Bulletin of Electrical Engineering and Informatics
Vol. 14, No. 5, October 2025, pp. 3704~3715
ISSN: 2302-9285, DOI: 10.11591/eei.v14i5.10063

Enhancing data integrity in internet of things-based healthcare applications: a visualization approach for duplicate detection

Siti Noor Basirah Md Isa[1], Nurul A. Emran[1], Norharyati Harum[1], Logenthiran Machap[2], Azlin Nordin[3]
[1]Department of Software Engineering, Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia
[2]Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, Panchor, Malaysia
[3]Department of Computer Science, Kulliyyah Information and Communication Technology, International Islamic University Malaysia, Kuala Lumpur, Malaysia

Article Info

Article history:

Received Feb 13, 2025
Revised Aug 8, 2025
Accepted Sep 1, 2025

Keywords:

Data duplication
Duplicates detection
Healthcare
Internet of things data
Visualization

ABSTRACT

This study addresses the critical issue of data duplication in healthcare-related internet of things (IoT) datasets, which can compromise the reliability of analyses and patient outcomes. A Python-based visualization framework using Pandas and Matplotlib was developed to detect and represent duplicate records. The methodology was applied to six cancer-related datasets sourced from Kaggle, ranging from 300 to 55,000 records, encompassing numerical, textual, and categorical data types. The visualization technique provided clear insights into duplication patterns, identifying specific counts such as 7 duplicates in the wearable device dataset, 19 in the thyroid recurrence dataset, and 534 in the synthetic healthcare electronic health record (EHR) dataset. Compared to traditional detection methods, the visualization tool facilitated faster and more intuitive initial data assessment, demonstrating its effectiveness for rapid quality checks in healthcare datasets. However, scalability limitations were observed in larger datasets, where visual clarity declined. These findings highlight the value of visualization as a preliminary data quality assessment tool and suggest future integration with advanced detection algorithms to enhance robustness and scalability.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Nurul A. Emran
Department of Software Engineering, Fakulti Teknologi Maklumat dan Komunikasi
Universiti Teknikal Malaysia Melaka
Ayer Keroh, 76100 Melaka, Malaysia
Email: nurulakmar@utem.edu.my

1. INTRODUCTION

The internet of things (IoT) has transformed how data is generated, shared, and utilized across various domains, including healthcare, manufacturing, and smart cities. By enabling autonomous communication among interconnected devices, IoT systems produce vast amounts of data that drive automation and intelligent decision-making. However, this data deluge presents significant challenges, particularly in maintaining data quality. Among these challenges, data duplication stands out as a critical issue that can compromise the reliability of analyses, especially in domains where accuracy is paramount, such as healthcare.

Several studies have highlighted the unique characteristics of IoT data, including its heterogeneity, high velocity, and dynamic nature [1], [2]. Inherent features align IoT data with big data aspects, such as high

volume, velocity, and heterogeneity. One important feature of IoT data is its heterogeneity, which contains various data types and sources within complex networks. Singh and Mahapatra [1] note that IoT devices generate structured, semi-structured, and unstructured data, collectively known as big data. The data formats are also heterogeneous, which can include numerical, text, extensible markup language (XML), and multimedia data [2].

The open-source datasets form the foundation of IoT research and development. These freely accessible data collections enable researchers to analyze and innovate without barriers. Shahid *et al*. [3] emphasize the importance of these datasets in providing realistic traffic patterns, network performances, and demand scenarios for both fixed and dynamic user terminals. These features make IoT data prone to quality issues such as missing values, anomalies, and duplication [4], [5]. Open-source datasets, which are widely used for research and development, are especially vulnerable to duplication due to inconsistent data collection and integration practices [6]–[9]. Despite the growing awareness of these issues, systematic comparative studies on data duplication in IoT datasets remain limited, particularly in the healthcare sector. Previous works have explored various methods for duplicate detection, including hash-based techniques [10]–[12], content-aware approaches [10], [13], and hybrid models [8], [14]. While these methods offer varying degrees of accuracy and efficiency, they often require significant computational resources and may not be suitable for preliminary data quality assessments. Moreover, the implications of duplication in healthcare data are profound, affecting patient safety, operational efficiency, and equity in care delivery [7], [15], [16], even though other issue like enabling unified access to verified patient data also has been highlighted [17].

While numerous studies have focused on algorithmic approaches to duplicate detection, such as hash-based, content-aware, and hybrid methods there is a noticeable lack of emphasis on visualization techniques that support interpretability and rapid assessment. Several researchers have advocated for the use of data quality visualization as part of the data preprocessing phase (see [18]–[20]), reinforcing the relevance of integrating visualization into early-stage data quality workflows. This gap highlights the need for lightweight, visual tools that can complement existing methods, particularly in exploratory and resource-constrained environments.

Thus, this study addresses the gap by proposing a visualization-based approach to detect and represent data duplication in IoT-based healthcare datasets. Unlike traditional detection algorithms, the proposed method leverages Python-based tools which are Pandas and Matplotlib to provide an intuitive graphical overview of duplicate records. The approach is tested on six cancer-related datasets sourced from Kaggle, ranging from 300 to 55,000 records, encompassing numerical, textual, and categorical data types.

The remainder of this paper is organized as follows: section 2 reviews related work on duplicate detection methods. Section 3 describes the methodology, including dataset selection, visualization techniques, and validation procedures. Section 4 presents the results and discusses the effectiveness and limitations of the proposed approach. Finally, section 5 concludes the study and outlines directions for future research.

2. RELATED WORK

Data duplication stands out as a significant concern in IoT datasets. In IoT datasets, several types of data duplication, namely exact, similar, spatial, temporal, and semantic, significantly affect data quality and analysis. Exact duplicates are identical records in all aspects, often arising from multiple sensors capturing the same conditions [21]. Similar duplicates involve records with minor variations in values or timestamps [22], while spatial duplicates occur when sensor nodes nearby capture the same data [23]. Temporal duplicates refer to identical data collected at different times [24]. Semantic duplicates occur when data is represented differently across datasets, such as through varied formats or naming conventions [8]. Duplicate records in healthcare data present significant challenges, impacting both patient safety and healthcare efficiency. These records arise when a single patient is associated with multiple medical record numbers, leading to fragmented patient information and potential medical errors [7]. Duplicate records can lead to patient safety issues, such as incorrect treatment due to incomplete or inaccurate patient information. This is particularly concerning in pediatric hospitals and organizations implementing new information systems [25].

