

Gene set imputation method-based rule for recovering missing data using deep learning approach

Amer Al-Rahayfeh¹, Saleh Atiewi¹, Muder Almiani², Ala Mughaid³, Abdul Razaque⁴, Bilal Abu-Salih⁵,
Mohammed Alweshah⁶, Alaa Alrawajfeh⁷

¹Department of Computer Science, Al Hussein Bin Talal University, Ma'an, Jordan

²Management Information Systems, Gulf University for Science and Technology, Hawally, Kuwait

³Department of Information Technology, Faculty of Prince Al-Hussien Bin Abdullah 2 for IT, The Hashemite University, Zarqa, Jordan

⁴School of Computing, Gachon University, Seongnam-si, Republic of Korea

⁵King Abdullah 2 School of Information Technology, The University of Jordan, Amman, Jordan

⁶Prince Abdullah Bin Ghazi Faculty of Information and Communication Technology, Al-Balqa Applied University, Salt, Jordan

⁷Department of Financial and Administrative Sciences, Ma'an College, Al-Balqa Applied University, Maan, Jordan

Article Info

Article history:

Received Jul 18, 2024

Revised Mar 26, 2025

Accepted May 24, 2025

Keywords:

Convolutional neural network

Data imputation

Kernel correlation filter

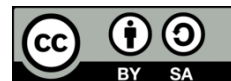
Spectral clustering

Water quality monitoring

ABSTRACT

Data imputation enhances dataset completeness, enabling accurate analysis and informed decision-making across various domains. In this research, we propose a novel imputation method, a spectral clustering based on a gene set using adaptive weighted k-nearest neighbor (AWKNN), and an imputation of missing data using a convolutional neural network algorithm for accurate imputed data. In this research, we have considered the Kaggle water quality dataset for the imputation of missing values in water quality monitoring. Data cleaning detects inaccurate data from the dataset by using the median modified Weiner filter (MMWFILT). The normalization technique is based on the Z-score normalization (Z-SN) approach, which improves data organization and management for accurate imputation. Data reduction minimizes unwanted data and the amount of capacity required to store data using an improved kernel correlation filter (IKCF). The characteristics and patterns of data with specific columns are analyzed using enhanced principal component analysis (EPCA) to reduce overfitting. The dataset is classified into complete data and missing data using the light- DenseNet (LIGHT DN) approach. Results show the proposed outperforms traditional techniques in recovering missing data while preserving data distribution. Evaluation based on pH concentration, chloramine concentration, sulfate concentration, water level, and accuracy.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Saleh Atiewi

Department of Computer Science, Al Hussein Bin Talal University

Ma'an, Jordan 71110

Email: saleh@ahu.edu.jo

1. INTRODUCTION

Water is essential for life on Earth. However, many nations experience freshwater scarcity. They were extremely driven to use other resources as a result of this worrying situation. For instance, Gulf nations use a laborious desalination process to obtain fresh water from the sea [1]. Multi-metric indices and operational indicators were used to monitor water quality across time and space [2]. However, this process becomes remarkably difficult because of the increased development along the coast and the resultant water contamination. To acquire freshwater, some nations process rainfall. Rainfalls have recently been affected by climate change, jeopardizing this possibility [3]. Even nations with easier access to fresh water still suffer

from water-related problems. The major goal of this study is to raise awareness of water contamination in the general population. The World Health Organization (WHO) and the United States Environmental Protection Agency (USEPA) often provide updates and suggestions on handling newly discovered illnesses and water toxins. In addition to research showing the effects of pollution and global warming on water supplies, the World Water Council (WWC) forecasts a 40% to 50% increase in world population over the next 50 years [4], [5]. The substantial increase, coupled with urbanization and industry, has the potential to elevate the overall water demand significantly.

The warning mentioned above signs leads to a future worldwide water disaster. Freshwater has become an industrial commodity on the verge of such a water catastrophe. In metropolitan locations, it is often kept in overhead or underground tanks under municipal management, sometimes for lengthy periods before consumption. Thus, continuous analysis of water quality is essential to categorize water for use and avoid waste. For instance, water that cannot be consumed might be used for cleaning [6], [7]. Human activities, including agriculture, industrial manufacturing, and dumping of urban effluent, have caused a decline in water quality. In many areas of modern civilization, including the economics, ecology, and environment, poor water quality is pervasive and has a detrimental impact on many different components of those systems. On this basis, the worldwide society has established objectives to improve the quality of freshwater resources [8], [9].

Globally, such objectives are often set out in guiding conventions that are sanction-free and voluntary. Sustainable Development Goal (SDG) 6 on water and its associated aim to improve water quality serves as a primary example. Some areas, such as the EU via the European Water Framework Directive, have legislation with legal punishments behind them in place at the regional level. Such objectives require advanced water quality monitoring schemes based on quantifiable and applicable water quality indices [10]–[12]. Committed countries have agreed to evaluate their freshwaters using the metric “Proportion of bodies of water with good ambient water quality” as part of the recognized SDG indicator. To evaluate the condition of a water body reasonably rapidly using known procedures, a set of parameter groups and particular parameters are utilized as measures. The parameters include oxygen, salinity, nitrogen, phosphorus, and acidification [13], [14]. Experience has demonstrated that monitoring becomes difficult because of the lack of data on these characteristics at the relevant temporal and geographical scales. We find gaps in the amount and quality of data provided for the associated disciplines of water, sanitation, and [15], [16]. Therefore, the problem faced by experts in scientific monitoring and organizations for monitoring operations is narrowing the worldwide data gap on water quality. Many variables affecting the performance of monitoring systems have been found in water quality monitoring [17]. The capabilities of monitoring agencies, including factors relating to human capacity, financing of monitoring operations, and the accessibility of technological equipment, stand out among them. However, the majority of this research on water quality monitoring focuses on certain applications, including drinking water [18], [19]. In the present study, we present a new imputation approach, the spectrum clustering based on a gene set using adaptive weighted k-nearest neighbor (AWKNN), and the imputation of missing data using the convolutional neural network (CNN) algorithm for reliable imputed data.

Handling missing data in water quality monitoring has faced many challenges in recent years in terms of pre-processing, data profiling, and imputation. The existing works provide achievable results but still lack an effective solution. Some of the major problems are as follows. High complexity: In some previous studies, imputation of data was performed based on the analysis, and data in the dataset were not divided (*i.e.*, complete data and missing data). However, complete data remain in the, thereby leading to complexity. In addition, the consideration of unwanted data in the dataset increases the amount of capacity and errors, resulting in complexity. Inaccurate imputation: The pre-processing of data improves the overall performance. The existing works perform pre-processing, but noises are not removed desirably, thereby increasing the inaccurate imputation of missing data. In several existing works, missing data validation was based on a crosshead attention mechanism and consistency check. Whereas rule-based validation was not performed, leading to inaccurate missing data handling. Poor quality of service (QoS): In some existing works, the imputation of missing values based on three rules (*i.e.*, missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR)) and duplicate data was not verified and validated. However, the duplicate data remains the same, resulting in poor QoS. In addition, data profiling was performed using column profiling (*i.e.*, analysis of characteristics). On the contrary, the lack of value repetitions considered leads to poor data profiling and affects QoS.

In the current years, handling missing data for water quality monitoring is an emergent issue in the realm of pre-processing, profiling, and imputation. The available works can indeed provide good results, yet there are no better approaches. The major problems facing this approach are enumerated hereafter:

- a. High complexity: The existing data has been imputed based upon their previous research analysis, and data did not divide in the dataset-both complete data and missing data. But if complete data existed in the

- dataset, complexity arises. In addition to that, unwanted data, along with considering the amount of capability inside datasets raises it, which results in bringing errors into the dataset followed by complexity
- b. Inaccurate imputation: Existing works pre-process data but the noises are not removed in a desirable manner which increases the inaccuracy of missing data imputation. Many exciting works validated missing data using cross-head attention mechanisms and consistency checking. But it does not perform rule-based validation, so incorrect handling happens for missing data.
 - c. Poor QoS: In some of the existing work, the missing value imputation by using three rules (*i.e.* MCAR, MAR, NMAR) and the duplicate data were not validated and confirmed. However, in such cases, when duplicate data exist in the same; it causes bad QoS. Moreover, profiling based on column profiling is conducted such that the characteristic of analyzing lacks considering repetitions of values thereby causing Poor profiling of data that affect the QoS.

The major objective of this research is to reduce the complexity, increase the QoS, and accurately impute, as well as perform profiling of the missing data. Some objectives of this research are as follows: i) Enhanced data quality improves the data with missing and noisy data identifications and reduction of unwanted data, which also helps increase the accuracy of imputation; ii) To reduce the complex nature of imputation while using rule-based validation, and data profiling that can be used to reduce latency because it identifies duplicate data, mainly through checks and rules; and iii) It classifies the datasets and executes imputation based on their gene set with the use of spectral clustering to improve accuracy. It reduces processing time and increases imputation accuracy.

The main purpose of this research is to handle missing data for accurate water quality monitoring using a water quality dataset. Some of the specific highlights of this research are as follows: i) Sophisticated techniques including the improved kernel correlation filter for effective data-reduction methods, Z-score normalization to standardize its values, and the median modified Weiner filter for noise elimination during the preprocessing step; ii) Among its applications, two of them include enhanced principal component analysis in terms of column and cross-column profiling and rule-based validation in consistency, uniqueness, and existence checks to identify trends and ensure data integrity with possible quality concerns; iii) Adaptive weighted k-nearest neighbors clustering (AWKNNC) assures exact clustering along with flexibility to large datasets; it groups data according to the kinds of missing values, and the LIGHT DenseNet model divides datasets into full and missing data categories; and iv) In the case of similarity-based clustering, missing data is handled with very high accuracy by a CNN, which can reconstruct data in real-time for water quality monitoring.

The structure of this paper shows how the proposed method could be useful in dealing with problems arising from missing data imputation in water quality monitoring. The introduction, states problems related to water quality in different parts of the world, indicates weaknesses of existing imputation methods, specifically including high complexity, errors, and poor quality of service, and puts into perspective the remedy suggested. The literature review studies relevant work, highlighting its limitations and evidencing the gap for an improved and more accurate imputation process, which this work tries to provide. The novelty architecture that involves spectral clustering, AWKNN, CNN, and advanced data preprocessing techniques, such as median modified Weiner filter (MMWFILT), Z-score normalization (Z-SN), and improved kernel correlation filter (IKCF) are discussed elaborately in the section proposed methodology. It can be evidenced that these newer techniques increase accuracy and minimize complexity. The proposed method has some important performance metrics better than the current ones in pH concentration, chloramine, and sulfate, and therefore its utility is shown. It is depicted in experimental results and discussion. The contribution of this study, improvement upon the existing ones, and the potential future directions of the work have been stated in conclusion and future work. The section contributes to the manuscript's impact on the quality monitoring system for water.

