10%	0.96	0.96	0.91	0.93	0.88
25%	0.97	0.98	0.92	0.92	0.85
50%	0.98	0.98	0.92	0.91	0.84

TABLE 5. R^2score details for humidity.

	SNN1	SNN2	NN1	NN2	I- KNN
5%	0.55	0.66	0.5	0.55	0.7
10%	0.77	0.78	0.6	0.62	0.68
25%	0.8	0.87	0.7	0.73	0.62
50%	0.9	0.96	0.88	0.91	0.84

TABLE 6. R^2score details for temperature.

	SNN1	SNN2	NN1	NN2	I- KNN
5%	0.66	0.77	0.61	0.62	0.78
10%	0.75	0.8	0.72	0.72	0.7
25%	0.85	0.93	0.81	0.82	0.68
50%	0.91	0.99	0.92	0.91	0.8

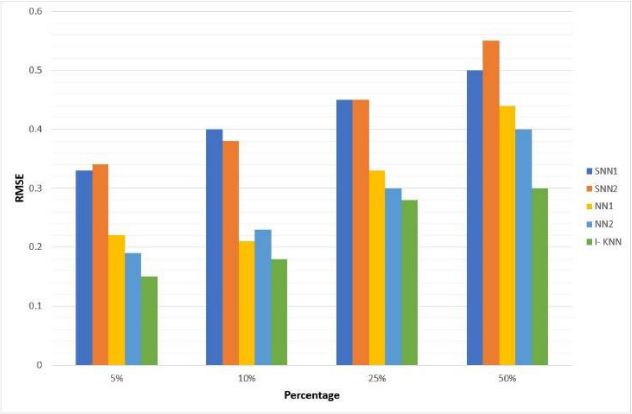

FIGURE 7. RMSE for atmospheric pressure attribute.

TABLE 7. RMSE details for atmospheric pressure.

	SNN1	SNN2	NN1	NN2	I- KNN
5%	0.33	0.34	0.22	0.19	0.15
10%	0.4	0.38	0.21	0.23	0.18
25%	0.45	0.45	0.33	0.3	0.28
50%	0.5	0.55	0.44	0.4	0.3

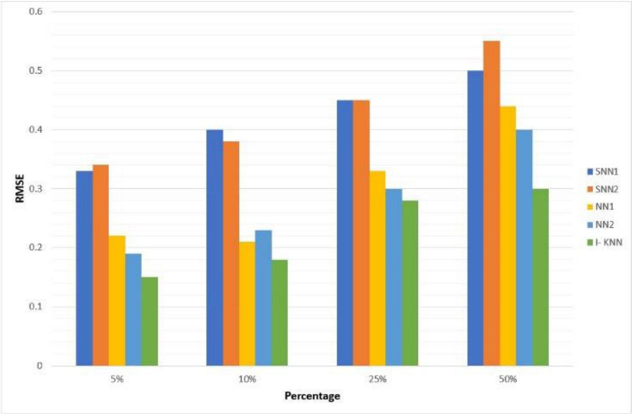

FIGURE 8. RMSE for humidity attribute.

TABLE 8. RMSE details for humidity.

	SNN1	SNN2	NN1	NN2	I- KNN
5%	0.28	0.29	0.22	0.22	0.18
10%	0.32	0.31	0.2	0.21	0.18
25%	0.44	0.45	0.32	0.32	0.28
50%	0.55	0.55	0.44	0.4	0.33

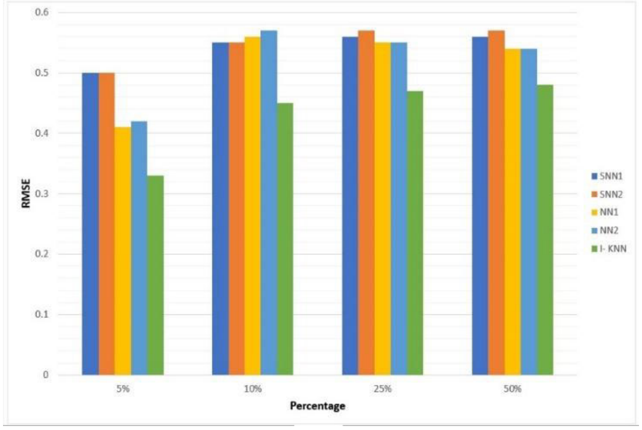

FIGURE 9. RMSE for temperature attribute.