2.1. Duplicate detection methods

There are at least three primary categories of duplicate detection methods: hash-based methods, content-aware methods, and hybrid/advanced approaches. Each category has its strengths and limitations, making them suitable for different scenarios within IoT data management. Table 1 summarizes these three main categories of duplicate detection methods by comparing their advantages, disadvantages, and suitable use cases. Hash-based methods are efficient for detecting exact duplicates and are computationally lightweight, making them ideal for storage optimization tasks. However, they struggle with near-duplicate detection. Content-aware methods, while more accurate for identifying similar or semantically related records, require significant processing power and are sensitive to data quality. Hybrid methods combine strengths from both categories, offering robust detection in complex scenarios but at the cost of

implementation complexity. This comparison helps readers understand which method is most appropriate depending on dataset characteristics and resource availability.

Table 1. Summary table for duplicate detection methods

Category	Advantages	Disadvantages	Suitable scenarios	Ref.
Hash-based	– Quick identification of identical data – Low computational cost – Efficient for exact duplicate detection	– Ineffective for detecting near-duplicates – Vulnerable to hash collisions with weaker hash functions	– File deduplication in backup systems – Storage optimization	[10], [11], [26]
Content-aware	– Effective for detecting near-duplicates and semantically similar data – Comprehensive similarity calculations (e.g., MSRD algorithm)	– Computationally intensive – Requires significant processing power and resources – Accuracy is influenced by data quality, noise, and incomplete information	– Plagiarism detection – Copyright infringement detection – Data integration tasks	[10], [27]-[29]
Hybrid	– Addresses specific challenges (e.g., frequently modified data, incomplete data) – Robust, accurate detection	– Complex to implement – Requires additional computational resources and specialized knowledge	– Scenarios with specific requirements or challenges – Applications needing robust and accurate duplicate detection	[9], [13], [28]

2.2. Comparative analysis of methods for data duplicate detection

In this section, a comparative analysis is presented. The analysis examines further key approaches to duplicate detection that cover data preparation and preprocessing, evaluation metrics, and performance assessment. The process flow of data duplicate detection plays a crucial role in ensuring data quality and consistency across systems. Various methodologies follow distinct workflows, incorporating processes such as data preprocessing, clustering, and similarity matching. This comparative analysis also explores the process flows used in duplicate detection.

2.2.1. Data preparation and preprocessing

Data preparation is a critical step before applying duplicate detection, as it ensures the accuracy and efficiency of the detection process. Proper data preparation involves cleaning and transforming the data to eliminate errors and inconsistencies that could hinder the detection of duplicates. This process is essential for integrating data from multiple sources, where the risk of duplicate records is higher due to variations in data representation. Effective data preparation is crucial before applying duplicate detection algorithms. Ali *et al*. [29], Long *et al*. [30] highlight this importance. Data preprocessing includes data cleaning that involves removing whitespace, normalizing addresses, and ensuring consistent data formats [31]. This step is crucial to reduce the search space for duplicate detection and improve the accuracy of the process.

2.2.2. Blocking and clustering

Clustering is a crucial step before applying duplicate detection as it significantly enhances the efficiency and accuracy of the process. By organizing data into clusters, the number of comparisons needed to identify duplicates is reduced, which is particularly beneficial when dealing with large datasets. Clustering helps in grouping similar records, thereby minimizing unnecessary comparisons and focusing computational resources on the most promising candidate pairs. This approach not only improves the speed of duplicate detection but also increases accuracy by ensuring that similar records are evaluated together. Huang and Chiang [32] propose a k-means clustering-based blocking technique to partition data, cutting down comparisons. Similarly, Chen *et al*. [28] apply locality sensitive hashing (LSH) for clustering, using selected keys for efficient candidate selection. Ali *et al*. [29] use attribute ranking, the Hot-Deck method for missing values, and a sort-neighborhood algorithm to optimize detection.

2.2.3. Similarity matching and feature extraction

Similarity matching and feature extraction are essential in duplicate detection, as they enable accurate identification and differentiation of similar entities, such as images or text, within large datasets. Long *et al*. [30] use a two-step similarity calculation approach, considering numerical, literal, and semantic aspects. Huang and Chiang [32] apply edit distance and q-grams with a weighted scheme for rare terms, while Chen *et al*. [28] utilize feature extraction with predicates for improved similarity measures. Advanced techniques like metric functional dependencies (MFDs) by Huang and Chiang [32] allow for tolerance in

attribute values, enhancing flexibility. Xia *et al.* [9] introduce deduplication and resemblance detection engine (DARE), combining deduplication with resemblance detection, using modules to identify candidate chunks for compression and feature detection for greater accuracy.

2.2.4. Evaluation metrics and performance assessment

Evaluation metrics and performance assessment are critical in duplicate detection as they provide a structured approach to measure the effectiveness and reliability of detection algorithms. These metrics help in understanding the strengths and weaknesses of algorithms, guiding improvements, and ensuring that the algorithms meet the desired performance standards. Studies in duplicate detection commonly use precision, recall, F1-measure, and accuracy to evaluate performance [29], [30]. These metrics standardize comparisons across different data scenarios.

In summary, duplicate detection is a multi-layered approach, that consists of multiple key steps that can be designed to cover numerical, literal, and semantic similarities. While existing studies have explored various algorithmic approaches to duplicate detection, ranging from hash-based and content-aware methods to hybrid models, these techniques often require substantial computational resources and lack interpretability in early-stage data quality assessments. Moreover, few works have systematically evaluated duplication patterns across diverse healthcare IoT datasets. Traditional detection algorithms, while effective, often require significant computational resources and are less accessible for preliminary data quality assessments. Visualization offers a rapid and intuitive alternative, especially valuable in resource-constrained environments and during early-stage data exploration. Figure 1 illustrates the generic flow of duplicate detection based on the analysis. The visualization step, previously absent in conventional workflows, is now explicitly highlighted in the proposed process flow to address a key gap in duplicate detection practices. This study uniquely contributes a lightweight, Python-based visual detection method for early-stage duplicate identification, validated against traditional techniques.

