DATA QUALITY IN IOT OPEN DATASETS: A METHODOLOGICAL REVIEW

S.N.B.M. Isa[1], **N.A. Emran**[1]

[1]Fakulti Teknologi Maklumat dan Komunikasi,
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian
Tunggal, Melaka, Malaysia.

Corresponding Author's Email: [1]nurulakmar@utem.edu.my

Article History: Received 24 September 2024; Revised 24 October 2024; Accepted 29 November 2024

ABSTRACT: The growth of the Internet of Things (IoT) has significantly increased data generated by connected devices, leading to challenges in data duplication that threaten data quality and reliability. The purpose of this study is to assess and thoroughly examine the quality of open-source IoT datasets, focusing on the occurrence and impact of duplicate data. By employing a Systematic Literature Review (SLR) and a literature-based comparative analysis, we reviewed and compared existing techniques for detecting these issues. Our findings reveal that while various methods have been proposed, there remains a lack of standardized approaches specifically designed for the unique characteristics of IoT environments. The study concludes by highlighting the need for more reliable and scalable solutions that are capable of handling the diverse and dynamic nature of IoT data, also offering insights into future research directions.

KEYWORDS: *IoT Data Quality, Open-Source Datasets, Data Duplication, Systematic Literature Review (SLR) and Comparative Study*

1.0 INTRODUCTION

The Internet of Things (IoT) has transformed how we interact with technology, enabling devices from industrial sensors to home appliances to communicate and share data autonomously (Gowda et al., 2022; Rajamohan et al., 2023). This vast network generates large amounts of data, driving automation and smarter decision-making, but also raises challenges like data duplication (Nguyen et al., 2023).

With IoT devices expected to grow from 15.9 billion in 2023 to 32.1 billion by 2030 (Statista, 2024), managing and ensuring data quality is critical. Data duplication, in particular, impacts the reliability and usefulness of IoT datasets, especially in open-source datasets (Mansouri et al., 2021). Despite its importance, systematic studies on data duplication in open-source IoT datasets are limited.

This review addresses the gap by exploring current studies and methods for detecting data duplication in IoT data. It evaluates detection approaches, highlights the impact of duplication, and stresses the need for better solutions.

Key contributions include analyzing data duplication in open-source IoT datasets to improve data quality and reliability. The review also provides practical insights for researchers, offering data management strategies that can enhance IoT technologies across various domains. Ultimately, this review serves as a guide for future research in IoT data quality management.

2.0 LITERATURE REVIEW

2.1 IoT data types and characteristics

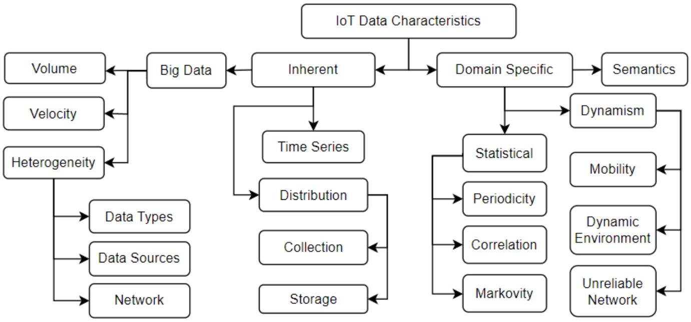

Figure 1: IoT Data Characteristics

IoT data has distinct characteristics that set it apart from traditional big data. Zubair et al. (2019) classify these into Inherent and Domain-Specific Features. Inherent features include high volume, velocity, and heterogeneity, as IoT data comes from various types and sources. Singh & Mahapatra (2020) highlight that IoT devices produce structured, semi-structured, and unstructured data, while formats range from numerical to multimedia (Vongsingthong & Smanchat, 2015).

Domain-specific features like mobility, dynamic conditions, and unreliable networks reflect IoT's constantly changing environment.

Additionally, IoT data shows statistical properties such as periodicity and correlation, with continuous streams generated by sensor devices, forming time series data Chowdhury et al. (2024) and Puschmann et al. (2017).