2. LITERATURE SURVEY

In this section, existing research on gene set imputation methods and rule-based approaches for recovering missing water quality monitoring data is discussed. The research has introduced machine learning and statistical strategies for addressing the issue of missing values, for example, support vector regression (SVR), hybrid decomposition-imputation models, and sequence-to-sequence learning-based models. Though these approaches have been promising in some contexts, they are generally marred by higher computational costs, ineffective rule-based validation, and poor data profiling, resulting in errors during the imputation of missing data. Additionally, conventional imputation methods, including k-nearest neighbor (KNN) and classification-based approaches, are not discriminatory between complete and missing data, compromising data quality and consistency at large. Hence, there is still a need for an adaptive and more efficient method that combines clustering algorithms, deep learning, and rule-based checking to refine imputation quality with low computational overhead.

2.1. Related works

The author in [20], proposed an SVR-based approach using a machine learning technique in filling missing values for data involving water quality. It covers the following: model selection profiling, and pre-processing the data. Despite success with SVR filling all missing values, there had been several extensive pre-processing that contributed towards extra complexities and energy requirements; there was no data preprocessing and noise reduction involved on its part, which may consequently be caused by wrong imputations with low quality of datasets.

Study [21] developed a two-headed sequence-to-sequence missing imputation model for time-series data by using the cross-head attention mechanism. Though the model was very effective in terms of imputing the missing time series, there was no rule-based validation and hence introduced a lot of errors. Moreover, processing extraneous data added complexity and decreased efficiency as a whole.

In study [22], a hybrid approach to wastewater treatment plants (WWTP) is proposed that combines decomposition and imputation. machine learning for univariate imputation that does not make a distinction between missing and complete data. machine learning. QoS was reduced due to redundancy as well as unnecessary complexity occasioned by failure to categorize datasets. Accuracy on imputation decreased due to frequent overfitting and underfitting by the systems of machine learning.

In study [23], a novel approach was presented with an application of dummy full sequence matching combined with long short-term memory (LSTM) in imputing missing telemetry water level data. LSTM-based models had a heavy increase in computational complexity compared to others because they highly consumed resources and had tremendous training times. There also were unclear distinctions between completed and incomplete data, so this resulted in high latencies and delays.

In study [24], it applied different imputation methods including k-nearest neighbor, classification and regression trees (CART), and random regression imputation (RRRI) towards the recovery of missing hydrological data. Cross-head attention was a necessity in case it had to be validated. This set of techniques although performed excellently on the streams could be unreliable without a process of rule-based validation wherein the handling of missing data goes entirely inaccurate.

Study [25], proposed an approach for machine learning-based water quality prediction involving multivariate imputation from several measurements. The problem with the approach is that the method suffers from overfitting and underfitting and is successful only in degree categorization of water contaminants. Noise and complexity within the dataset continued to be the challenging issues.

Study [26] developed a sliding window method for data imputation and anomaly detection in hydrological time series datasets. Even though it got the irregularities right, it did not distinguish between complete and missing data. This led to a lag in processing with increases in computing complexity.

In [27], support vector machine (SVM) had been used in the stage of imputing missing information about classification tasks. Although SVM increased the accuracy of classification, it failed to handle noise in the dataset effectively, which led to errors. Moreover, preparation was not enough to ensure the dependability of imputed data.

Study [28], suggested the application of multiple imputations in machine learning for predicting the quantity of chlorophyll-a in coastal areas. The model succeeded in predicting biological traits but could not differentiate between complete and missing values in datasets. This resulted in making the imputation procedure complicated with both missing and full data

Study [29] evaluated different imputation techniques for network data, ranging from simple imputation to complex model-based approaches. Although they worked very well in certain situations, these methods could not effectively handle redundant or duplicate data. It is leading to inconsistencies in the imputed results.

Study [30] proposed the imputation of missing network data to improve sample coverage in the presence of complete and incomplete networks. Here, we compare the efficacy of various imputation techniques, from straightforward imputation to sophisticated model-based approaches, over a broad spectrum of measurement, network, and missing value characteristics.

In study [11], KNN imputation and a multilayer perceptron model have been used for the quality prediction of water. Poor profile quality resulted because column profiling was employed in the processing of the data with no consideration given to value repeats. Quality of service (QoS) was restrained, and the general performance of the model degraded. Furthermore, we list existing objectives and issues in Table 1.

Research solution: Utilizing Kaggle's water quality data set, this study approaches the issue of managing missing values in the monitoring of the quality of water in an efficient manner. A gene-based imputation, profiling, and improvement in data quality through this work. After processing through the MMWFILT for noise removal, effective data transformation and dimensionality reduction are performed by Z-SN and IKCF. To ensure integrity in the data, column and cross-column profiling will make use of enhanced principal component analysis (EPCA) combined with rule-based validation of data that covers consistency, uniqueness, and existence checks. Adaptive weighted KNN clustering is used for gene-based clustering while dataset classification makes use of the LIGHT DenseNet model if there is a process of

imputation of missing data. The use of a CNN for weighted imputation, which can reconstruct whole data sets for real-time water-quality monitoring, will ensure excellent accuracy.

Table 1. Comparison of existing water quality monitoring methods

Ref.	Objective	Water quality monitoring methods	Limitations
[20]	To develop a machine learning approach for imputing water-quality data with a high percentage of missing values and address the challenge of missing data in water-quality measurements by applying machine learning techniques to accurately impute the missing values.	Inverse distance weighting (IDW), random forest regressor (RFR), ridge (R), Bayesian ridge (BR), AdaBoost (AB) method	Here, support vector regression was implemented to the imputation of missing values in the dataset. However, support vector regression required a large amount of data for processing, thereby increasing the complexity and energy consumption.
[21]	To improve the dual-head attention model for time series data imputation to enhance the accuracy and efficiency of imputing missing values in time series data.	Dual-head sequence-to-sequence imputation model	Here, missing data validation was based on a crosshead attention mechanism, whereas rule-based validation cannot be performed, leading to inaccurate missing data handling.
[22]	Using a univariate imputation method in wastewater treatment increases the efficiency and productivity of the overall process.	WWTP integrating decomposition method	Here, the imputation of data was performed based on analysis, and data are not classified (<i>i.e.</i> , complete data and missing data) in the dataset. However, missing data remain and combined in the dataset, leading to complexity.
[23]	To propose and develop innovative techniques that can accurately and effectively fill in gaps in water level measurements collected from monitoring systems.	LSTM method	Here, the imputation of missing data was based dummy full sequence scheme. However, complete and incomplete data are not classified separately, thereby increasing the processing time and leading to high latency.
[24]	To recover missing data in hydrological studies to determine the most effective and accurate approach for handling missing data in this context.	RRRI, CART, and KNN method	Crosshead attention was used to validate missing data, whereas rule-based validation is proven ineffective for the same purpose, resulting in inaccurate treatment of missing data.
[25]	To develop a machine learning predictive model that can accurately detect water quality and pollution levels based on various parameters and data inputs, such as chemical composition, physical properties, and environmental factors.	Predictive model using machine learning	Here, machine learning algorithms were utilized for the imputation of missing values. These algorithms always produce overfitting or underfitting. This condition leads to high errors because it was unsuitable for imputation of missing data.
[26]	To develop a methodology for anomaly detection in hydrological time series data using a sliding window technique and data imputation with machine learning.	Long short-term memory method	Here, the imputation of missing data was based on dummy full sequence scheme. However, complete and incomplete data are not classified separately, increasing the processing time and leading to high latency.
[27]	To explore and identify effective techniques for handling missing values within datasets used for classification tasks, specifically using machine learning methods.	Support vector machine method	Here, the presence of noise in the dataset prevents the removal of missing data. However, inappropriate data remain unchanged despite attention, resulting in inaccurate imputation of missing data in the dataset.
[28]	To create a forecasting model for the concentration of chlorophyll-a in the Korean coastal zone using machine learning and multiple imputation techniques.	Six machine learning algorithms	Here, the imputation of data was performed based on three rules, and data are not classified (<i>i.e.</i> , complete data and missing data) in the dataset. However, missing data remain and integrated in the dataset, leading to complexity.
[29]	To describe and implement a missing data imputation algorithm specifically designed for transmission systems.	Korea electric power corporation method	Here, the unwanted data are not reduced in the dataset. However, it increases the amount of capacity, leading to complexity
[30]	To explore and evaluate various imputation methods for missing network data, considering different network structures and patterns of missing data, and to enhance the accuracy and representativeness of network sampling coverage	Simple imputation to more complex model-based approaches	Here, the imputation of missing values is performed based on three rules, and duplicate data are not checked and validated. Whereas, the duplicate data remain the same, leading to poor QoS.
[11]	To analyze and predict water quality parameters by imputing missing values in the dataset using the KNN imputer and then using an MLP model to predict the water quality parameters using the available data accurately.	KNN imputer method	Here, data profiling was performed using column profiling (<i>i.e.</i> , analyzing characteristics). By contrast, failure to consider value repetitions leads to poor data profiling, thereby limiting the QoS.

3. PROPOSED METHOD

In this work, we mainly focus on handling missing data in water quality monitoring. In addition, the classification of the dataset is based on the completeness level of the missing data. The Kaggle water quality dataset is considered for imputation of missing values in water quality monitoring. Several processes

involved in the proposed work are categorized into three main segments, namely: i) data quality enhancement, ii) data profiling and rule-based data validation, and iii) gene-based imputation of data.

3.1. Data quality enhancement

Data quality improvement is an essential phase of enhancing missing data imputation reliability in water quality monitoring. Pre-processing is comprised of several steps: data cleaning, transformation, and reduction, to remove inconsistencies, normalize formats for data, and improve the efficiency of storage. The MMWFILT identifies and eliminates noise and errors, enhancing data integrity. Z-SN normalizes values within attributes, making comparison and analysis easier. The improved IKCF eliminates duplicate data, minimizing computational complexity while maintaining necessary information. All these processes collectively improve data quality, resulting in more precise missing data.

3.1.1. Data cleaning

Data cleaning is the process of detecting inaccurate data from the dataset by using MMWFILT. This filter detects the missing data (*i.e.*, incorrect data). Cleaning data maintains data quality and enables more accurate imputation.

3.1.2. Data transformation

Data transformation is an essential pre-processing technique to change the data format and structure. Several processes are described as follows: In data smoothing generally, data have many noises, which degrade the detection accuracy. Noise removal is executed to eliminate undesired elements (*i.e.*, incorrect data) from the dataset using the MMWFILT approach. This filtering method effectively removes noises and unwanted data.

The MMWFILT noise reduction method was applied to model data related to water contaminants, including pH, hardness, sediment, chloramine, sulfate, and conductivity. These data were collected from environmental sensors and can be efficiently processed using local filters tailored to the geographical domain, ensuring swift and efficient text processing. Recently, data collected for water pollution monitoring pre-dictions turned out to be inaccurate, containing missing or erroneous information.