TABLE 9. RMSE details for temperature.

	SNN1	SNN2	NN1	NN2	I- KNN
5%	0.5	0.5	0.41	0.42	0.33
10%	0.55	0.55	0.56	0.57	0.45
25%	0.56	0.57	0.55	0.55	0.47
50%	0.56	0.57	0.54	0.54	0.48

(IMD-SNN, I-MLP and I-KNN) to impute missing values, that are in the dataset. Furthermore, all approaches are applied on a single record at a time. Figure 10 demonstrates the lowest total memory usage kilobytes (KiB) value for NN1 (4100 KiB), NN2 (300 KiB), SNN1(3900 KiB) and SNN2 (4100 KiB) and the highest total memory usage value 4400 KiB, 4600 KiB, 4300 KiB, 4400 KiB for NN1, NN2, SNN1 and SNN2 respectively.

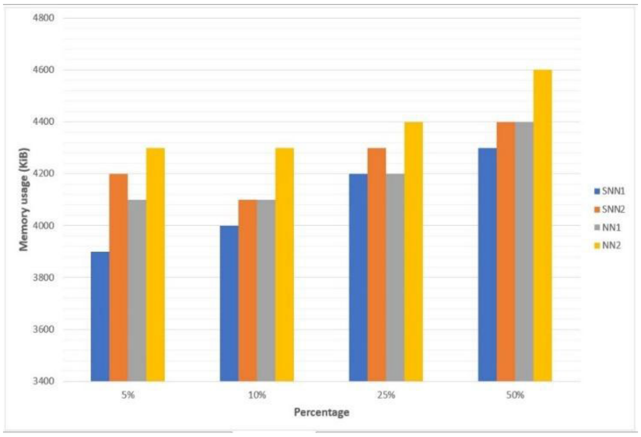

FIGURE 10. Illustrate memory usage IMD-SNN and I-MLP in training phase.

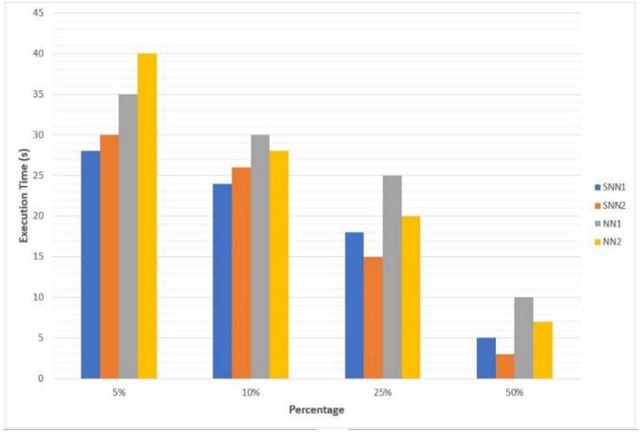

FIGURE 11. Illustrates execution time for IMD-SNN and I-MLP in training phase.

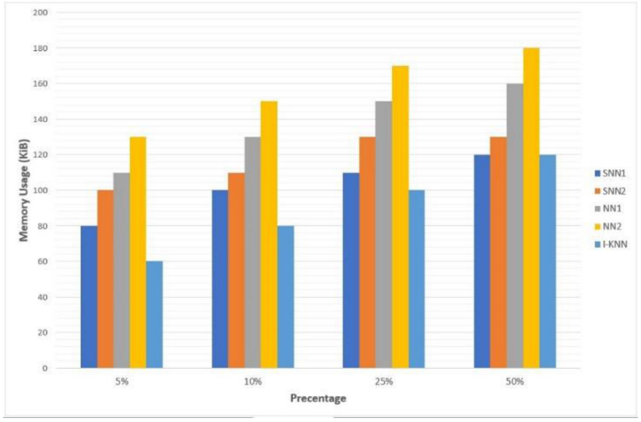

FIGURE 12. Illustrates memory usage for IMD-SNN and I-MLP in testing phase.