2.2 Open-Source Datasets

Open-source IoT datasets play a crucial role in IoT research and development, providing accessible data for analysis and innovation (Salian, 2023). Shahid et al. (2024) highlight their value in offering realistic traffic patterns, network performance, and user demand scenarios. However, the quality of these datasets is a critical concern. Bertrand et al. (2023) and Kabi et al. (2023) identify main quality issues such as outliers, duplicates, anomalies, and missing values due to massive amount of data by IoT devices, that often leads to noise, uncertainty, and incompleteness which can undermine data reliability and result in poor decision-making (Byabazaire et al., 2022).

2.3 Data Duplication

Data duplication is a major issue in IoT datasets, varying across different domains. Widad et al. (2023) define duplication in big data as redundant entries referring to the same real-world entity. In edge computing, Tverdal et al. (2023) describe duplicates as adjacent records with identical timestamps. Similarly, in environmental datasets, Buelvas et al. (2023) identify redundancy as multiple items sharing the same timestamp.

In the IoT context, data duplication takes on specific forms, reflecting the unique characteristics of IoT data. Table 1 shows the categories of duplicate data in IoT and its example.

Table 1: Categories of duplicate data in IoT

Category	Definition & Example
Exact Duplicates	Records identical in all aspects, including data values and timestamps, can be from multiple sensors, thereby skewing data accuracy and reliability (Jiang et al., 2020).
Example	Location: 1, Sensor: A, Temperature: 25.0°C, Humidity: 60%, Time: 12:00 PM Location: 1, Sensor: A, Temperature: 25.0°C, Humidity: 60%, Time: 12:00 PM
Near Duplicates	Records with minor variations but representing the same real-world entity or data points with slightly different content, where retaining any one does not impact the analysis outcome (Gao et al., 2023).
Example	Location: 2, Sensor: B, Temperature: 24.8°C, Humidity: 58%, Time: 12:05 PM Location: 2, Sensor: C, Temperature: 25.1°C, Humidity: 57%, Time: 12:05 PM

Spatial Duplicates *Example*	Data from multiple sensors in close proximity reporting similar information, potentially causing inaccuracies in data analysis (Li et al., 2022). Location: Park, Coordinates: (40.7128, -74.0060), Sensor: D, Temperature: 23.0°C, Humidity: 55%, Time: 12:10 PM
	Location: Park, Coordinates: (40.7129, -74.0061), Sensor: E, Temperature: 23.0°C, Humidity: 55%, Time: 12:10 PM
Temporal Duplicates *Example*	Nearly identical data is reported over short time intervals by the same sensor, collected at different time points, which can occur in time-series databases and impact the accuracy of data analysis (An-Dong & Fang, 2021; Wang et al., 2020). Location: 3, Sensor: F, Temperature: 26.0°C, Humidity: 62%, Time: 12:15:01 PM Location: 3, Sensor: F, Temperature: 26.0°C, Humidity: 62%, Time: 12:15:02 PM
Semantic Duplicates *Example*	Data points representing the same information but expressed differently (Aydin & Aydin, 2020), occur due to variations in data formats, naming conventions, or contextual usage of terms across different datasets (Muralidharan et al., 2020). Location: 4, Sensor: G, Temperature: 22.0°C, Humidity: 65%, Time: 12:20 PM Location: 4, Sensor: H, Temperature: 71.6°F, Humidity: 65%, Time: 12:20 PM

The impact of data duplication is broad, affecting data quality, consistency, analysis, and decision-making. The following table summarizes the key impacts of data duplication across different domains.

Table 2: Summary of key impacts of data duplication

Impact	Descriptions
Data Quality and Consistency	•Challenges in updating/ modifying across every instance •Propagation of errors and misreporting of information •Erroneous results in analysis •Negative impact on data utilization and decision quality (Ding et al., 2022; Long et al., 2023; Nguyen et al., 2023; Searle et al., 2021)
Analysis and Decision-Making	•Difficult to identify recent and reliable data across multiple instances •Anomalies and unexpected query results •Distort analysis results and inflated statistical measures lead to inaccurate conclusions •Affecting reliability of data-driven decision (Firmani et al., 2016; Frank, 2024; Nguyen et al., 2023)
Resource Utilization and Efficiency	•Utilizes high storage space •Higher energy consumption and additional communication costs •Inefficiencies (Firmani et al., 2016; Frank, 2024; Nguyen et al., 2023)
Genetic and Biological Implications	•Selection advantage in changing environments •Increased genetic diversity and appearance of new functions •Evolution of viruses (like HIV) and complex traits (such as fertility) (Fouchécourt et al., 2022; Girard-Madoux et al., 2018; Müller & Sauter, 2023)
Signal Processing	• Add redundant information to prevent channel failures and packet losses, allowing for error detection and correction to reconstruct the received signal with minimal distortion (Ma & Labeau, 2012)

Despite these challenges, addressing data duplication through deduplication techniques offers numerous benefits across various domains. The following table summarizes the benefits of eliminating data duplication.