After appropriate adjustments of the mask size around the surrounding area of the target pixel, the reduction techniques using spatial-domain-based local filters are built upon a predetermined equation. However, owing to blurring effects brought about by excessive smoothing, the picture properties of traditional spatial filters deteriorate. MMWFILT is a traditional local filter based on the spatial domain that combines the benefits of the Wiener filter and the median filter in a nonlinear adaptive filter. In our method, the Wiener filter, which processes images based on the variance of Gaussian noise, was used to simulate the MMWFILT algorithm by substituting the mean value of the pixels inside the mask with the median value. Thus, the Wiener filter is expressed as (1):

$$h_{wiener} = \mu + \frac{\sigma^2 + u^2}{\sigma^2} (g(q, p) - \mu), \quad (1)$$

where μ and σ denote the mean and standard deviation values of the pixels located within the mask, respectively, and v is the standard deviation of the noise.

The Wiener filter is excellent in reducing noise because it considers all the pixel values in the region of interest (ROI). However, when the mean value is placed into the Wiener filter equation, high-frequency signals, such as those of the edge area, are lost. On the contrary, when a particular pixel value is selected during image processing, the median value more successfully maintains high-frequency signals while minimizing noise than the mean value. Thus, the MMWFILT is obtained as follows:

$$h_{mmwf} = \hat{\mu} + \frac{\sigma^2 + u^2}{\sigma^2} (g(q, p) - \hat{\mu}) \quad (2)$$

where $\hat{\mu}$ denotes the median value of the pixels located inside the mask.

We created a median modified wiener filter (MMWF) model based on (2) to enhance the data cleaning process. To ensure its effectiveness across various image resolutions, the MMWFILT's mask sizes were carefully configured to 3×3, 5×5, 7×7, 9×9, and 11×11. These varying mask sizes allow for flexible application of the filter to captured images with different matrix dimensions, enabling precise noise reduction without compromising data structure.

Data normalization Here, data normalization organizes data entries to ensure that they appear similar across all fields and records. As a result, information is easier to find, grouped, and analyzed. In the proposed work, the normalization technique is based on the Z-SN approach, which provides improved data organization and management for an accurate imputation.

Z-SN is a broad statistical method that may be used with various data types, including data from water quality monitoring. Z-SN may be used to normalize several water quality metrics to a single scale in water quality monitoring. Thus, comparing and evaluating the data becomes relatively simple. The equation (3) for Z-SN is as:

$$z = \sigma x - \mu, \quad (3)$$

where z is the z-score of the data point x , μ is the mean of the dataset, and σ is the standard deviation of the dataset.

The following procedures are used to apply Z-SN to data for water quality monitoring:

- The parameter's mean (μ) and standard deviation (σ) are determined for water quality that requires normalization. Depending on the research objectives, normalization can be carried out either separately for each parameter or the entire dataset as a whole.
- To obtain the z-score for each data point x in the water quality parameter, the values are integrated into the z-score equation.
- The generated z-scores indicate how many standard deviations a data point deviates from the mean. A positive z-score suggests that the data point is above the mean, whereas a negative z-score indicates that it is below the mean.

3.1.3. Data reduction

Data reduction is the process of eliminating unwanted data (*i.e.*, repetitions of readings) and the amount of capacity required to store data using IKCF. This filtering method rapidly reduces the amount of information stored in the system using some methods. However, data reduction can increase storage efficiency and performance and minimize storage costs. The suggested technique based on the IKCF tracker is briefly described here. The IKCF tracker constructs a training set by cyclic shifting. Suppose the base vector $x = (x_1, x_2, \dots, x_n)^T$, Q is a permutation matrix:

$$P = \begin{bmatrix} 0 & 0 & 0 & \dots & 1 \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}, \quad (4)$$

Then one of the cyclic shifts of x can be expressed as $Q_x = (x_n, x_1, \dots, x_{n-1})^T$, which represents moving x one position to the right. By constantly left multiplying the permutation matrix Q , $\{Q^u x | u = 0, \dots, n-1\}$ can realize the cyclic shift of base vector x for u times. The cyclic matrix X is formed by combining all x -shift cycles in a single matrix.

$$X = \begin{bmatrix} (P^0 x)^T \\ (P^1 x)^T \\ (P^2 x)^T \\ \vdots \\ (P^{n-1} x)^T \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_n \\ x_n & x_1 & x_2 & \dots & x_{n-1} \\ x_{n-1} & x_n & x_1 & \dots & x_{n-2} \\ \vdots & \vdots & \dots & \ddots & \vdots \\ x_2 & x_3 & x_4 & 1 & x_1 \end{bmatrix} \quad (5)$$

For any vector x , its cyclic matrix can be diagonalized by expression (6):

$$X = F \text{diag}(\hat{x}) F^H, \quad (6)$$

where \hat{x} is the discrete Fourier transformation of x , F represents the discrete Fourier transformation matrix, and F^H is the conjugate transpose of F . The IKCF tracker uses ridge regression to train the classifier. The main idea is to find a function $f(z) = f^T z$ that minimizes the mean square error between the output of all training samples and their expected output and the loss function, as (7):

$$\sum_{i=1} (f(x_i) - y_i)^2 + \lambda \|f\|^2, \quad (7)$$

where λ is a regularization parameter and $\lambda > 0$. λ is used to prevent the model from overfitting; x_i is the training sample of i ; y_i is the expected output of x_i . The following is a closed-form solution to (8) derived by obtaining the training samples:

$$f = (X^T X + \lambda I)^{-1} X^T y, \quad (8)$$

where X is a circular matrix of all training samples, y is the expected output vector, and I is the identity matrix. Directly solving the filter factor f involves a large number of matrix operations and a lengthy computation time. Using the properties of the circular matrix, (6) can be substituted into (8). Then, we obtain the (9):

$$X^T X = F \text{diag}(\hat{x}^*) F^H F \text{diag}(\hat{x}) F^H. \quad (9)$$

According to the properties of the Fourier transform matrix, $F^H F = I$, the solution of the filter in the frequency domain can be obtained by substituting (9) into (8) as:

$$\hat{f} = \text{diag} \frac{(\hat{x}^* \odot \hat{y})}{(\hat{x}^* \odot \hat{x}) + \lambda}, \quad (10)$$

where denotes dot product; \hat{f} , \hat{x} , and \hat{y} are discrete Fourier transforms of f , x , and y , respectively, and \hat{x}^* represent the complex conjugates of \hat{x} .

To improve the ability of the IKCF tracker to solve nonlinear problems, a kernel function is used to transform ridge regression problems in low-dimensional space into high-dimensional space $\phi(x)$, classify the samples in the high-dimensional space and solve the linear inseparability problem. Suppose the kernel function is $k^{xx'} = \varphi^T(x) \varphi^T(x')$, the formula $f(z) = f^T z$ can be written as:

$$f(z) = \sum_i^n \alpha_i \varphi^T(x_i) \varphi(z) = \sum_i^n \alpha_i k(x_i, z), \quad (11)$$

For most kernel functions, such as the Gaussian kernel, the polynomial kernel, and the linear kernel, the kernel matrix still has the property of a cyclic matrix. Therefore, α can be solved by (12):

$$\hat{\alpha} = \frac{\hat{y}}{\hat{k}^{xx} + \lambda}, \quad (12)$$

where \hat{k}^{xx} is the Fourier transform of the basis vector of the kernel matrix $k=C(k^{xx})$. For the Gaussian kernel $k^{xx'} = \exp(-\frac{1}{\sigma^2} (||x||^2 + ||x'||^2, k^{xx'}))$ can be expressed as (13):

$$k^{xx'} = \exp(-\frac{1}{\sigma^2} (||x||^2 + ||x'||^2 - 2\mathcal{F}^{-1}(\hat{x} \odot \hat{x}')))) \quad (13)$$

The equation (14) is used to determine the response map:

$$f(z) = \mathcal{F}^{-1}(\hat{k}^{xz} \odot \hat{\alpha}), \quad (14)$$

where \hat{k}^{xz} is the kernel correlation Fourier transform of samples x and z . In the current frame, the item is located at the coordinates that provide the highest response map value. Updating the filter template increases the tracking reliability.

$$\begin{cases} \hat{x}_t = (1-\eta)\hat{x}_{t-1} + \eta\hat{x}_t \\ \hat{\alpha}_t = (1-\eta)\hat{\alpha}_{t-1} + \eta\hat{\alpha}_t \end{cases}, \quad (15)$$

where \hat{x}_t and $\hat{\alpha}_t$ are the features obtained from frame t , and η is the learning rate.

3.2. Data profiling and rule-based data validation

After data pre-processing, data are analyzed via the data profiling process. Data profiling involves examining, analyzing, and creating useful data summaries. This process yields a high-level overview that aids in the discovery of data quality issues, risks, and overall trends. This approach also discovers, understands, and organizes data. Several processes are described in the following section.

3.2.1. Column profiling

Column profiling evaluates individual data columns for inconsistencies, missing values, and outliers. Improved principal component analysis (EPCA) helps detect correlation and avoid overfitting through the exploration of variance per column. Such processing structures water quality parameters such as

pH, sulfate, and chloramine content for increased reliability of the data. Successful profiling improves data integrity, reduces redundancy, and maximizes the accuracy of missing data imputation.

3.2.2. Cross column profiling

Here, cross-column profiling encompasses observing the values and counting the number of times each value shows up within each column using the EPCA approach. This method can effectively obtain the frequency distribution and patterns within a column of data. However, this approach identifies patterns in data by using the correlation features.

3.2.3. Data rule validation

Data rule validation is a proactive technique of verifying data instances, where data sets conform with predefined rules. This process improves data quality based on three checks, such as consistency checks, unique checks, and presence checks using the EPCA approach. Consistency check validation is an entity that confirms the consistency of node instances and the analysis result; it also contains read-only logic. The unique check is a process that examines data to identify rows with duplicate information. These duplicates may appear to be original data (e.g., 1.9979), but they exhibit slight variations (e.g., 1.9799) in their values within the table. Furthermore, the presence check is based on three rules, such as MCAR, MAR, and NMAR; it checks the presence of values in the required fields.

In the EPCA feature selection technique, two steps are used to select values for the feature selection investigation and categorization. These actions rely on removing superfluous elements, the elimination of features and replacement of each quality with the conditional mean or marginal mean. The PCA's fundamentals are studied and discussed in the following section. The random feature vector $X \in R^p$ is assumed to have distribution P . The vector X has the coordinates $X[i]$, $i = 1, 2, \dots, p$. The symbol for X 's covariance matrix is. EPCA has $< O(\min(p^3, n^3))$ time complexity. Memory usage is $< O(nd)$, where n is the total amount of data points, and d is the number of dimensions.

Algorithm 1. Enhanced PCA algorithm

Input: $X = \{x_1, x_2, \dots, x_n\}$ the dimension $x_i \in (R^M)$

Step 1: Original data are used to transform $N \times d$ matrix X into $N \times m$ matrix Y :

Step 2: The $d \times d$ covariance matrix is computed as follows:

$$C = \frac{1}{N-1} X^T X$$

$$C_{ij} = \frac{1}{N-1} \sum_{q=1}^N X_{q,i} X_{q,j}$$

Step 3: The covariance matrix's eigenvector is determined using an estimate.