The execution time is decreased as the missing value percentage raises for the for NN1, SNN1 and SNN2. Also, SNN1 and SNN2 required less time in comparison with NN1 and NN2 simultaneous, see figure 11. In the testing phase, lowest total memory usage kilobytes (KiB) value for SNN1 (80 KiB), SNN2 (100 KiB), NN1(110 KiB), NN2 ((130 KiB) and I-KNN (60 KiB) and the highest total memory allocated value 120 KiB, 130 KiB, 160 KiB, 180 KiB, 120 Kib for SNN1, SNN2, NN1, NN2 and I-KNN respectively, see figure 12. For execution time is increased as the missing value percentage raises for the SNN1, SNN2, NN1, NN2 and I-KNN and I-KNN achieves the lowest value in comparisons with I-MLP and IMD-SNN. Moreover, SNN1 and SNN2 needs less time in comparison with NN1 and NN2 respectively, see figure 14. Also, in this study, energy usage metric has been used to evaluate the three model's performance. Where, the energy usage is calculated by utilizing equation 7 and 8 respectively [38]. The Eenegy_Tx represents the amount of energy usage, that required to convey (k) data packet for (d) distance between a pair of nodes, Eenegy_Rx represents the amount of energy usage that needed to get (k) for (d) among a pair of nodes. However, the implementation results of three methods have been shown that IMD-SNN model consumed less energy in comparison with I-KNN and I-MLP method, see figure 14.

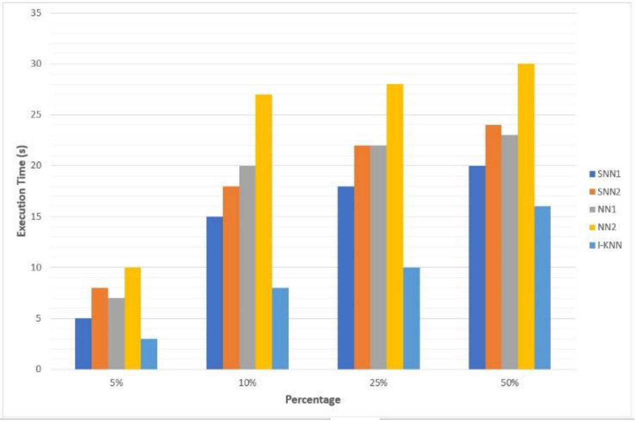

FIGURE 13. Illustrates execution time for IMD-SNN and I-MLP in testing phase.

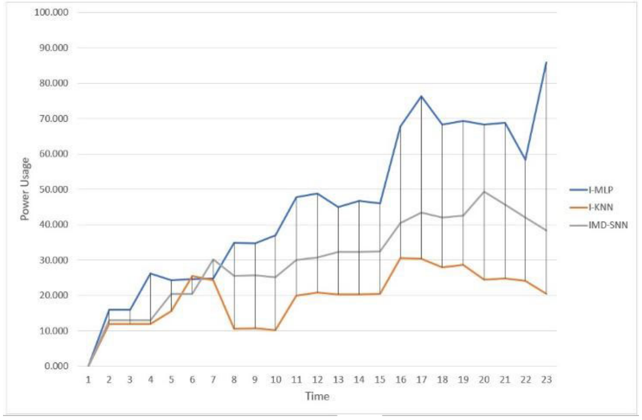

FIGURE 14. Illustrates power usage for IMD-SNN, I-KNN and I-MLP.

$$Eenegy_{Tx}\left(d,k\right)=\begin{cases}kElec+k\varepsilon ampd^{2}, d<d_{0}\\ kElec+k\varepsilon ampd^{4}, d\geq d_{0}\end{cases}\tag{7}$$

$$Eenegy_{Rx}\left(k\right)=kElec+kEpa\tag{8}$$

V. CONCLUSION

In this study, a new imputation technique is proposed based on the multi spike neural network to predict missing values in IoT dataset. The technique consists of three layers: inserting of missing data, SNN and performance evaluation. In the first layer, a distribution uniform (CDF) approach has been used to select the missing data that used as input to the second phase. While, two spike neural networks: SNN1, SNN2 have been utilized to predict the missing data. Finally, in the last phase a performance of the model is evaluates the imputation accuracy by using R^2-scors and RMSE in comparison with two models: imputation based KNN and Imputation based MLP. Also, three models are evaluated according to resource usage (total memory utilization, execution time).