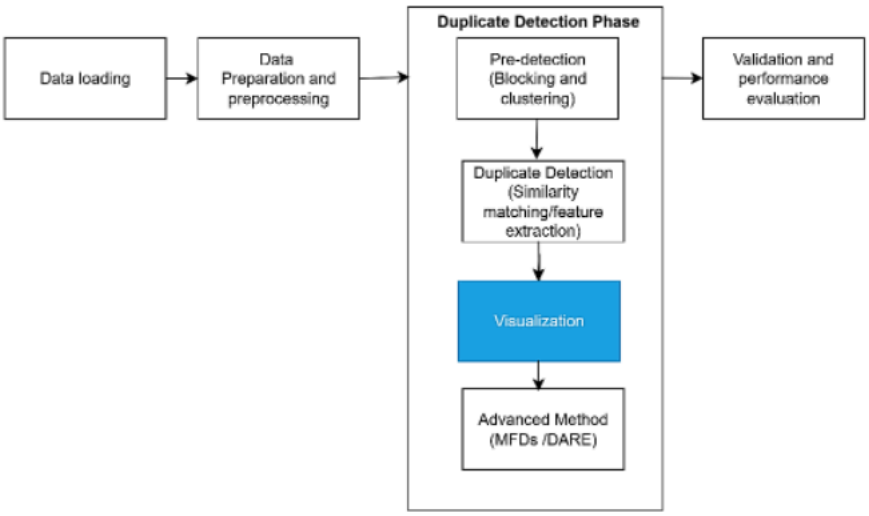

Figure 1. Process flow of duplicate detection with visualization

3. METHOD

This section outlines the detailed methodology used in this study. The choice of a visualization-based approach was driven by the need for a lightweight, interpretable, and scalable method to detect data duplication in healthcare IoT datasets. Python was selected due to its extensive ecosystem of data analysis libraries, and the initial codebase for data deduplication was adapted from a publicly available Kaggle notebook by Tatman [33]. Pandas provides robust data manipulation capabilities, while Matplotlib enables flexible and clear graphical representations. These tools are widely adopted in both academic and industry settings, ensuring reproducibility and accessibility. As highlighted in the introduction section, existing literature lacks systematic comparisons of duplication patterns across open-source healthcare IoT datasets. This study addresses that gap by applying a visualization framework across six diverse datasets, enabling comparative analysis of duplication characteristics. This section also outlines how the visualization technique is tested across datasets of varying sizes (300 to 55,000 records), directly responding to this concern and evaluating the method's performance under different data volumes. By combining visualization with

validation through baseline detection, the methodology ensures both interpretability and accuracy, aligning with the study's goal of enhancing data integrity in healthcare IoT applications.

3.1. Visualization of potential duplicates

The visualization method was implemented using the Python programming language. Two libraries were used namely Pandas (for data manipulation and preprocessing) and Matplotlib (for graphical representation). The following pseudocode outlines the steps used in the duplicate detection process:

Algorithm 1. Duplicate detection and visualization workflow

```
1: Load dataset D using Pandas
2: Preprocess D:
     for each column c in D do
         if c is irrelevant (e.g., primary key, metadata) then
             remove c
         end if
     end for
     for each categorical column c in D do
         convert c to numerical format (label encoding or one-hot encoding)
     end for
     for each missing value in D do
         if missingness is low and predictable then
             impute using mean or mode
         else
             remove record
         end if
     end for
     normalize and standardize data formats
3: Detect duplicates:
     apply duplicated() function to D
     store duplicate indices in list I
4: Visualize duplicates:
     if I is not empty then
         plot vertical lines at each index in I using Matplotlib
     end if
5: Validate visualization:
     compare plotted indices with terminal-based detection results
6: Evaluate effectiveness:
     assess clarity, interpretability, and alignment of visualization record evaluation
  metrics (e.g., detection  time, clarity score)
```

Visualization is not a detection method itself but a supportive tool that enhances interpretability and provides a quick overview of duplication patterns. Data quality visualization is commonly advocated in the data preprocessing phase (see [18]–[20]). Nevertheless, embedding it after detection (as illustrated in Figure 1) allows researchers to visually confirm the presence and distribution of duplicates, which is especially useful for exploratory analysis and quality assurance. It also helps in identifying potential anomalies or patterns that may not be obvious through terminal-based or algorithmic outputs. This graphical approach provides an intuitive overview of duplication patterns, enabling rapid assessment of data quality.

3.2. Validation of visualization results

To validate the effectiveness of the visualization, a baseline duplicate detection method was applied using Pandas. The `duplicated()` function was used to identify exact duplicates. The results from this method were compared with the visual output to ensure consistency. This validation step ensures that the visualization accurately reflects the underlying duplication in the dataset.

3.3. Experiment configuration

The experiment was configured to test the visualization method across six cancer-related healthcare datasets sourced mainly from Kaggle, which varied in size (ranging from 300 to 55,000 records) and data types.

3.3.1. Data selection and preparation

Datasets were selected based on their relevance to healthcare applications and their structural diversity, which included a mix of numerical, categorical, and textual data. This diversity was essential to evaluate the flexibility and robustness of the visualization method across different data types and formats. This variety was chosen to assess the visualization tool's flexibility in detecting duplicates across contexts, ensuring it effectively highlights duplicate issues in both large and small datasets and across different data

types. After selecting the datasets, a comprehensive data-cleaning process was undertaken to ensure data quality and relevance for the experiment. Irrelevant columns, such as automatically generated primary keys, were removed to streamline the datasets and maintain analytical focus. Categorical data are all converted to numerical type. This step was essential, as extraneous data could introduce noise and distort the results. Additionally, handling missing values was a critical aspect of the conclusions. By systematically addressing these issues, the datasets were refined to support precise and reliable outcomes in the evaluation of data duplication. Table 2 provides a summary of the six datasets used in this experiment along with their characteristics. The URLs of the datasets are available in the data availability section. Data were included if they contained relevant healthcare information and had sufficient structure for analysis. Records with excessive missing values or irrelevant metadata were excluded to maintain analytical focus. The decision to exclude certain columns or records was based on their impact on duplication detection accuracy and visualization clarity.