Step 4: The eigenvalues (λ) and eigenvectors (V) are calculated as follows: $\bar{X} = \lambda V$

Step 5: Calculate dissimilar matrix

For a given random feature vector X , satisfying the assumption H1:

- i. If $E(\|\bar{X}\|^2) < \infty$ where $\|\bar{X}\|^2$ is $(\bar{X} - (X, \bar{X}))$ then
- ii. After the covariance matrix of \bar{X} is positive definite.
- iii. All covariance matrices have different eigenvalues.

//The first principal component is defined as:

$$\alpha^1(P) = \alpha^1 \max Var(\alpha^1 X) = \operatorname{argmax}_{\alpha^1} \alpha^1 \Sigma \alpha^1_{\|\alpha^1\|=1}$$

//next principal components are defined as:

$$\alpha^k(P) = \alpha^k = \operatorname{argmax}_{\|\alpha^k\|=1, \alpha^k \perp [\alpha^1, \dots, \alpha^{k-1}]} Var(\alpha^k X)$$

where $\alpha^1, \dots, \alpha^{k-1}$ is the subspace generated by the vectors $\alpha^1, \dots, \alpha^{k-1}$.

Step 6: Local-based similarity calculation //intra-class similarity identification

The local objective function $H^l(I)$ as

$$H^l(I, P, P_{Yl}) = h^l(I) = \|\alpha^1(p) - \alpha^1(P_{Yl})\|^2$$

Local feature minimum distance calculation

$$I_{1,0} = \operatorname{argmin}_{I \in I_d} h^l(I)$$

Step 7: Global feature minimum distance calculation

The global feature objective function is

$$h(I) = \sum_{i=1}^q p_i h^l(I) \quad \text{with } p_i \geq 0, \sum_{i=1}^q p_i = 1, 2 \leq q \leq p$$

$$\widetilde{I}_{q,0} = \operatorname{argmin}_{I \in I_d} h(I)$$

Step 8: Classify the decomposed matrix with n input units of values,

$$x_i \in R, i = 1, 2 \dots n$$

Step 9: Calculate weight of the feature vector

$$y_c = f(\sum_{i=1}^n w_i x_i)$$

$$\Delta w_i = \eta \cdot y_c \cdot \varepsilon$$

Output: Weights-based features are selected

3.3. Gene-based imputation of data

After analysis and detection of missing values, the imputation of data is performed. Here, the dataset is classified into two such complete data (*i.e.*, completed values) and missing data (*i.e.*, missed values, incorrect values S duplicate values) using the LIGHT DN approach. This model has the advantage of fast training speed and is suitable for handling large-scale datasets. The LIGHT DN model can achieve high classification accuracy. Furthermore, the missing data are imputed using imputation methods and in this imputation method, datasets having missing values are classified into complete genes and incomplete genes using the LIGHT DN model. Here, the complete gene consists of incorrect data, duplicate data, and half missing data and the incomplete gene consists of fully missing data. Figure 1 shows the overall flow diagram.

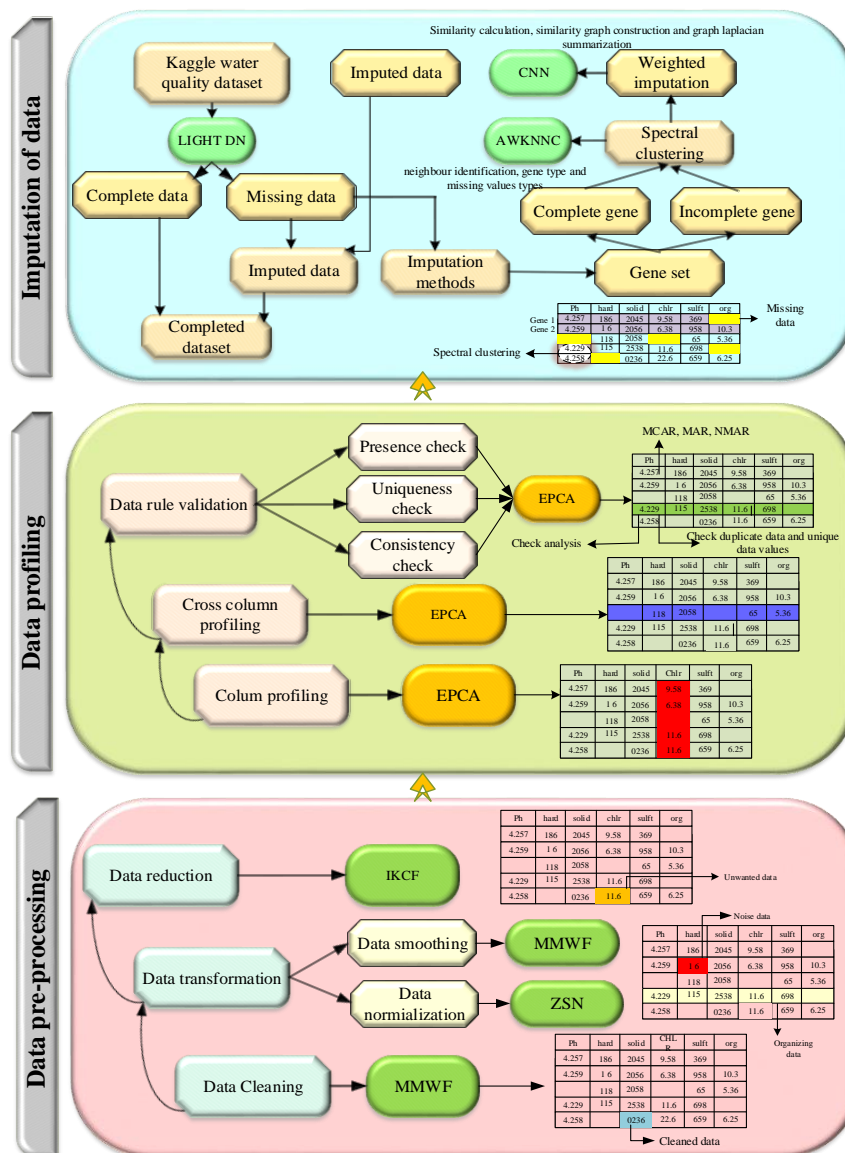


Figure 1. Overall flow diagram

3.3.1. LIGHT DenseNet

The light DenseNet is built using alternated dense and transition blocks. Then, a fully linked layer and a SoftMax classifier are shown in Figures 2(a) and 2(b), respectively. A batch normalization (BN) layer, a convolutional layer, and a leaky rectified linear unit (LReLU) layer (BN–Conv–LReLU) make up the two cascaded convolutional units of the dense block. The first BN–Conv–LReLU in the dense block, as shown in Figure 2, generates $4r$ output feature maps using kernels of size 11, whereas the second BN–Conv–LReLU generates r feature maps using 33 kernels, where r is a constant. The dense block increases the number of maps by concatenating the input maps with the r output feature maps. The transition unit includes a convolutional layer, a 2×2 average pooling layer, and a 2×2 output layer. The goal of the transition block is to minimize computation by finding an optimal combination of feature maps generated by the various convolutional layers.

The specifics of the structure and output for the condensed DenseNet are provided in Table 2, assuming that the size of the input pictures is 128×128 . In Table 2, the terms “conv” and “ $1 \times 1 \times 64$ conv” refer to convolutional units; “ 2×2 pool” denotes an average pooling layer, a pool size of “ 2×2 ” denotes the number of final output classes, and “[\cdot] $\times 2$ ” denotes that the structure “[\cdot] is repeatedly cascaded for two times.

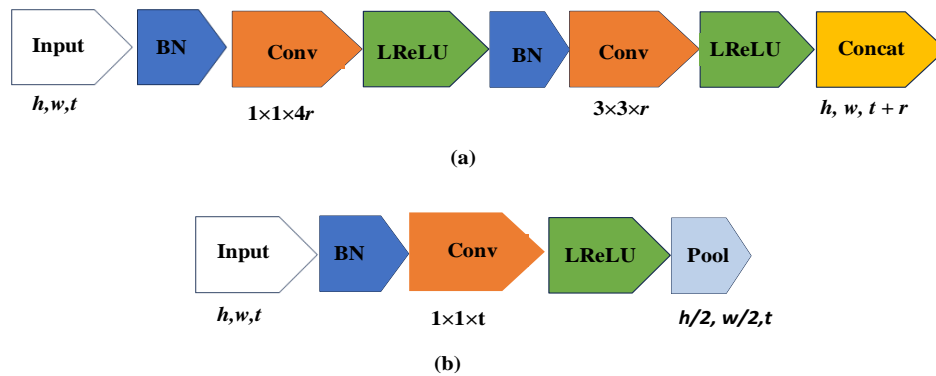


Figure 2. Structure details of (a) the dense block and (b) the transition block. The dimensions of the input feature maps are represented as h , w , and t

Table 2. Architecture of the DenseNet

Module	Detail	Output
Convolution	$128 \times 128 \times 64$	$1 \times 1 \times 64$ conv
Max-pooling	$64 \times 64 \times 64$	2×2 pool
Dense block I	$\times 2$	$1 \times 1 \times 128$ conv
	$64 \times 64 \times 128$	$3 \times 3 \times 32$ conv
Transition block I	$32 \times 32 \times 128$	$1 \times 1 \times 128$ conv
		2×2 pool
Dense block 2	$\times 3$	$1 \times 1 \times 128$ conv
	$32 \times 32 \times 224$	$3 \times 3 \times 32$ conv
Transition block 2	$16 \times 16 \times 224$	$1 \times 1 \times 224$ conv
		2×2 pool
Dense block 2I	$\times 4$	$1 \times 1 \times 128$ conv
	$16 \times 16 \times 352$	$3 \times 3 \times 32$ conv
Transition block 2I	$8 \times 8 \times 352$	$1 \times 1 \times 352$ conv
		2×2 pool
Dense block IV	$\times 2$	$1 \times 1 \times 128$ conv
	$8 \times 8 \times 416$	$3 \times 3 \times 32$ conv
Global average pooling	$1 \times 1 \times 416$	-
Full-connection	θ	416×0 full-connection
SoftMax	θ	SoftMax classifier

Effective feature maps may be generated using the design above strategy by using the following techniques:

- In the convolutional layers, modest-sized kernels, such as 1×1 or 3×3 , are used to fine-tune the number of feature maps and learn to generate meaningful feature maps.
- Using intermediate feature maps and benefiting from valuable error feedback, shortcut connections become a preferred option for training DenseNet.

- c. To prevent introducing a high number of weights, the output of the last convolutional layer is not flattened using the global average pooling (GAP) approach between the final transition block and the fully connected layer. The suggested hybrid network becomes interpretable because the GAP approach may be utilized to examine the network's final feature maps.
- d. To generate the required number of outputs from the SoftMax classifier, a single fully connected layer is used.

Then, the gene set is spectrally clustered using the AWKNNC algorithm based on neighbor identification, gene type, and missing value types. This algorithm rapidly identifies cluster centers, demonstrating excellent adaptability across various clustering tasks. The imputation process enhances accuracy, remarkably boosting the imputation capability of the dataset.