The three models have been implemented by applying three scenarios: IMD-SNN, I-MLP and I-KNN utilizing python libraries on the MonitorAr dataset for three attributes: atmospheric Pressure, temperature and humidity. The implementation results have shown, that IMD-SNN gives high prediction accuracy in comparison with I-MLP and I-KNN for three attributes. For the resource usage, the IMD-

SNN model utilized less memory allocated and needs minimum execution time in contrast to I-MLP. Nevertheless, the I-KNN gives lowest memory usage and needs less execution time in comparison with the IMD-SNN and I-MLP model. Also, the IMD-SNN model has utilized less energy in comparison with I-MLP and I-KNN.

REFERENCES

[1] Y. Hajjaji, W. Boulila, I. R. Farah, I. Romdhani, and A. Hussain, "Big data and IoT-based applications in smart environments: A systematic review," *Comput. Sci. Rev.*, vol. 39, pp. 1–39, Feb. 2021.

[2] S. C. Santos, R. M. Firmino, D. M. F. Mattos, and D. S. V. Medeiros, "An IoT rainfall monitoring application based on wireless communication technologies," in *Proc. 4th Conf. Cloud Internet Things (CIoT)*, Oct. 2020, pp. 53–56.

[3] L. Gantert, M. Sammarco, M. Detyniecki, and M. E. M. Campista, "A supervised approach for corrective maintenance using spectral features from industrial sounds," in *Proc. IEEE 7th World Forum Internet Things (WF-IoT)*, Jun. 2021, pp. 723–728.

[4] Y. Liu, T. Dillon, W. Yu, W. Rahayu, and F. Mostafa, "Missing value imputation for industrial IoT sensor data with large gaps," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 6855–6867, Aug. 2020.

[5] J.-C. Kim and K. Chung, "Recurrent neural network-based multimodal deep learning for estimating missing values in healthcare," *Appl. Sci.*, vol. 12, no. 15, pp. 1–15, 2022.

[6] D. Adhikari, W. Jiang, J. Zhan, Z. He, D. B. Rawat, U. Aickelin, and H. A. Khorshidi, "A comprehensive survey on imputation of missing data in Internet of Things," *ACM Comput. Surv.*, vol. 55, no. 7, pp. 1–38, Jul. 2023.

[7] H. Pan, Z. Ye, Q. He, C. Yan, J. Yuan, X. Lai, J. Su, and R. Li, "Discrete missing data imputation using multilayer perceptron and momentum gradient descent," *Sensors*, vol. 22, no. 15, pp. 1–23, 2022.

[8] H. Ahn, K. Sun, and K. Pio Kim, "Comparison of missing data imputation methods in time series forecasting," *Comput., Mater. Continua*, vol. 70, no. 1, pp. 767–779, 2022.

[9] S. Rouzinov and A. Berchtold, "Regression-based approach to test missing data mechanisms," *Data*, vol. 7, no. 16, pp. 1–28, 2022.

[10] L. Weed, R. Lok, D. Chawra, and J. Zeitzer, "The impact of missing data and imputation methods on the analysis of 24-hour activity patterns," *Clocks Sleep*, vol. 4, no. 4, pp. 497–507, Sep. 2022.

[11] A. Hammon, "Multiple imputation of ordinal missing not at random data," *AStA Adv. Stat. Anal.*, pp. 1–22, Aug. 2022.

[12] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *J. Big Data*, vol. 8, no. 1, pp. 1–23, Oct. 2021.

[13] M. Chen, H. Zhu, Y. Chen, and Y. Wang, "A novel missing data imputation approach for time series air quality data based on logistic regression," *Atmosphere*, vol. 13, no. 7, pp. 1–22, 2022.

[14] S. Wang, M. Kim, X. Jiang, and A. O. Harmanci, "Evaluation of vicinity-based hidden Markov models for genotype imputation," *BMC Bioinf.*, vol. 23, no. 1, pp. 1–26, Aug. 2022.

[15] E. Thulare, R. Ajoodha, and A. Jadhav, "An empirical analysis and application of the expectation-maximization and matrix completion algorithms for varying degrees of missing data," in *Proc. Southern Afr. Univ. Power Eng. Conf./Robot. Mechatronics/Pattern Recognit. Assoc. South Afr. (SAUPEC/RobMech/PRASA)*, Jan. 2021, pp. 1–7.