Table 3: Summary of benefits of data deduplication

Category	Description
Cost	Helps lower storage costs & costs associated with maintaining duplicated data (Adhab & Hussien, 2022; Deshingkar et al., 2023; Khan et al., 2024; Tahir et al., 2020)
Storage	Effective data storage, keeps only useful data in storage, saving storage space (Nguyen et al., 2023; Tahir et al., 2020)
Data Management	•efficient data migrations, reducing the volume of data to be transferred (Tahir et al., 2020) •allow easy updating and cross-referencing (Searle et al., 2021) •efficient data updating, error-free query processing (Nguyen et al., 2023)
Performance	•improves backup and restore time, improving performance and efficiency (Tahir et al., 2020) •reduces data volume, enhancing overall system's performance (Khan et al., 2024)
Miscellaneous	•facilitating effective disaster recovery by replicating data after removing redundant data (Tahir et al., 2020) •keep precise info to help medical professionals focus on only critical readings (Khan et al., 2024)

The unique characteristics of IoT data, combined with the prevalence of data quality issues in open-source datasets, underscore the need for effective strategies to manage and mitigate data duplication. This paper will compare existing methods that allow researchers and practitioners to make informed decisions about which approaches are suitable for specific IoT applications and data characteristics.

3.0 METHODOLOGY

This study employs a Systematic Literature Review (SLR) to analyze research on data duplication detection in IoT open-source datasets. Searches were conducted in IEEE Xplore, ACM Digital Library, Scopus, and MDPI, utilizing keywords such as "Internet of Things" OR "IoT," AND ("data quality" OR "data integrity"), AND ("open-source dataset"), AND ("duplicate data" OR "data redundancy"). This yielded 110 studies, with 95 unique records remaining after duplicate removal. The screening process involved two stages: an initial review of titles and abstracts that excluded 30 irrelevant papers, followed by a full-text review of 65 studies. Ultimately, 47 studies relevant to data duplication in IoT datasets were selected.

Data was extracted using a standardized form, focusing on study details, methods for handling duplicates, and key findings. This information was synthesized to identify themes, trends, and gaps in research. A comparative analysis evaluated methods for detecting and managing data duplication, categorizing them by strengths and limitations to provide insights into their effectiveness in IoT environments.

4.0 RESULT AND DISCUSSION

The analysis of the literature reveals three primary categories of duplicate data 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.

4.1 Hash-based Methods

Hash-based methods calculate unique hash values or fingerprints for each data chunk or file. Vijayalakshmi & Jayalakshmi (2021) explain this process in their study for cloud storage, while Deshingkar et al. (2023) used SHA-256 (Python code) to remove duplicate files, keeping only the latest copy.

Advantages:

- Quick identification of identical data
- Low computational cost
- Efficient for exact duplicate detection

Limitations:

- Ineffective for detecting near or semantically similar data
- Vulnerable to hash collisions with weaker hash functions

These methods are ideal for exact duplicate detection, such as in backup systems or storage optimization (Tahir et al., 2020). However, Chen et al. (2019) note that these methods may not be effective in detecting similar duplicates or near-duplicates that have minor differences in content.

4.2 Content-aware Methods

Content-aware methods analyze data content to detect similarities, employing techniques from fields like computer vision, NLP, and machine learning (Chen et al., 2019, 2022; Long et al., 2023).

Long et al. (2023) proposed the Multidimensional Similarity Redundancy Detection Algorithm (MSRD), that combines numerical similarity, literal similarity, and semantic similarity for comprehensive duplicate detection. Chen et al. (2022) developed a two-stage detection approach utilizing locality-sensitive hashing (LSH) to efficiently select candidate pairs in RDF data and perform similarity analysis on the selected pairs to identify duplicates. Chen et al. (2019) introduced the CompoundEyes method, which detects near-duplicate videos by utilizing multiple features and classifiers in parallel.