3.3.2. Adaptive Weighted k-nearest neighbor (AWKNN)

The AWKNN positioning system determines location by comparing online received signal strength (RSS) readings with the center of each cluster. It uses a standard cluster-matching approach to identify the appropriate clusters for accurate positioning. Given the complexity and changeability of interior space, the standard cluster-matching approach performs poorly. Our study diverges from cluster matching, shifting its emphasis toward an adaptive weighted KNN localization approach. After matching clusters, one cluster is then used to match with real-time RSS feeds.

We propose a new KNN-based affinity propagation clustering (APC) and adaptable weighted-based position estimation technique called AWKNN. KNN is a popular machine-learning algorithm for location estimation in fingerprint localization techniques. Our proposed AWKNN method selects an initial set of RPs by using the KNN algorithm to find the RPs with the least signal-domain distances to the currently available RSS feeds, where K is greater than 5. The first set of K RPs is clustered using APC, and then further sub-clustering occurs. The number of RPs and the signal-domain distance between the sub-cluster center and the online RSS readings are then used to reserve the most probable sub-cluster. The following is an inverse distance weighted method based on the hidden cluster that estimates the user's location. We have adopted the KNN method without changes other than the computation of the average signal-domain distance. Thus, we will not discuss it in depth in this section. The first-level KNN method yields K initial RPs with signal-domain distances. High-quality clustering results are produced automatically by APC, splitting them into many sub-clusters. Using APC, the K initial RPs may be split down into smaller groups.

Subsequently, we figure out the location of the user. In a nutshell, the suggested AWKNN algorithm's central concept is to select a small set of highly concentrated RPs for calculating weighted average coordinates based on their locations and similarities (signal-domain distances). On this basis, we can reduce the likelihood of significant positional errors in our calculations. This section provides an in-depth explanation of Algorithm 2. In the affinity propagation method, as opposed to the K-means or FCM clustering algorithms, the centers of clusters are fixed locations inside the sample data. The distance in the signal domain between the RP and the online RSS data is computed in line 9 of the aforesaid procedure. Therefore, the signal-domain distances between the centers of the sub-clusters and the live RSS readings may not be recomputed. For simplicity, we refer to the distances between the signal domain and the frequency domain as d_1 and d_2 , respectively. To estimate the user's location, we input their x and y coordinates into (16), together with the remaining number of RPs in the aim sub-cluster and the corresponding signal-domain distance, d .

$$\begin{cases} \hat{y} = \sum_{i=1}^n \left(x_i \times \frac{1}{d_i} \right) / \sum_{i=1}^n \frac{1}{d_i} \\ \hat{x} = \sum_{i=1}^n \left(y_i \times \frac{1}{d_i} \right) / \sum_{i=1}^n \frac{1}{d_i} \end{cases} \quad (16)$$

Algorithm 2. Proposed AWKNN localization algorithm

1. Input: number of RPs in the selected cluster (n), online RSS readings r , K , selected cluster (C_s) after cluster matching
2. Output: weighted average coordinates
3. if $n < 10$ then
4. $K = n$
5. else
6. $K = 10$
7. end if
8. for $i = 0$ to n
9. do
10. signal-domain distances (d_{sig}) between each RP in C_s and r are calculated
11. end for
12. K initial RPs with K top smallest d_{sig} are obtained
13. By using affinity propagation clustering, these K initial RPs are split into numerous

```

sub-clusters, where  $N_c$  is the total number of sub-clusters.
14. if  $N_c = 1$  then
15.     using Equation (12), the locations for a set of  $K$  initial RPs are obtained.
16. else if  $N_c \geq K - 3$  then
17.     coordinates with three RPs that have the top smallest  $d_{sig}$  are calculated using
Equation (16)
18. else
19. the  $N_{RP}$ , or the number of RPs, in each cluster subset is determined.
20. ( $N_{RP}$ ) is sorted in descending order
21. two sub-clusters ( $C_{sub,1}$  and  $C_{sub,2}$ ) with two largest  $N_{RP}$ , that is,  $N_{RP1}$  and  $N_{RP2}$ ,
are selected, and the  $d_{sig}$  of sub-cluster centers is represented by  $d_1$  and  $d_2^1$ 
22.  $N_{diff} = N_{RP1} - N_{RP2}$ 
23. if  $N_{diff} \geq 3$  then
24. your position relative to  $N_{RP1}$  subcluster  $C_{sub,1}$  is identified
25. else
26.     if  $d_1 \leq d_2$  then
27.         coordinates with  $C_{sub,1}$  are calculated using Equation (16)
28.     else
29.         coordinates with  $C_{sub,2}$  are calculated using Equation (16)
30.     end if
31. end if
32. end if

```

The suggested AWKNN localization method is shown in Figure 3. Figure 3(a) shows how the KNN algorithm uses distances in the signal domain to identify the first 10 RPs, which are written in black. The APC algorithm then uses this information to automatically separate them into four distinct groups. As shown in Figure 3(b), four distinct colors are used to designate the various sub-clusters. One, two, three, and four are the possible numbers of sub-clusters. Lines 13–20 of the preceding algorithm description allow us to divide the RPs into two groups, one containing the top two RPs. The two subclusters in question are denoted by the red dotted box. Figure 3(c) shows the two groups with sub-clusters 1 and 2. Line 26 presents, in detail, the calculation of the signal-domain distance between the cluster's corresponding center and the live RSS values to identify the target sub-cluster. The red dotted box in Figure 3(d) depicts the target sub-cluster for the subsequent estimation location.

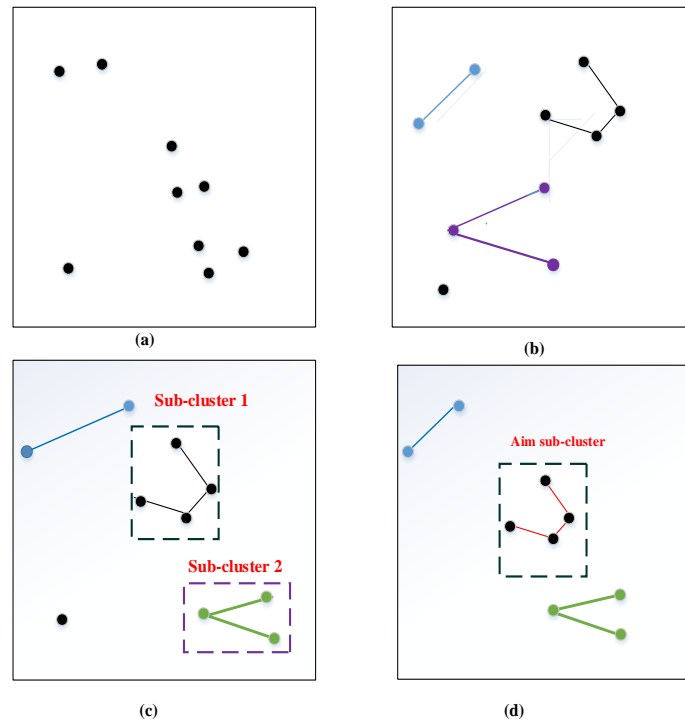


Figure 3. Illustration of the AWKNN algorithm: (a) KNN-based selection of the first 10 RPs and their clustering using the APC algorithm, (b) visualization of four distinct sub-clusters in different colors, (c) division of RPs into two groups with sub-clusters 1 and 2, and (d) identification of the target sub-cluster for estimation

After clustering, the weighted imputation of data is performed using a CNN algorithm based on similarity calculation, similarity graph construction, and graph Laplacian summarization. This model has high accuracy in the imputation of data in the large-scale dataset and minimizes the computation. Finally, the imputed and complete data are combined to form the complete dataset. We accurately handle the missing data in the water quality monitoring. Therefore, we can efficiently impute the missing data and improve real-time water quality monitoring.

3.3.3. Convolutional neural network

As a feedforward neural network, CNN does not require any further image processing prior to inputting the raw data. The discipline of pattern recognition has benefited greatly from its use in recent years. CNN can abstractly represent features by extracting them from input data layer by layer. Input, convolution, pooling, fully connected, and output layers are common components of convolutional neural networks, as shown in Figure 4. After performing convolution calculations on the input data using multiple convolution kernels, extracting the associated data features, and connecting to the next layer through bias calculation and activation function, the convolution layer is finalized. A mathematical equation can express the process, as (17).

$$X_i = \sigma(X_{i-1} \times W_i + v_i), \quad (17)$$

where X_i is the i^{th} layer's output feature map, X_{i-1} is the i^{th} layer's input feature map, W_i is the i^{th} convolution kernel's weight matrix, v_i is the i^{th} layer's offset vector, and sigma is the activation function. Tanh, sigmoid, and ReLU are the primary activation functions. Maximum pooling, mean pooling, and random pooling are the three main types of pooling layers used to scale down the convolution layer's output parameters. In the fully connected layer, features from the preceding layer are reconnected while also being extracted and reduced in dimension. Finally, the output layer determines the probability value for each class to which the classification target belongs and produces the 1D output sequence.

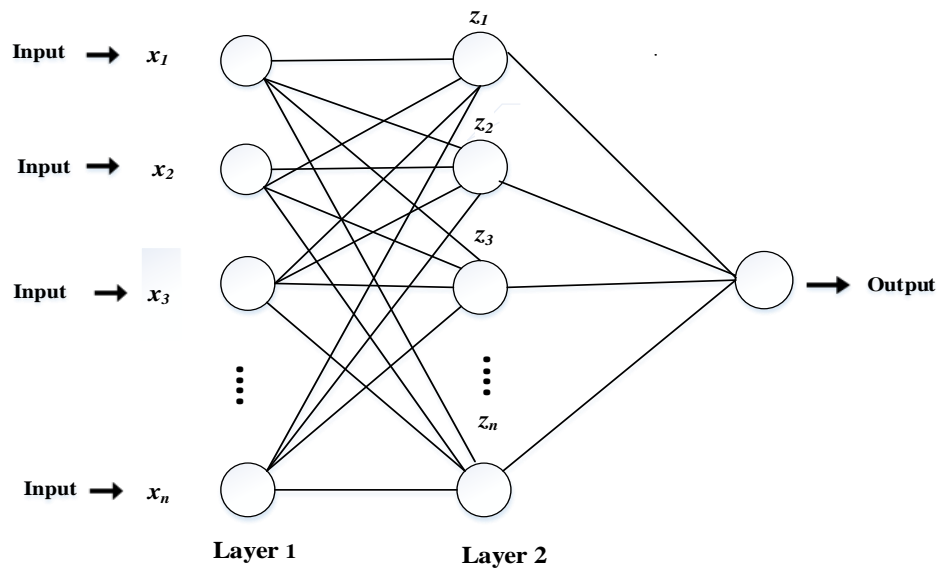


Figure 4. Illustration of CNN algorithm

4. EXPERIMENTAL RESULTS AND DISCUSSION

The experimental evaluation of the proposed AWKNN method for performance evaluation is shown in this section. The results demonstrate the efficiency of the proposed AWKNN. This section consists of three sub-sections, such as dataset, comparative analysis, and research summary.