[16] B. Al-Helali, Q. Chen, B. Xue, and M. Zhang, "A new imputation method based on genetic programming and weighted KNN for symbolic regression with incomplete data," *Soft Comput.*, vol. 25, pp. 1–20, Apr. 2021.

[17] B. Agbo, H. Al-Aqrabi, R. Hill, and T. Alsboui, "Missing data imputation in the Internet of Things sensor networks," *Future Internet*, vol. 14, no. 5, pp. 1–16, 2022.

[18] I. Belachsen and D. M. Broday, "Imputation of missing $PM_{2.5}$ observations in a network of air quality monitoring stations by a new kNN method," *Atmosphere*, vol. 13, no. 11, pp. 1–16, 2022.

[19] Y. Deng, C. Han, J. Guo, L. Li, and L. Sun, "Online missing data imputation using virtual temporal neighbor in wireless sensor networks," *Wireless Commun. Mobile Comput.*, vol. 2022, pp. 1–20, Feb. 2022.

[20] T. Moiseeva and T. Ledeneva, "Missing data imputation using fuzzy system," in *Proc. 4th Int. Conf. Control Syst., Math. Modeling, Autom. Energy Efficiency (SUMMA)*, Nov. 2022, pp. 350–354.

[21] A. Ali, M. Abu-Elkheir, A. Atwan, and M. Elmogy, "Missing values imputation using fuzzy K-top matching value," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 35, no. 1, pp. 1–12, 2022.

[22] F. K. Karim, H. Elmannai, A. Seleem, S. Hamad, and S. M. Mostafa, "Handling missing values based on similarity classifiers and fuzzy entropy measures," *Electronics*, vol. 11, no. 23, p. 3928, 2022.

[23] W.-C. Lin, C.-F. Tsai, and J. R. Zhong, "Deep learning for missing value imputation of continuous data and the effect of data discretization," *Knowl.-Based Syst.*, vol. 239, Mar. 2022, Art. no. 108079.

[24] N. Ipsen, P. Mattei, and J. Frellsen, "How to deal with missing data in supervised deep learning?" *Proc. 10th Int. Conf. Learn. Represent. (ICLR)*, 2022, pp. 1–30.

[25] K. Phiwhorm, C. Saikaew, C. K. Leung, P. Polpinit, and K. R. Saikaew, "Adaptive multiple imputations of missing values using the class center," *J. Big Data*, vol. 9, no. 1, pp. 1–25, Apr. 2022.

[26] C. M. França, R. S. Couto, and P. B. Velloso, "Missing data imputation in Internet of Things gateways," *Information*, vol. 12, no. 10, pp. 1–22, 2021.

[27] S. Nikfalazar, C.-H. Yeh, S. Bedingfield, and H. A. Khorshidi, "Missing data imputation using decision trees and fuzzy clustering with iterative learning," *Knowl. Inf. Syst.*, vol. 62, no. 6, pp. 2419–2437, Jun. 2020.

[28] B. Agbo, Y. Qin, and R. Hill, "Best fit missing value imputation (BFMVI) algorithm for incomplete data in the Internet of Things," in *Proc. 5th Int. Conf. Internet Things, Big Data Secur. (IoTBDS)*, 2020, pp. 130–137.

[29] P. Vedavalli and D. Ch, "A deep learning based data recovery approach for missing and erroneous data of IoT nodes," *Sensors*, vol. 23, no. 1, pp. 1–16, 2023.

[30] T. Kim, J. Kim, W. Yang, H. Lee, and J. Choo, "Missing value imputation of time-series air-quality data via deep neural networks," *Int. J. Environ. Res. Public Health*, vol. 8, no. 22, pp. 1–8, Oct. 2021.

[31] P. S. Maciąg, M. Kryszkiewicz, R. Bembenik, J. L. Lobo, and J. Del Ser, "Unsupervised anomaly detection in stream data with online evolving spiking neural networks," *Neural Netw.*, vol. 139, pp. 118–139, Jul. 2021.

[32] R. A. Zarzoor, N. A. S. Al-Jamali, and D. A. A. Qader, "Intrusion detection method for Internet of Things based on the spiking neural network and decision tree method," *Int. J. Elect. Comput. Eng.*, vol. 13, no. 2, pp. 2278–2288, 2023.