Table 2. The characteristics of datasets under study

Dataset	Size (number of records)	Data type
Wearable device	5,767	Numerical
Thyroid recurrence	387	Numerical, textual, and categorical
Patient healthcare HER (synthetic)	55,500	Numerical, textual, and categorical
Clinical glioma grading	839	Numerical and categorical
Online survey lung cancer	309	Numerical and categorical
Real breast cancer	334	Numerical, textual, and categorical

3.3.2. Visualization of duplicated rows and duplicate detection

As outlined in the pseudocode, the `duplicated()` function in Pandas was used to identify exact duplicates, and their indices were extracted for visualization. Using Matplotlib, vertical line plots were generated to mark the positions of these duplicates along the x-axis, representing row indices. The visual output was then cross-checked against terminal-based results to ensure consistency and accuracy.

3.3.3. Visualization clarity evaluation

To assess the interpretability and usability of the visualization tool, a visualization clarity score was introduced. This score reflects the subjective evaluation of each dataset's visual output based on clarity, readability, and ease of pattern recognition. Three independent data analysts with experience in healthcare data quality were asked to rate each visualization on a scale from 1 (poor clarity) to 5 (excellent clarity). Ratings were based on visual density and crowding, ability to distinguish duplicate clusters and overall readability of the plot. The average score across evaluators was recorded for each dataset. Notes were collected to contextualize the ratings and identify factors affecting clarity (e.g., dataset size, data type heterogeneity). This metric provides insight into the practical usability of the visualization tool across diverse dataset conditions.

3.4. Methodological assumptions

This method assumes that duplicate records are exact matches across all columns. As such, it is designed to detect only exact duplicates. Semantic duplicates such as records with similar but not identical values, formatting differences, or contextual equivalence are not captured by this approach. Additionally, fuzzy matches and temporal or spatial redundancies require more advanced techniques, such as clustering or AI-based similarity scoring, which are recommended for future integration.

4. RESULTS AND DISCUSSION

To evaluate the effectiveness of visualization in detecting duplicate records, a Python-based graphical approach was applied across six healthcare-related datasets. The visualization tool, built using Pandas and Matplotlib, successfully identified duplication patterns, offering immediate insights into data quality. However, the results also revealed varying degrees of effectiveness depending on dataset size and structure.

In the wearable device dataset (5,767 records), the visualization clearly highlighted 7 duplicates, as shown in Figure 2. This dataset, being uniformly numerical, allowed for straightforward detection, confirming the tool's suitability for homogeneous data types. The clarity of the visualization aligned with the terminal-based detection results, reinforcing its reliability. The thyroid recurrence dataset (387 records), which included numerical, textual, and categorical data, revealed 19 duplicates (see Figure 3). Despite the mixed data types, the visualization remained interpretable, demonstrating the tool's adaptability. This supports findings by Ali *et al.* [34], who emphasized the importance of preprocessing in enhancing duplicate detection accuracy.

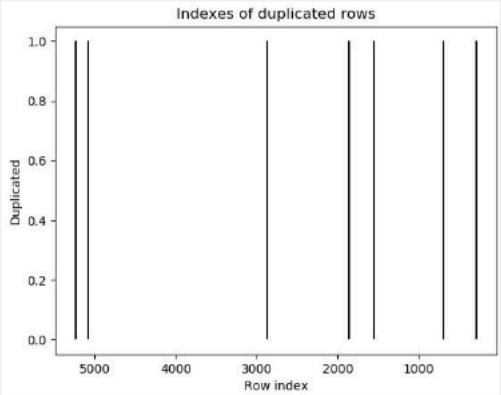

Figure 2. Duplicate data detection visualization graph for wearable device dataset

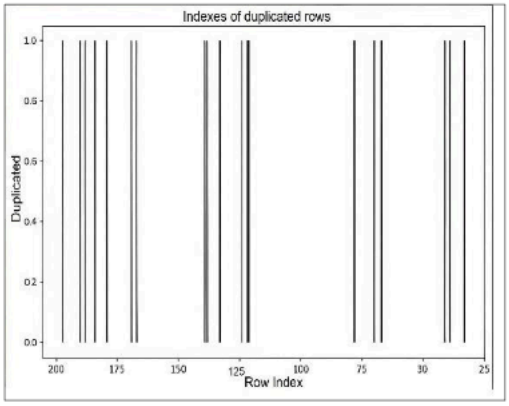

Figure 3. Duplicate data detection visualization graph for thyroid recurrence dataset

In contrast, the patient healthcare electronic health record (EHR) (synthetic) dataset (55,500 records) presented scalability challenges. Although 534 duplicates were detected (as shown in Figure 4), the visualization became overcrowded, reducing interpretability. This limitation echoes concerns raised by Huang and Chiang [35] regarding the need for clustering and blocking techniques to manage large-scale data. The results suggest that while visualization is effective for preliminary assessments, it should be complemented with advanced methods for high-volume datasets. The clinical glioma grading dataset (839 records) in Figure 5 contained only one duplicate, which was easily visualized. This case illustrates the tool's precision in low-duplication scenarios, offering a quick validation mechanism. Similarly, the online survey lung cancer dataset (309 records) in Figure 6 revealed 33 duplicates, with clear visual patterns that matched terminal outputs. Interestingly, the real breast cancer dataset (334 records) in Figure 7 showed no exact duplicates. This could indicate high data quality or the presence of non-exact (e.g., semantic or temporal) duplicates that the current visualization method could not detect. This limitation aligns with the observations by Chen *et al*. [28], who noted that semantic redundancy requires more sophisticated detection techniques.

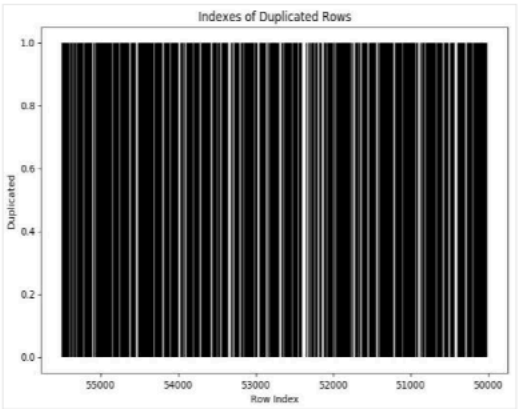

Figure 4. Duplicate data detection visualization graph for healthcare EHR dataset