Advantages:

- Effective for near-duplicates and semantically similar data

- Comprehensive similarity calculations
- Suitable for complex data types and structures

Limitations:

- Computationally intensive
- Requires significant processing power and resources
- Accuracy can be influenced by the quality of data

Content-aware methods excel in detecting near-duplicates or semantically similar data, that ideal in plagiarism detection, copyright infringement detection, or data integration (Chen et al., 2022; Lu & Wang, 2019).

4.3 Hybrid and Advanced Approaches

Hybrid and advanced approaches combine multiple techniques or incorporate additional processing steps to enhance data redundancy detection accuracy and efficiency. These methods often aim to address specific challenges or requirements, such as handling frequently modified data, dealing with incomplete or missing data, or optimizing performance for large-scale datasets.

Xia et al. (2016) proposed the Deduplication-Aware Resemblance Detection and Elimination (DARE) scheme, using duplicate-adjacency (DupAdj), an improved super-feature approach, followed by delta compression to remove redundancy. Huang & Chiang (2017) developed a Deduplication Framework with Metric Functional Dependencies (MFDs) and a weighting scheme for term frequency. Ali et al. (2021) introduced the Duplicate Detection within Incomplete Datasets (DDID) method, which selects attributes for generating sort keys for clustering and comparison strings for matching records and Hot Deck imputation to compensate for missing values.

Advantages:

- Handles complex challenges (e.g., modified or incomplete data)
- Potentially more robust and accurate detection
- Suitable for complex scenarios with specific requirements

Limitations:

- Complex to implement
- Requires additional computational resources and expertise
- May be overkill for simpler duplicate detection needs

The effectiveness of these methods depends on the specific use case. For example, the DARE scheme by Xia et al. (2016) may be effective for backup or archiving systems that deal with frequently changing data, while the Deduplication Framework proposed by Huang & Chiang

(2017) can be effective in scenarios where numeric data or term frequencies play a crucial role in detecting duplicates.

Table 4: Summary of key impacts of data duplication

Category	Advantages	Disadvantages	Suitable	References
Hash-based Methods	- Quick identification of identical data - Low computational cost	- Ineffective for detecting near-duplicates - Vulnerable to hash	- File deduplication in backup systems	(Chen et al., 2019; Deshingkar et al., 2023; Tahir et al., 2020;
	- Efficient for exact duplicate detection	collisions with weaker hash functions	- Storage optimization	Vijayalakshmi & Jayalakshmi, 2021)
Content-aware Methods	- Effective for detecting near-duplicates and semantically similar data - Comprehensive similarity calculations	- Computationally intensive - Requires significant processing power and resources - Accuracy influenced by quality of data	- Plagiarism detection - Copyright infringement detection - Data integration tasks	(Ali et al., 2021; Chen et al., 2019, 2022; Long et al., 2023; Lu & Wang, 2019)
Hybrid/ Advanced Methods	- 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	(Chen et al., 2022; Huang & Chiang, 2017; Xia et al., 2016)

The choice of duplicate detection method in IoT depends on the application's needs. Hash-based methods work well for storage optimization, while content-aware or hybrid methods are better for detailed duplicate detection, such as data integration or handling sensor data from multiple sources. Hybrid and advanced methods are increasingly popular as they balance computational efficiency with detection accuracy, addressing challenges in frequently modified or incomplete datasets. As IoT technologies evolve, ensuring data quality must also progress. Future research should aim for scalable, adaptive methods that uphold accuracy as IoT datasets expand in size and complexity.

5.0 CONCLUSIONS

This review paper highlights the multilayered nature of data duplication across multiple domains, emphasizing its impact in IoT environments. The analysis reveals a spectrum of duplicate detection methods, from basic hash-based techniques to advanced content-aware and hybrid approaches. Each method presents unique advantages and limitations, highlighting the need for context-specific solutions to manage data redundancy. Developing adaptive and scalable detection methods is crucial for future research, aiming to balance computational efficiency with detection accuracy in increasingly complex data ecosystems as IoT technologies evolve.