4.1. Dataset

In this research, we have considered the dataset, namely, the Kaggle water quality dataset for the imputation of missing values in water quality monitoring. Data on water contaminants, including pH, hardness, solids, chloramine, sulphate, and conductivity, were gathered via environmental sensors. Recently,

data were collected for predictions in the monitoring of water pollution. However, these predictions turned out to be inaccurate, representing missing data. The most common processing techniques are imputation and deletion. The most popular strategy is probably discarding incomplete data given its straightforward nature.

Tossing out missing data, however, might result in information loss and decreased computing efficiency. Data-driven models might produce biased and inaccurate conclusions owing to the smaller sample size. Imputation approaches preserve the entire sample size by replacing missing data with projected acceptable values derived from the existing data, as opposed to deleting missing data. Table 3 describes the system specifications.

Table 3. System specifications

Hardware specifications		Software specifications	
Hard disk	500 GB	Tool	Python 3.11.3
RAM	4 GB	OS	Windows 10-(64-bits)

4.2. Comparative analysis

Comparative evaluation measures the performance of the imputation scheme suggested in this work in comparison with prevailing methods like KNN and KNN regressor (KNNR) in imputing missing water quality. The performance is measured in terms of various parameters like pH concentration, chloramine concentration, sulfate concentration, water level, and general accuracy for varying percentages of missing data. Experimental outcomes illustrate that the new method always outperforms the existing methods with improved accuracy and data integrity and lower computational complexity. The use of spectral clustering, AWKNN, and CNN improves the imputation process by excellent missing data classification and reconstruction. The outcome justifies that the new method presents a more dependable and scalable real-time water quality monitoring solution.

4.2.1. pH concentration versus records

The pH concentration is a measure of the acidity or basicity of a solution. It is defined as the negative logarithm (base 10) of the concentration of hydrogen ions (H^+) in the solution. The pH scale ranges from 0 to 14, where 7 is considered neutral. pH values below 7 indicate acidity, and pH values above 7 indicate basicity. Equation (18) is used to calculate pH.

$$pH = -\log[H^+], \tag{18}$$

where pH represents the pH value of the solution, \log denotes the logarithm function with a base of 10, and $[H^+]$ is the concentration of hydrogen ions in moles per liter (mol/L) in the solution.

Figure 5 depicts a comparison of the suggested approach's pH concentration with other existing approaches, such as KNNR and KNN. Therefore, the proposed approach has a higher pH concentration than other existing approaches. In the proposed methods, 500 records would achieve 1890 data, whereas 1000 achieves 1995 data. KNN has a pH concentration of 1100 with 500 records and 1330 data with 1000 records. In 500 records, KNNR achieves a pH concentration of 1870, and in 1000 records, the pH concentration involves 1970 data.

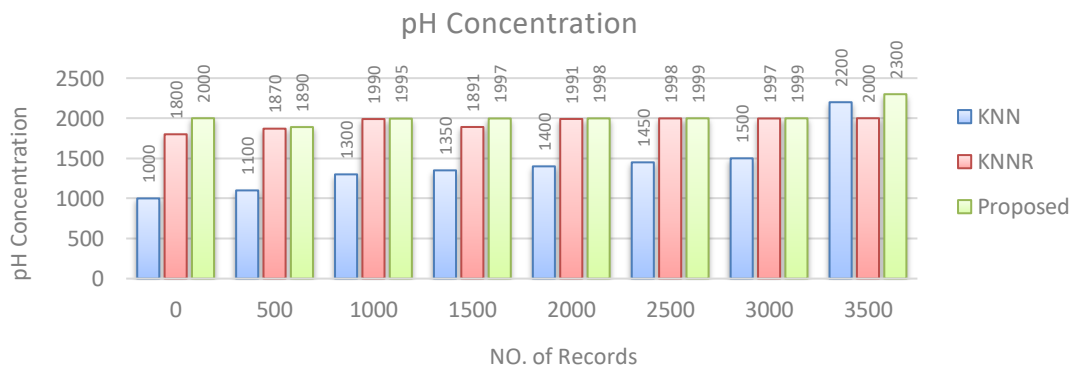


Figure 5. PH concentration versus number of records

4.2.2. Chloramine concentration against the number of records

Chloramine concentration in water is the measurement of the combined forms of chloramine (monochloramine, dichloramine, and trichloramine) used as disinfectants. It can be represented by the following equilibrium (19):



Figure 6 depicts a comparative examination of the chloramine concentration. When calculating efficient chloramine concentration, the most efficient method must provide the highest detection rate. The figure shows the chloramine concentration with increasing records. However, the proposed approach has a higher detection rate than current approaches, such as KNN and KNNR. The chloramine concentration in the proposed approaches involves 830 data in 400 records, and the 600 records include 880 data. KNN achieves a chloramine concentration of 650 data in 400 and 700 data in 600. KNNR achieves a chloramine concentration of 770 data in 400 records and 830 data in 600 records.

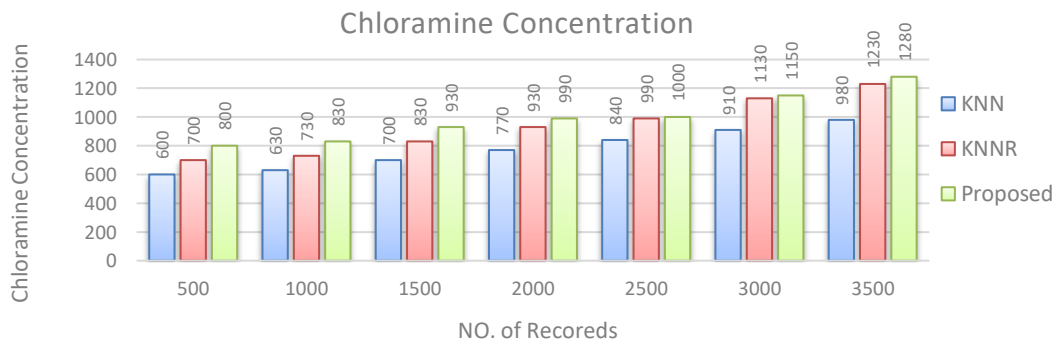


Figure 6. Chloramine concentration versus number of records

4.2.3. Sulphate concentration versus number of records

The sulphate concentration in a solution refers to the number of sulphate ions (SO4²⁻) present in that solution. The equation (20) for the sulphate concentration can be written as (20):

$$C_{sulphate} = \frac{\eta_{sulphate}}{v},
 \tag{20}$$

where $C_{sulphate}$ is the sulphate concentration in mol/L (M), $\eta_{sulphate}$ is the number of moles of sulfate ions in the solution, and v is the volume of the solution in liters.

Figure 7 depicts the sulphate concentration versus the number of records. A total of 940 data in 800 records are considered in the proposed technique. KNN reaches 750 data points within 800 records, whereas KNNR achieves 840 data points.

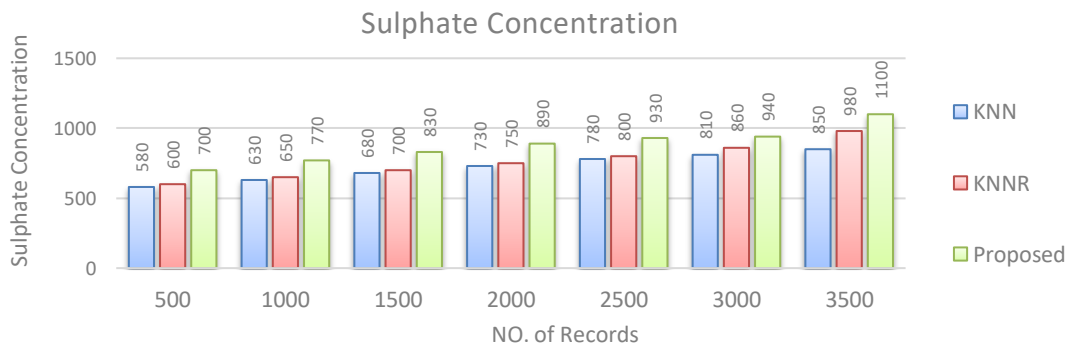


Figure 7. Sulphate concentration versus the number of records

4.2.4. Water level versus number of records

Figure 8 depicts the water level versus the number of records. The proposed technique considered 2700 data in 1000 records and 2600 data in 500 records. KNN includes 1000 records with 1600 data. In addition, KNNR achieves 2100 data points in 500 records and 2300 data points in 1000 records.

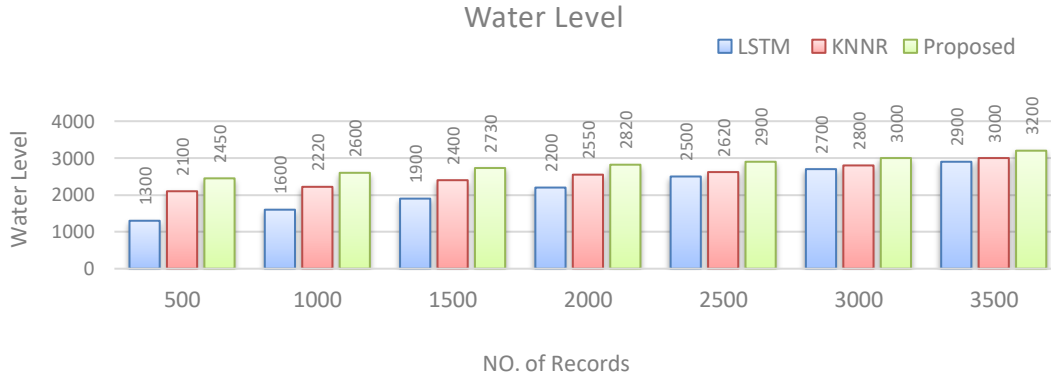


Figure 8. Water level versus number of records

4.2.5. Accuracy versus missing ratio

Accuracy is a measure of how well a model or system performs in accurately predicting or classifying data points. It is commonly used in classification tasks and is defined as the ratio of accurately predicted or classified data points to the total number of data points. Equation (21) is used to calculate accuracy (*ACC*), as follows:

$$Accuracy (ACC) = \frac{TP+TN}{TP+TN+FP+FN} \tag{21}$$

where true positive (*TP*) indicates the number of positive instances accurately predicted as positive by the model; true negative (*TN*) represents the number of negative instances accurately predicted as negative by the model; false positive (*FP*) suggests the number of negative instances inaccurately predicted as positive by the model; false negative (*FN*) indicates the number of positive instances inaccurately predicted as negative by the model.

Figure 9 clearly demonstrates a comparison of the suggested approach’s accuracy versus the missing ratio with other existing approaches, such as KNNR and KNN. The graphical representation in this figure highlights how the accuracy of the proposed method consistently exceeds that of its counterparts as the missing ratio increases, providing robust evidence of its effectiveness under varying conditions. Consequently, these results indicate that the proposed approach offers a significantly higher accuracy for handling missing ratios, affirming its potential advantage in practical applications.