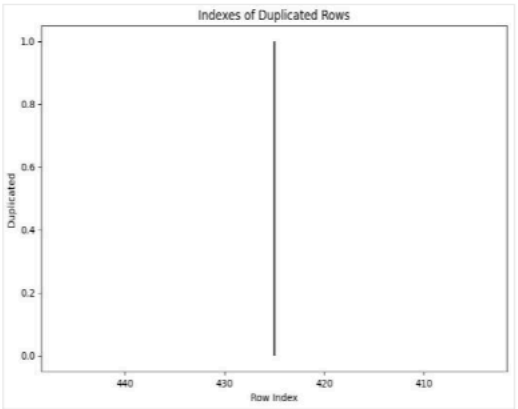

Figure 5. Duplicate data detection visualization graph for glioma grading dataset

Across all datasets, the visualization provided spatial cues about duplication concentration, even when exact row indices were difficult to discern due to axis scaling. This spatial overview is particularly valuable for stakeholders conducting initial data quality assessments, especially in resource-constrained environments. The high number of duplicates in the synthetic healthcare dataset may reflect common data generation practices that prioritize volume over uniqueness. This raises concerns about the validity of predictive models trained on such data. Compared to traditional duplicate detection methods, the visualization tool can reduce initial data assessment time, making it suitable for rapid quality checks in resource-constrained environments.

Table 3 summarizes the findings, highlighting the number of duplicates, data types, and visualization effectiveness. The results confirm that the tool is most effective for small to moderate datasets and homogeneous data structures. For larger or more complex datasets, integration with clustering, similarity matching, or semantic detection algorithms is recommended. While Nguyen *et al*. [36] emphasized the importance of detecting functional redundancy using generative adversarial networks (GANs), our approach focuses on visual interpretability, offering a lightweight alternative for preliminary assessments. Our findings support the observations by Bertrand *et al*. [37] regarding the prevalence of duplication in open-source datasets, but extend their work by offering a scalable visualization method for early detection. Even though the visualization method effectively highlights exact duplicates, it may produce false negatives by missing near or semantic duplicates with slight variations, indicating the need for more advanced similarity-based techniques.

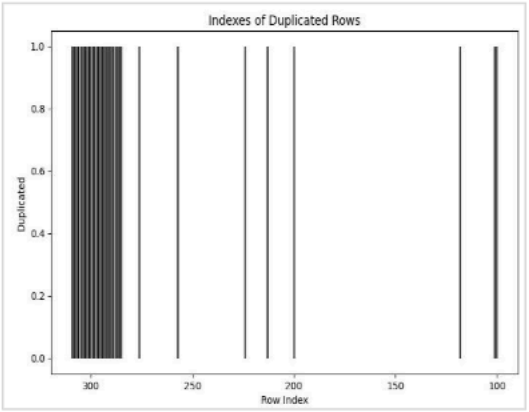

Figure 6. Duplicate data detection visualization graph for online survey lung cancer dataset

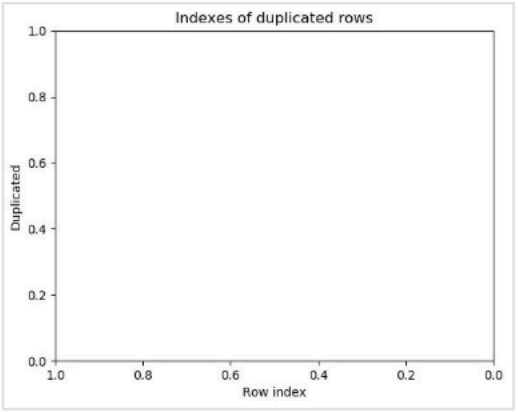

Figure 7. Duplicate data detection visualization graph for breast cancer dataset

Table 3. The key insights derived from the results

Dataset	Size (number of records)	Data type	Visualization capability	Number of duplicates	Visualization effectiveness
Wearable device	5,767	Numerical	√	7	Effective for managing uniform numerical data
Thyroid recurrence	387	Numerical, textual, and categorical	√	19	Mixed data types can be handled
Patient healthcare HER (synthetic)	55,500	Numerical, textual, and categorical	√	534	Detected duplicates but visualization became overcrowded (scalability issues)
Clinical glioma grading	839	Numerical and categorical	√	1	Straightforward and effective
Online survey lung cancer	309	Numerical and categorical	√	33	Clear and accurate insights
Real breast cancer	334	Numerical, textual, and categorical	Null	Null	The presence of duplicates is low causing unclear insights

The results of visualization clarity score is as shown in Table 4. The clarity scores, ranging from 1 to 5, reflect how effectively the visualization tool conveyed duplication patterns across datasets of varying sizes and structures. Datasets with fewer records and more homogeneous data types generally achieved higher clarity scores. For instance, the clinical glioma grading dataset (839 records) received a perfect score of 5, as its single duplicate was easily visualized. Similarly, the wearable device dataset (5,767 records) scored 4.5, demonstrating that the tool performs well with uniform numerical data. In contrast, the patient healthcare EHR (synthetic) dataset, which contains 55,500 records, scored only 2. The visualization became overcrowded, making it difficult to interpret duplication patterns. This result highlights a key limitation of the approach scalability as visual clarity diminishes with increasing dataset size.

Mixed-type datasets such as thyroid recurrence (387 records, score 4) and online survey lung cancer (309 records, score 4.2) showed that the tool remains effective even with heterogeneous data, provided the dataset size is moderate. the real breast cancer dataset, despite having no exact duplicates, was noted as clean and interpretable, suggesting that the visualization tool can still provide value in confirming data integrity, even when duplicates are absent. Overall, these results support the use of visualization as a preliminary data quality assessment tool, particularly for small to medium-sized datasets. However, for large-scale data,

complementary techniques such as clustering or filtering may be necessary to maintain interpretability. This study demonstrated that a Python-based visualization tool can effectively identify exact duplicates in healthcare IoT datasets, particularly those with small to moderate record sizes. The tool showed high clarity and usability in datasets under 1,000 records, with visualization clarity scores ranging from 4.0 to 5.0. In larger datasets, such as the synthetic healthcare EHR dataset with over 55,000 records, visualization became less interpretable, highlighting scalability limitations.