[33] *Dados horários do monitoramento da qualidade do ar—MonitorAr*. Accessed: Jan. 16, 2023. [Online]. Available: https://www.data.rio/datasets/dados-hor%C3%A1rios-do-monitoramento-da-qualidade-do-ar-monitorar/explore

[34] M. O. Arowolo, R. O. Ogundokun, S. Misra, J. Oluranti, and A. F. Kadri, "K-nearest neighbour algorithm for classification of IoT-based edge computing device," in *Artificial Intelligence for Cloud and Edge Computing, Internet of Things*, S. Misra, A. K. Tyagi, V. Piuri, and L. Garg, Eds. Springer, 2022, pp. 161–179.

[35] H. Pan, Z. Ye, Q. He, C. Yan, J. Yuan, X. Lai, J. Su, and R. Li, "Discrete missing data imputation using multilayer perceptron and momentum gradient descent," *Sensors*, vol. 22, no. 15, pp. 1–23, 2022.

[36] N. U. Okafor, Y. Alghorani, and D. T. Delaney, "Improving data quality of low-cost IoT sensors in environmental monitoring networks using data fusion and machine learning approach," *ICT Exp.*, vol. 6, no. 3, pp. 220–228, Sep. 2020.

[37] H. Tsai, L. P. Truong, and W. Hsieh, "Design and evaluation of wireless power monitoring IoT system for AC appliances," *Energies*, vol. 16, no. 1, pp. 1–27, 2023.

[38] A. R. Zarzoor, N. A. S. Al-Jamali, and I. R. K. Al-Saedi, "Traffic classification of IoT devices by utilizing spike neural network learning approach," *Math. Model. Eng. Problems*, vol. 10, no. 2, pp. 639–646, Apr. 2023.

NADIA ADNAN SHILTAGH AL-JAMALI (Member, IEEE) received the B.Sc. degree in control and systems engineering, the M.Sc. degree in control engineering, and the Ph.D. degree in computer engineering from the University of Technology, Baghdad, Iraq. Her current research interests include computer control, wireless sensor networks, intelligent systems, neural networks, and robotics.

IBTESAM R. K. AL-SAEDI received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Technology (UoT), Baghdad, Iraq, in 1987, 1994, and 2003, respectively.

From 2004 to 2010, she was a Scientific Researcher with Freiburg University, Germany, IMTEK, and Informatica Institutes. From 2003 to 2023, she was a Faculty Member and an Assistant Professor with the Electromechanical Engineering Department and the Communication Engineering Department, UoT. Her duties include teaching, supervision, consulting, and management. She was a Visiting Researcher for many international universities, such as Technical University Berlin, in 2011; Auburn University, AL, USA, in 2013; and Melbourne University, Australia, in 2018. She has extensive international experience, having been invited to teach and conduct research at prominent academic institutions with the University of Louisville (UofL), KY, USA. She is involved in several research projects, including a study on the Internet of Things and wireless communication with the Department of Electrical and Computer Engineering.

AHMED R. ZARZOOR received the M.Sc. degree in software engineering from the University of Bradford, Bradford, U.K., in 2006, and the Ph.D. degree in computer science from the Informatics Institute for Post-graduation Studies Iraqi Commission for Computer and Informatics, Baghdad, Iraq, in 2019. He is currently the Director of information technology with the Ministry of Health, Baghdad, Iraq. His current research interests include WSN, the IoT, MANET, computer networks and security, and soft computing.

HONGXIANG LI received the Ph.D. degree in electrical engineering from the University of Washington-Seattle, in 2008. Currently, he is an Associate Professor with the Department of Electrical and Computer Engineering, University of Louisville. He has more than 20 years of experience in the research and development of wireless communication and networking systems. His research has been funded by the National Science Foundation (NSF), the National Aeronautics and Space Administration (NASA), and the Office of Naval Research (ONR). His current research interests include big data analytics and the application of machine learning to communication systems and spectrum optimization. He was a recipient of the ORAU Ralph E. Powe Junior Faculty Enhancement Award, in 2012. He received the NASA Glenn Faculty Fellowship Program (NGFFP) Award, in 2013, 2018, and 2019. He served as an Editor for IEEE Communications Letter.