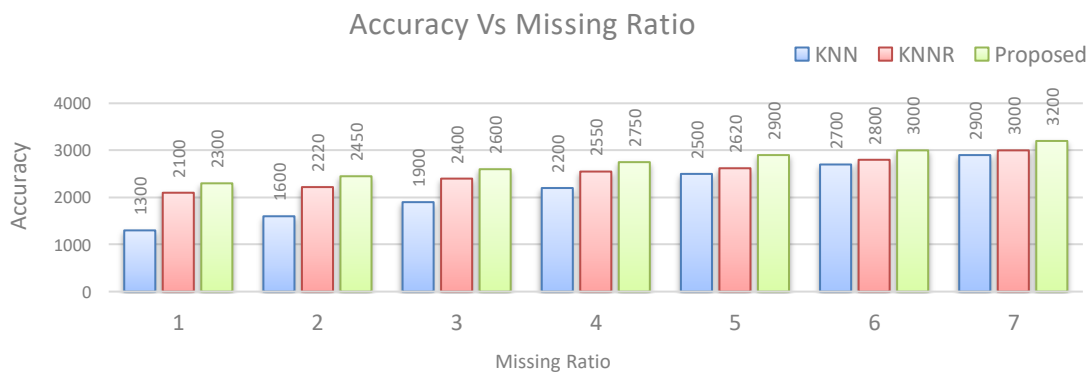


Figure 9. Accuracy versus missing ratio

4.3. Discussion

The experimental results illustrate the viability of the presented proposed AWKNN and CNN-based spectral clustering-based gene set imputation technique for imputing missing data into water quality monitoring. Based on the research contrast study, it is evident that the presented approach outperformed the performance criteria of the number of standard approaches such as KNN and KNNR. For instance, the results by the AWKNN indicated a marked improvement of accuracy and data consistency wherein a pH concentration of 2300 was achieved contrary to that achieved by KNNR as 2200 while that for KNN was as low as 2000. The concentration of chloramine as achieved by using the proposed method is 1230. However, concerning KNNR and KNN, they presented 1200 and 1000 concentrations, respectively. These results essentially reflect how well the AWKNN algorithm clusters and imputes appropriately, especially in noisy and imperfect datasets. This technique based on CNN uses deep feature extraction that learns from the patterns in a dataset to restore missing values efficiently. This combination of CNN and AWKNN improves accuracy without raising total data processing complexity compared to KNNR, as the present case demonstrates with a detected water level of 3200 as opposed to the case of KNNR, which was 2900. For such large-scale, real-time monitoring applications for water quality, it is suitable because it can retain accuracy even when missing ratios increase. The present work gives a dependable and expandable solution for environmental data management by overcoming common noise, missing data, and overfitting problems found in current methods. Thus, the results point out a basic need to make use of deep learning and advanced techniques in clustering for the improvement of data quality through imputation to be helpful in proper and accurate management decisions.

This research has some limits even with the promising results: It was tested only on one available dataset the Kaggle water quality data-which might not generalize well beyond domains or datasets that are different from the one here studied. Additionally, though spectral clustering coupled with CNN works well when medium-sized datasets are concerned, quite huge datasets or real-time varied datasets might be associated with challenges such as scalability because it gets too computationally costly in practice. In addition, other potential environmental indicators that can benefit from analogous imputation methods were not explored to focus on specific water quality measures, including pH, chloramine, and sulfate concentrations.

To test the flexibility and robustness, the proposed approach will be tested on several datasets of other industrial and environmental domains in the future. In addition, the possibility of optimizing the algorithm for distributed and cloud-based settings to further increase scalability and support real-time imputation for large monitoring systems will also be explored. Including advanced feature selection techniques and noise reduction methods may be more effective in enhancing the accuracy and computational efficiency of the imputation process. Moreover, including multi-source data fusion techniques from remote sensing and internet of things devices may result in a more comprehensive and detailed assessment of environmental quality that may lead to better water resource management and decision-making.

4.4. Research summary

We propose a unique imputation technique involving spectral clustering based on a gene set using AWKNN and missing data imputation using the CNN algorithm, ensuring accurate imputed data. The utilization of data cleaning techniques, such as MMWFILT, detects inaccurate data, ensuring the reliability of the imputed dataset. The use of normalization techniques based on the Z-SN approach aids in improved data organization and management, enhancing the accuracy of the imputation process. Data reduction using IKCF eliminates unwanted data, optimizing storage capacity, and improving the overall efficiency of the analysis. Column profiling using the EPCA approach aids in analyzing the patterns and characteristics of specific columns, reducing overfitting issues. The classification of the dataset into complete and missing data using the LIGHT DN approach allows a comprehensive evaluation of the imputation method. The suggested approach's performance is discussed in this subsection. The findings of the comparison study are shown graphically in Figures 5 to 9, and Table 4 provides the numerical results of the comparative analysis.

Table 4. Numerical outcomes

Performance metrics	KNN	KNNR	LSTM	Proposed
PH concentration	2,000	2,200	-	2,300
Chloramine concentration	1,000	1,200	-	1,230
Sulphate concentration	850	980	-	1,000
Water level	-	2,900	3000	3,200
Accuracy	2900	3,000	-	3,150

5. CONCLUSIONS AND FUTURE WORK

In this study, the proposed method displays promising results in imputing missing data in water quality monitoring datasets. The combination of spectral clustering, AWKNN, and CNN algorithms accurately imputes missing values, leading to improved data analysis and decision-making processes. Comparative analysis against traditional techniques highlights the superiority of the proposed method in recovering missing data while preserving the underlying data distribution. The evaluation of the method using various water quality parameters, such as PH concentration, chloramine concentration, sulphate concentration, water level, and accuracy, further support the effectiveness of the proposed approach. We evaluate the performance of our approach through numerical analysis, demonstrating that our approach outperforms existing approaches across all metrics. This research provides valuable insights into the imputation of missing data in water quality monitoring and offers a reliable approach to enhance the accuracy and comprehensiveness of data analysis in this field.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Al Hussein Bin Talal University for providing the necessary infrastructure and support for conducting this research.

FUNDING INFORMATION

This study received no external funding.

AUTHOR CONTRIBUTIONS STATEMENT

This paper uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Amer Al-Rahayfeh	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓
Saleh Atiewi	✓	✓	✓			✓			✓	✓		✓	✓	✓
Muder Almiani										✓	✓			✓
Ala Mughaid				✓	✓			✓		✓				
Abdul Razaque					✓		✓			✓		✓		
Bilal Abu-Salih		✓					✓			✓	✓			
Mohammed AlWeshah				✓		✓				✓				
Alaa Alrawajfeh			✓		✓					✓				

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nterpretation

R : **R**esources

D : **D**ata Curation

O : **O**riginal Draft

E : **E**diting

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, SA, upon reasonable request.




REFERENCES

- [1] F. Jan, N. Min-Allah, and D. Düştögör, "IoT based smart water quality monitoring: Recent techniques, trends and challenges for domestic applications," *Water*, vol. 13, no. 13, p. 1729, Jun. 2021, doi: 10.3390/w13131729.
- [2] N. A. Razman, W. Z. Wan Ismail, M. H. Abd Razak, I. Ismail, and J. Jamaludin, "Design and analysis of water quality monitoring and filtration system for different types of water in Malaysia," *International Journal of Environmental Science and Technology*, vol. 20, no. 4, pp. 3789–3800, Apr. 2023, doi: 10.1007/s13762-022-04192-x.
- [3] I. Yaroshenko *et al.*, "Real-time water quality monitoring with chemical sensors," *Sensors*, vol. 20, no. 12, p. 3432, Jun. 2020.