Table 4. Visualization clarity score

Dataset	Number of records	Clarity score (1-5)	Notes
Wearable device	5,767	4.5	Clear visualization of 7 duplicates; suitable for homogeneous data.
Thyroid recurrence	387	4	Visualization interpretable despite mixed types; 19 duplicates detected.
Patient healthcare EHR (synthetic)	55,500	2	Overcrowded visualization due to large size; 534 duplicates detected.
Clinical glioma grading	839	5	Single duplicate easily visualized; high clarity.
Online survey lung cancer	309	4.2	33 duplicates clearly visualized; effective for small mixed-type data.
Real breast cancer	334	5	No exact duplicates; visualization clean and interpretable.

It is important to note that the visualization framework in this study is not intended to function as an independent detection algorithm, but rather as a qualitative tool to support early-stage data quality assessment. The visual output is directly derived from the results of the `duplicated()` function in Pandas, which identifies exact duplicates. Therefore, the visual and actual duplicates are inherently aligned by design. Given that the visualization framework directly reflects the output of the `duplicated()` function in Pandas, designed to identify exact duplicates, the inclusion of a metric such as detection precision, which typically evaluates the correctness of predicted duplicates, would be redundant in this context. Instead, the emphasis of this study is on interpretability, speed of assessment, and practical usability in resource-constrained environments. This clarification is essential to ensure the tool's purpose is accurately understood as a qualitative aid for early-stage data quality evaluation, rather than a standalone detection algorithm.

This study has several limitations that may affect the results. First, the visualization method works best with small to medium-sized datasets; in larger datasets, the plots become overcrowded and harder to interpret. Second, the tool only detects exact duplicates and may miss near-duplicates or semantic duplicates, which could lead to incomplete assessments. Third, the clarity scores used to evaluate visualization effectiveness are based on subjective ratings from a small group of reviewers, which may not reflect broader user experiences. Lastly, the method is designed for static datasets and does not support real-time data, limiting its use in dynamic IoT environments. To enhance the robustness of duplicate detection, future studies should explore integrating the visualization framework with clustering algorithms or AI-based detection models, such as semantic similarity scoring or embedding-based matching. These approaches could help identify complex duplication patterns that are not visually apparent or detectable through exact matching.

The visualization tool can be particularly useful for healthcare data analysts conducting initial data audits. By quickly identifying clusters of duplicates, analysts can prioritize cleaning efforts and ensure higher data integrity before deploying predictive models or conducting statistical analyses. Future research can extend this approach by integrating it with semantic similarity algorithms or embedding-based models to detect more complex forms of duplication. Additionally, adapting the visualization for streaming IoT data could enable real-time quality monitoring in clinical environments.

4 CONCLUSION

This study investigated the effectiveness of a Python-based visualization tool for detecting data duplication in IoT-based healthcare datasets. By applying the method across six cancer-related datasets of varying sizes and data types, the study demonstrated that visualization can serve as a rapid and intuitive approach for preliminary data quality assessment. The results confirmed that the tool is particularly effective for small to moderately sized datasets, where duplication patterns are clearly represented. In these cases, visualization provided immediate insights into data integrity, aligning with prior research that emphasizes the value of visual analytics in early-stage data exploration. However, in larger datasets, such as the synthetic healthcare EHR dataset, the visualization became less interpretable due to overcrowding, highlighting a scalability limitation. This observation supports the findings that advocate for clustering and blocking techniques to manage large-scale data quality issues. The study also revealed that while exact duplicates are easily detected, semantic and temporal duplicates may require more advanced techniques. This underscores the need for hybrid approaches that combine visualization with content-aware or semantic detection

algorithms to improve robustness. Overall, the visualization method offers a lightweight, interpretable, and generalizable tool for initial duplicate detection, especially in resource-constrained environments. Its adaptability across diverse data types and domains makes it a valuable addition to the data quality toolkit. Future research should explore integrating this approach with machine learning-based detection methods to enhance scalability and accuracy. Additionally, extending the framework to address other data quality dimensions such as missing values, inconsistencies, and outliers could further improve its utility in real-world IoT applications. For the research field, this method offers a lightweight alternative to complex algorithms, making early-stage data checks more accessible. For the healthcare community, it supports better data integrity, which is crucial for accurate patient care and decision-making.

ACKNOWLEDGMENTS

The authors would like to thank the Centre for Research and Innovation Management (CRIM), UTeM, and members of the IDEAL research group for all their support.

FUNDING INFORMATION

This study is funded by Malaysian Technical University Network (MTUN) Strategic Collaboration Research Grant (MTUN/2024/UTEM-FTMK/CRG/MS0008).

AUTHOR CONTRIBUTIONS STATEMENT

This journal 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
Siti Noor Basirah Md Isa	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Nurul A. Emran		✓				✓		✓	✓	✓	✓	✓	✓	
Norharyati Harum	✓			✓						✓	✓			
Logenthiran Machap	✓									✓				✓
Azlin Nordin	✓						✓			✓				

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest regarding this research.

DATA AVAILABILITY

The data that were used in this study are openly available from the URLs as follows:

- Wearable device: https://figshare.com/articles/dataset/Bern2019WDP_data/13471338
- Thyroid recurrence: https://www.kaggle.com/datasets/jainaru/thyroid-disease-data
- Patient healthcare HER (synthetic): https://www.kaggle.com/datasets/prasad22/healthcare-dataset
- Clinical glioma grading: https://www.kaggle.com/datasets/tanshihjen/clinical-gliomagrading
- Online survey lung cancer: https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer
- Real breast cancer: https://www.kaggle.com/datasets/amandam1/breastcancerdataset/data

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

REFERENCES

[1] A. Singh and S. Mahapatra, "Network-Based Applications of Multimedia Big Data Computing in IoT Environment," in *Multimedia Big Data Computing for IoT Applications: Concepts, Paradigms and Solutions*, 2020, pp. 435–452, doi: 10.1007/978-981-13-8759-3_17.
[2] S. Vongsingthong and S. Smanchat, "A Review of Data Management in Internet of Things," *KKU Research Journal*, vol. 20, no.

2, pp. 215–240, 2015, doi: 10.14456/kkurj.2015.18.
[3] H. Shahid, M. A. Vazquez, L. Reynaud, F. Parzysz, and M. Shaat, "Open Datasets for AI-Enabled Radio Resource Control in Non-Terrestrial Networks," in *Proceedings - 2023 IEEE Future Networks World Forum: Future Networks: Imagining the Network of the Future, FNWF 2023*, 2023, pp. 1–6, doi: 10.1109/fnwf58287.2023.10520488.
[4] T. Mansouri, M. R. S. Moghadam, F. Monshizadeh, and A. Zareravasan, "IoT Data Quality Issues and Potential Solutions: A Literature Review," *The Computer Journal*, vol. 66, no. 3, pp. 615–625, 2021, doi: 10.1093/comjnl/bxab183.
[5] H. Y. Teh, A. W. Kempa-Liehr, and K. I. K. Wang, "Sensor data quality: a systematic review," *Journal of Big Data*, vol. 7, no. 1. Springer Science and Business Media Deutschland GmbH, 2020, doi: 10.1186/s40537-020-0285-1.
[6] D. Salian, "Usability of Open Data," in *Open-Source Horizons*, L. M. Castro, Ed. Rijeka: IntechOpen, 2023.
[7] M. A. McClellan, "Duplicate medical records: a survey of Twin Cities healthcare organizations," in *Annual Symposium Proceedings/AMIA Symposium*, 2009, vol. 2009, pp. 421–425.
[8] S. Aydin and M. N. Aydin, "Semantic and Syntactic Interoperability for Agricultural Open-Data Platforms in the Context of IoT Using Crop-Specific Trait Ontologies," *Applied Sciences*, vol. 10, no. 13, pp. 1-27, Jun. 2020, doi: 10.3390/app10134460.
[9] W. Xia, H. Jiang, D. Feng, and L. Tian, "DARE: A Deduplication-Aware Resemblance Detection and Elimination Scheme for Data Reduction with Low Overheads," *IEEE Transactions on Computers*, vol. 65, no. 6, pp. 1692–1705, 2016, doi: 10.1109/tc.2015.2456015.
[10] Y. Chen, D. Li, Y. Hua, and W. He, "Effective and Efficient Content Redundancy Detection of Web Videos," *IEEE Transactions on Big Data*, vol. 7, no. 1, pp. 187–198, 2019, doi: 10.1109/tbdata.2019.2913674.
[11] K. Vijayalakshmi and V. Jayalakshmi, "Analysis on data deduplication techniques of storage of big data in cloud," in *Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021*, 2021, pp. 976–983, doi: 10.1109/ICCMC51019.2021.9418445.
[12] M. U. Tahir, M. R. Naqvi, S. K. Shahzad, and M. W. Iqbal, "Resolving Data De-Duplication issues on Cloud," in *2020 International Conference on Engineering and Emerging Technologies (ICEET),* Lahore, Pakistan, 2020, pp. 1–5, doi: 10.1109/iceet48479.2020.9048214.
[13] L. Lu and P. Wang, "Duplication detection in news articles based on big data," in *IEEE 4th International Conference on Cloud Computing and Big Data Analytics, ICCCBDA 2019*, 2019, pp. 15–19, doi: icccbda.2019.8725674.
[14] S. An-Dong and Z. Fang, "Research on Open Source Solutions of Data Collection for Industrial Internet of Things," in *Proceedings - 2021 7th International Symposium on Mechatronics and Industrial Informatics, ISMII 2021*, 2021, pp. 180–183, doi: 10.1109/ISMII52409.2021.00045.
[15] O. Sahin, A. Zhao, R. J. Applegate, T. R. Johnson, and E. V. Bernstam, "Epidemiology of Patient Record Duplication," *Applied Clinical Informatics*, vol. 16, no. 1, pp. 24–30, 2025, doi: 10.1055/a-2423-8499.
[16] L. F. Pinto and L. J. Dos Santos, "Electronic medical records in primary care: Management of duplicate records and a contribution to epidemiological studies," *Ciencia e Saude Coletiva*, vol. 25, no. 4, pp. 1305–1312, 2020, doi: 10.1590/1413-81232020254.34132019.
[17] F. N. M. Leza and N. A. Emran, "Data accessibility model using QR code for lifetime healthcare records," *World Applied Sciences Journal*, vol. 30, no. 30 A, pp. 395–402, 2014, doi: 10.5829/idosi.wasj.2014.30.icmrp.55.
[18] M. Vattulainen, "Data quality visualization for preprocessing," in *Lecture Notes in Computer Science*, 2016, vol. 9728, pp. 428–437, doi: 10.1007/978-3-319-41561-1_32.
[19] R. Sulo, S. Eick, and R. Grossman, "DaVis: a tool for visualizing data quality," *Posters Compendium of InfoVis*, pp. 1-6, 2005.
[20] S. Liu *et al.*, "Steering data quality with visual analytics: The complexity challenge," *Visual Informatics*, vol. 2, no. 4, pp. 191–197, 2018, doi: 0.1016/j.visinf.2018.12.001.
[21] W. Jiang, B. Wu, Z. Jiang, and S. Yang, "Cloning Vulnerability Detection in Driver Layer of IoT Devices," in *Lecture Notes in Computer Science*, 2020, pp. 89–104, doi: 10.1016/j.sysarc.2023.102961.
[22] Y. Gao, L. Chen, J. Han, G. Wu, and S. Liu, "Similarity-based deduplication and secure auditing in IoT decentralized storage," *Journal of Systems Architecture*, vol. 142, 2023, doi: 10.1016/j.sysarc.2023.102961.
[23] H. Li, H. Lu, C. S. Jensen, B. Tang, and M. A. Cheema, "Spatial Data Quality in the Internet of Things: Management, Exploitation, and Prospects," *ACM Computing Surveys*, vol. 55, no. 3, 2022, doi: 10.1016/j.sysarc.2023.102961.
[24] S. An-Dong and Z. Fang, "Research on Open Source Solutions of Data Collection for Industrial Internet of Things," in *Proceedings - 2021 7th International Symposium on Mechatronics and Industrial Informatics, ISMII 2021*, 2021, doi: 10.1109/ISMII52409.2021.00045.
[25] B. H. Just and K. Proffitt, "Do you know who's who in your EHR?," *Journal of the Healthcare Financial Management Association*, vol. 63, no. 8, pp. 68–73, 2009.
[26] M. U. Tahir, M. R. Naqvi, S. K. Shahzad, and M. W. Iqbal, "Resolving Data De-Duplication issues on Cloud," in *2020 International Conference on Engineering and Emerging Technologies (ICEET)*, 2020, pp. 1–5, doi: 10.1109/iceet48479.2020.9048214.
[27] L. Lu and P. Wang, "Duplication Detection in News Articles Based on Big Data," in *IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, 2019, pp. 15–19, doi: 10.1109/icccbda.2019.8725674.
[28] Y. Chen, D. Li, L. Yan, and Z. Ma, "Two-stage Detection of Semantic Redundancies in RDF Data," *Journal of Web Engineering*, vol. 21, no. 8, pp. 2313–2337, 2022, doi: 10.13052/jwe1540-9589.2184.
[29] A. Ali, N. A. Emran, and S. A. Asmai, "Missing values compensation in duplicates detection using hot deck method," *Journal of Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00502-1.
[30] Y. Long, H. Li, Z. Wan, and P. Tian, "Data Redundancy Detection Algorithm based on Multidimensional Similarity," in *Proceedings - 2023 International Conference on Frontiers of Robotics and Software Engineering, FRSE 2023*, 2023, pp. 180–187, doi: 10.1109/frse58934.2023.00032.
[31] I. Koumarelas, L. Jiang, and F. Naumann, "Data Preparation for Duplicate Detection," *Journal of Data and Information Quality*, vol. 12, no. 3, pp. 1–24, 2020, doi: 10.1145/3377878.
[32] Y. Huang and F. Chiang, "Refining duplicate detection for improved data quality," in *Proceedings of the International Workshop on Data Quality (DQ)*, 2017, pp. 1–10, doi: 10.1145/3377878.
[33] R. Tatman, "Data Cleaning Challenge: Deduplication," *kaggle*, 2018. [Online]. Available: https://www.kaggle.com/code/rtatman/data-cleaning-challenge-deduplication. (accessed: Aug. 05, 2025).
[34] A. Ali, N. A. Emran, S. A. Asmai, and A. Thabet, "Duplicates detection within incomplete data sets using blocking and dynamic sorting key methods," *International Journal of Advanced Computer Science and Applications*, 2018, doi: 10.14569/ijacsa.2018.090979.
[35] Y. Huang and F. Chiang, "Refining Duplicate Detection for Improved Data Quality," in *TDDL, MDQual and Futurity Workshops at TPDL 2017*, 2017, pp. 1–10, doi: 10.1145/3377878.
[36] T. T. Nguyen, T. T. Huynh, M. T. Pham, T. D. Hoang, T. T. Nguyen, and Q. V. H. Nguyen, "Validating functional redundancy with mixed generative adversarial networks," *Knowledge-Based Systems*, vol. 264, p. 110342, 2023, doi: 10.1016/j.knosys.2023.110342.
[37] Y. Bertrand, R. Van Belle, J. De Weerdt, and E. Serral, "Defining Data Quality Issues in Process Mining with IoT Data," in *Process Mining Workshops*, 2023, pp. 422–434, doi: 10.1007/978-3-031-27815-0_31.