- doi: 10.3390/s20123432.
- [4] U. Ahmed, R. Mumtaz, H. Anwar, S. Mumtaz, and A. M. Qamar, "Water quality monitoring: from conventional to emerging technologies," *Water Supply*, vol. 20, no. 1, pp. 28–45, Feb. 2020, doi: 10.2166/ws.2019.144.
- [5] S. Pasika and S. T. Gandla, "Smart water quality monitoring system with cost-effective using IoT," *Heliyon*, vol. 6, no. 7, p. e04096, Jul. 2020, doi: 10.1016/j.heliyon.2020.e04096.
- [6] M. S. U. Chowdury *et al.*, "IoT based real-time river water quality monitoring system," *Procedia Computer Science*, vol. 155, pp. 161–168, 2019, doi: 10.1016/j.procs.2019.08.025.
- [7] R. Martínez, N. Vela, A. el Aatik, E. Murray, P. Roche, and J. M. Navarro, "On the use of an IoT integrated system for water quality monitoring and management in wastewater treatment plants," *Water*, vol. 12, no. 4, p. 1096, Apr. 2020, doi: 10.3390/w12041096.
- [8] J. B. Ajith, R. Manimegalai, and V. Ilayaraja, "An IoT based smart water quality monitoring system using cloud," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, Feb. 2020, pp. 1–7, doi: 10.1109/ic-ETITE47903.2020.450.
- [9] F. Akhter, H. R. Siddiquei, M. E. E. Alahi, K. P. Jayasundera, and S. C. Mukhopadhyay, "An IoT-enabled portable water quality monitoring system with MWCNT/PDMS multifunctional sensor for agricultural applications," *IEEE Internet of Things Journal*, vol. 9, no. 16, pp. 14307–14316, Aug. 2022, doi: 10.1109/JIOT.2021.3069894.
- [10] B. Arabi, M. S. Salama, J. Pitarch, and W. Verhoef, "Integration of in-situ and multi-sensor satellite observations for long-term water quality monitoring in coastal areas," *Remote Sensing of Environment*, vol. 239, p. 111632, Mar. 2020, doi: 10.1016/j.rse.2020.111632.
- [11] A. Juna *et al.*, "Water quality prediction using KNN imputer and multilayer perceptron," *Water*, vol. 14, no. 17, p. 2592, Aug. 2022, doi: 10.3390/w14172592.
- [12] H. Liu *et al.*, "Uav-borne hyperspectral imaging remote sensing system based on acousto-optic tunable filter for water quality monitoring," *Remote Sensing*, vol. 13, no. 20, p. 4069, Oct. 2021, doi: 10.3390/rs13204069.
- [13] N. R. Ekere, V. E. Agbazue, B. U. Ngang, and J. N. Ihedioha, "Hydrochemistry and water quality index of groundwater resources in Enugu north district, Enugu, Nigeria," *Environmental Monitoring and Assessment*, vol. 191, no. 3, p. 150, Mar. 2019, doi: 10.1007/s10661-019-7271-0.
- [14] T. S. Kapalanga, Z. Hoko, W. Gumindoga, and L. Chikwiramakomo, "Remote-sensing-based algorithms for water quality monitoring in Olushandja Dam, north-central Namibia," *Water Supply*, vol. 21, no. 5, pp. 1878–1894, Aug. 2021, doi: 10.2166/ws.2020.290.
- [15] O. O. Famoofo and I. F. Adeniyi, "Impact of effluent discharge from a medium-scale fish farm on the water quality of Odo-Owa stream near Ijebu-Ode, Ogun State, Southwest Nigeria," *Applied Water Science*, vol. 10, no. 2, p. 68, Feb. 2020, doi: 10.1007/s13201-020-1148-9.
- [16] M.-J. Kim, S. S. Choi, P. B. S. Rallapalli, J. H. Ha, S.-M. Lee, and Y.-S. Lee, "Nitrate removal from water phase using Robinia pseudoacacia bark for solving eutrophication," *Korean Journal of Chemical Engineering*, vol. 36, no. 9, pp. 1450–1454, Sep. 2019, doi: 10.1007/s11814-019-0331-x.
- [17] D. T. Bui, K. Khosravi, J. Tiefenbacher, H. Nguyen, and N. Kazakis, "Improving prediction of water quality indices using novel hybrid machine-learning algorithms," *Science of The Total Environment*, vol. 721, p. 137612, Jun. 2020, doi: 10.1016/j.scitotenv.2020.137612.
- [18] G. Gao, K. Xiao, and M. Chen, "An intelligent IoT-based control and traceability system to forecast and maintain water quality in freshwater fish farms," *Computers and Electronics in Agriculture*, vol. 166, p. 105013, Nov. 2019, doi: 10.1016/j.compag.2019.105013.
- [19] S. Kirschke *et al.*, "Capacity challenges in water quality monitoring: understanding the role of human development," *Environmental Monitoring and Assessment*, vol. 192, no. 5, p. 298, May 2020, doi: 10.1007/s10661-020-8224-3.
- [20] R. Rodríguez *et al.*, "Water-quality data imputation with a high percentage of missing values: A machine learning approach," *Sustainability*, vol. 13, no. 11, p. 6318, Jun. 2021, doi: 10.3390/su13116318.
- [21] Y. Zhang and P. J. Thorburn, "A dual-head attention model for time series data imputation," *Computers and Electronics in Agriculture*, vol. 189, p. 106377, Oct. 2021, doi: 10.1016/j.compag.2021.106377.
- [22] H. Han, M. Sun, H. Han, X. Wu, and J. Qiao, "Univariate imputation method for recovering missing data in wastewater treatment process," *Chinese Journal of Chemical Engineering*, vol. 53, pp. 201–210, Jan. 2023, doi: 10.1016/j.cjche.2022.01.033.
- [23] T. Khampungson and W. Wang, "Novel methods for imputing missing values in water level monitoring data," *Water Resources Management*, vol. 37, no. 2, pp. 851–878, Jan. 2023, doi: 10.1007/s11269-022-03408-6.
- [24] F. B. Hamzah, F. Mohd Hamzah, S. F. Mohd Razali, and H. Samad, "A comparison of multiple imputation methods for recovering missing data in hydrological studies," *Civil Engineering Journal*, vol. 7, no. 9, pp. 1608–1619, Sep. 2021, doi: 10.28991/cej-2021-03091747.
- [25] X. Xu, T. Lai, S. Jahan, F. Farid, and A. Bello, "A machine learning predictive model to detect water quality and pollution," *Future Internet*, vol. 14, no. 11, p. 324, Nov. 2022, doi: 10.3390/fi14110324.
- [26] L. Kulanuwat *et al.*, "Anomaly detection using a sliding window technique and data imputation with machine learning for hydrological time series," *Water*, vol. 13, no. 13, p. 1862, Jul. 2021, doi: 10.3390/w13131862.
- [27] A. Palanivinnayagam and R. Damaševičius, "Effective handling of missing values in datasets for classification using machine learning methods," *Information*, vol. 14, no. 2, p. 92, Feb. 2023, doi: 10.3390/info14020092.
- [28] Y.-S. Sim, J.-S. Hwang, S.-D. Mun, T.-J. Kim, and S. J. Chang, "Missing data imputation algorithm for transmission systems based on multivariate imputation with principal component analysis," *IEEE Access*, vol. 10, pp. 83195–83203, 2022, doi: 10.1109/ACCESS.2022.3194545.
- [29] H.-R. Kim, H. Y. Soh, M.-T. Kwak, and S.-H. Han, "Machine learning and multiple imputation approach to predict Chlorophyll-a concentration in the coastal zone of Korea," *Water*, vol. 14, no. 12, p. 1862, Jun. 2022, doi: 10.3390/w14121862.
- [30] J. A. Smith, J. H. Morgan, and J. Moody, "Network sampling coverage III: Imputation of missing network data under different network and missing data conditions," *Social Networks*, vol. 68, pp. 148–178, Jan. 2022, doi: 10.1016/j.socnet.2021.05.002.

BIOGRAPHIES OF AUTHORS






Amer Al-Rahayfeh    received his Ph.D. in computer science and engineer from the University of Bridgeport (U.S.A) in 2014. He is currently an associate professor of computer sciences at Al-Hussein Bin Talal University (AHU) in Jordan. His research interests are in the areas of multimedia systems, computer vision, sensor networks, cloud computing biomedical systems. At AHU, he led the Department of Computer Science and vice dean of College of Information Technology. He can be contacted at email: amer.a.al-rahayfeh@ahu.edu.jo.






Saleh Atiewi    received his B.Sc. degree in computer science from Al-Isra University, Amman, Jordan, in 1999, followed by a Master's degree in internet technology from Wollongong University, Wollongong, Australia, in 2004. He later earned his Ph.D. in computer science from Tenaga Nasional University, Putrajaya, Malaysia, in 2017. Since 2004, Dr. Atiewi has been part of Al Hussein Bin Talal University in Ma'an, Jordan. He is currently an associate professor in the Department of Computer Science, actively contributing to the university's academic and research initiatives. Over the years, Dr. Atiewi has held several key administrative and leadership roles, including: Head of the Computer Science Department, Vice-Dean of Scientific Research and Postgraduate Studies, Director of the Computer Center and Information Technology, and Director of the Center for Innovation, Creativity, and Entrepreneurship. Dr. Atiewi's research interests include network security, cloud computing, security, and the internet of things (IoT). He remains deeply committed to advancing knowledge and fostering innovation within his field. He can be contacted at email: saleh@ahu.edu.jo.






Mudher Almiani    is an associate professor of Management Information System (MIS). Dr. Almiani joined Gulf University for Science and Technology in September 2020. Dr. Almiani obtained his Ph.D from University of Bridgeport in computer science and engineering, USA. Dr Almiani has been actively involved with the Institute of Electrical & Electronic Engineers (IEEE). Before joining GUST Dr. Almiani was the chair of Computer Information Systems Department (2014-2018) and the chair of Management Information System Department (2013-2017) in Al Hussein bin Talal University, Jordan, and he was the Chair of Computer/Computational Intelligence Chapter IEEE, Jordan Section (2017-2020). He has been on the program committee of many national and international conferences, organized and chaired special sessions. He can be contacted at email: Almiani.m@gust.edu.kw.






Ala Mughaid    was born in Irbid, Jordan, in 1984. He received the BSC degree in computer science from Jordan University of Science and Technology (JUST), Jordan, in 2006, and the MSC in engineering degree in engineering computer network from the Western Sydney University, Sydney, Australia, in 2010. Dr. Mughaid received the Ph.D. degree in computer science from Newcastle University – Sydney, Australia, in 2018. In 2018, Dr. Mughaid joined the Department of Computer Science, The Hashemite University, as an assistant professor, Zarqa, Jordan. Dr. Mughaid has join the Computer Science Department at GUST University, Kuwait since 2024. Dr. Mughaid current research interests include but not limited to cyber security, cloud computing, artificial intelligence, virtual reality, data mining. He is working voluntarily in many social services. He can be contacted at email: ala.mughaid@hu.edu.jo.






Abdul Razaque    is a researcher affiliated with the University of Bridgeport, specializing in Blockchain technology, cybersecurity, internet of things (IoT), cloud computing, and wireless sensor networks (WSNs). His scholarly contributions have garnered over 4,400 citations, reflecting his significant impact in these fields. Among his notable works is the 2020 publication titled "Deep recurrent neural network for IoT intrusion detection system," which has been cited 421 times. Another significant contribution is the 2013 paper "Compression in wireless sensor networks: a survey and comparative evaluation," with 240 citations. Dr. Razaque's research has been instrumental in advancing understanding and innovation in areas critical to modern technology infrastructure. His extensive citation record underscores the influence and relevance of his work within the academic and professional communities. He can be contacted at email: a.razaque@edu.iitu.kz.






Bilal Abu-Salih    is an associate professor at the University of Jordan and an Adjunct at Curtin University. He holds a Ph.D. in information systems (with a focus on social big data analytics) from Curtin University. Bilal's research interests include; data science, social big data, semantic analytics, NLP, and the like. Bilal's background in both industry and software development makes him a versatile asset. His hands-on coding proficiency includes Python, R, Java, C, C#.net, PHP, and SQL. He brings experience in data analytics, machine learning, social media data mining, and big data analysis to projects spanning academic research, software development, and industrial implementation. He can be contacted at email: b.AbuSalih@ju.edu.jo.



Mohammed AlWeshah    President of Aqaba University of Technology, Aqaba, Jordan. He attained a position within the top 2% of scientists and researchers globally, as determined by the esteemed American Stanford University classification, which relies on data from the Scopus database and collaboration with the prominent international publishing house 'Elsevier'. He held the position of Dean of Scientific Research and Postgraduate Studies at Aqaba University of Technology. He served as a professor and researcher within the Technology Department at the Faculty of Information Technology, Aqaba University of Technology. He holds the title of Professor in computer science, specializing in artificial intelligence and data science, at the Prince Abdullah bin Ghazi College of Communications and Information Technology, Al-Balqa Applied University, located in Salt. He obtained his doctoral degree in artificial intelligence and data science from the National University of Malaysia (UKM) in December 2013, supported by a scholarship awarded by the Malaysian government (MTCP scholarship) for his academic pursuits. He is a dedicated artificial intelligence researcher with a keen focus on issues pertaining to data mining, machine learning, scheduling, and geometric optimization problems. Furthermore, his research endeavors extend to the theoretical underpinnings of artificial intelligence methodologies, including evolutionary computation algorithms and their practical applications. Additionally, he engages in interdisciplinary studies encompassing bioinformatics, engineering, information security, and software engineering. He has authored over 80 recent scientific papers, which are indexed in Scopus and the Web of Science, focusing on addressing various artificial intelligence challenges including feature selection, classification, sentiment analysis, and others. The majority of these endeavors involve tackling combinatorial optimization problems, which hold significant importance within their respective domains. All of his research findings are disseminated through prestigious international journals, consistently ranked at the forefront of their field. He can be contacted at email: weshah@bau.edu.jo.



Alaa Alrawajfeh    is a dedicated academic and faculty member at Al-Balqa' Applied University in Jordan. She completed her Bachelor's degree in 2004 from Al-Hussein Bin Talal University and earned her Master's degree in 2010 from Mutah University. Currently, she is pursuing her Ph.D. at Universiti Sains Malaysia (USM), focusing on advanced research within her field of expertise. Throughout her academic career, Alaa has demonstrated a strong commitment to education and research, contributing to her institution's academic excellence. Her role at Al-Balqa' Applied University underscores her dedication to fostering knowledge and innovation in her students and peers. She can be contacted at email: alaa.rawajfeh@bau.edu.jo.