BIOGRAPHIES OF AUTHORS

Siti Noor Basirah Md Isa is a recent graduate with a Master's degree in Computer Science (Database Technology) from Universiti Teknikal Malaysia Melaka (UTeM) in 2024. She holds a B.Sc. in Statistics from Universiti Teknologi MARA (UiTM) in 2015 and has 6 years of experience in Marketing and Sales Operations Support across the automotive and logistics industries where she is responsible for data-driven decision-making, customer insights analysis, business reporting, and process optimization. Her expertise includes data quality, database systems, data visualization, and business intelligence, with research interests in IoT data quality, database optimization, and data analytics. She can be contacted at email: sitinoorbasirah.isa@gmail.com.

Associate Professor Dr. Nurul A. Emran is an Associate Professor at Universiti Teknikal Malaysia Melaka (UTeM). She earned her bachelor's degree in Management Information Systems (MIS) from the International Islamic University Malaysia in 2001, followed by an M.Sc. in Internet and Database Systems from London South Bank University in 2003. She later obtained her Ph.D. in Computer Science from the University of Manchester, UK, in 2011. Her academic career began in 2004 when she was appointed as a lecturer. She was promoted to Senior Lecturer in 2011 before attaining the rank of Associate Professor in 2017. Her research interests span data quality, IoT, database systems and security, storage space optimization, mobile analytics, and green computing. She can be contacted at email: nurulakmar@utem.edu.my.

Dr. Norharyati Harum is currently a senior lecturer at the Universiti Teknikal Malaysia (UTeM). She received bachelor's degree Information Engineering in 2003, Master in Engineering (Science for Open and Environmental Systems) in 2009. and Ph.D. in Engineering (Science for Open and Environmental Systems) in 2012 from Keio University, Japan. She has working experince in R & D Next Generation Mobile Communication Department at Panasonic Japan from 2005-2009. Her research area includes machine learning, internet of things (IoT), cloud computing, wireless sensor network, and signal processing. She can be contacted at email: norharyati@utem.edu.my.

Dr. Logenthiran Machap is a senior lecturer at Universiti Tun Hussein Onn Malaysia (UTHM), affiliated with the Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology. He obtained his BSc in Bioinformatics (Hons) from Universiti Kebangsaan Malaysia (UKM) in 2010, a Master of Computer Science (Database Technology) from Universiti Teknikal Malaysia Melaka (UTeM) in 2013, and a Ph.D. in Computer Science (Bioinformatics) from Universiti Teknologi Malaysia (UTM) in 2020. He joined UTHM academia in 2023 and has been actively involved in research and teaching. He leads the Computational Intelligence and Data Science (CIDaS) Focus Group, with research interests spanning bioinformatics, machine learning, deep learning for biomedical applications, and computational approaches in data science. He can be contacted at email: logen@uthm.edu.my.

Dr. Azlin Nordin is a senior lecturer at the Department of Computer Science, Kulliyyah of Information and Communication Technology (KICT), International Islamic University Malaysia (IIUM). She received her bachelor degree from B.IT (Hons), Universiti Utara Malaysia in 1998. Later, she obtained her M.Sc. in Real-Time Software Engineering, Universiti Teknologi Malaysia in 2001 and Ph.D. in Computer Science, University of Manchester in 2013. She is a Certified Professional in Requirements Engineering (CPRE) and Certified Testing Foundation Level (CTFL). Her research work specializing in requirements engineering, and software quality assurance. She is also active in postgraduate supervision, scholarly publications, and academic leadership roles. She can be contacted at email: azlinnordin@iium.edu.my.