ACKNOWLEDGMENTS

This work was conducted independently and did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] Adhab, A. H., & Hussien, N. A. (2022). *Techniques of Data Deduplication for Cloud Storage: A Review*. [Online]. Available: https://doi.org/10.31695/IJERAT.2022.8.4.2

[2] Ali, A., Emran, N. A., & Asmai, S. A. (2021). *Missing values compensation in duplicates detection using hot deck method*. [Online]. Available: https://doi.org/10.1186/s40537-021-00502-1

[3] Aydin, S., & Aydin, M. N. (2020). *Semantic and syntactic interoperability for agricultural open-data platforms in the context of IoT using crop-specific trait ontologies.* [Online]. Available: https://doi.org/10.3390/app10134460

[4] Bertrand, Y., Van Belle, R., De Weerdt, J., & Serral, E. (2023). *Defining Data Quality Issues in Process Mining with IoT Data*. [Online]. Available: https://doi.org/10.1007/978-3-031-27815-0_31

[5] Buelvas, J. H., Múnera, D., & Gaviria, N. (2023). *DQ-MAN: A tool for multi-dimensional data quality analysis in IoT-based air quality monitoring systems*. [Online]. Available: https://doi.org/10.1016/J.IOT.2023.100769

[6] Byabazaire, J., O'Hare, G. M. P., Collier, R., & Delaney, D. (2022). *Dynamic Data Source Selection: A Case of Weather Stations for IoT Applications*. [Online]. Available: https://doi.org/10.1109/WF-IoT54382.2022.10152030

[7] Chen, Y., Li, D., Hua, Y., & He, W. (2019). *Effective and Efficient Content Redundancy Detection of Web Videos*. [Online]. Available: https://doi.org/10.1109/tbdata.2019.2913674

[8] Chowdhury, A., Pal, A., Raut, A., & Kumar, M. (2024). *KIHCDP: An Incremental Hierarchical Clustering Approach for IoT Data Using Dirichlet Process*. [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3385628

[9] Deshingkar, H., Desai, P., Deshmane, S., Deshmukh, A., Desai, T., Desai, S., & Ubale, G. (2023). *Data Deduplication Using Python*. [Online]. Available: https://doi.org/10.1109/ICCUBEA58933.2023.10391968

[10] Ding, X., Wang, H., Li, G., Li, H., Li, Y., & Liu, Y. (2022). *IoT data cleaning techniques: A survey*. [Online]. Available: https://doi.org/10.23919/ICN.2022.0026

[11] Firmani, D., Mecella, M., Scannapieco, M., & Batini, C. (2016). *On the Meaningfulness of "Big Data Quality" (Invited Paper)*. [Online]. Available: https://doi.org/10.1007/s41019-015-0004-7

[12] Fouchécourt, S., Fillon, V., Marrauld, C., Callot, C., Ronsin, S., Picolo, F., Douet, C., Piégu, B., & Monget, P. (2022). *Expanding duplication of the testis PHD Finger Protein 7 (PHF7) gene in the chicken genome*. [Online]. Available: https://doi.org/10.1016/J.YGENO.2022.110411

[13] Frank, E. (2024). *Machine Learning Models for Data Quality Assessment*.

[14] Gao, Y., Chen, L., Han, J., Wu, G., & Liu, S. (2023). *Similarity-based deduplication and secure auditing in IoT decentralized storage*. [Online]. Available: https://doi.org/10.1016/j.sysarc.2023.102961

[15] Girard-Madoux, M. J. H., Gomez de Agüero, M., Ganal-Vonarburg, S. C., Mooser, C., Belz, G. T., Macpherson, A. J., & Vivier, E. (2018). *The immunological functions of the Appendix: An example of redundancy?* [Online]. Available: https://doi.org/10.1016/j.smim.2018.02.005

[16] Gowda, V. D., Bandaru, V. N. R., Begum, A. Y., Palanikkumar, D., & Jadhav, A. C. (2022). *Internet of Things (IoT): Definitions, Components, Characteristics and Applications*. [Online]. Available: https://doi.org/10.9734/bpi/costr/v8/3535c

[17] Huang, Y., & Chiang, F. (2017). *Refining Duplicate Detection for Improved Data Quality*. [Online]. Available: https://ceur-ws.org/Vol-2038/paper3.pdf

[18] Jiang, W. P., Wu, B., Jiang, Z., & Yang, S. B. (2020). *Cloning Vulnerability Detection in Driver Layer of IoT Devices*. [Online]. Available: https://doi.org/10.1007/978-3-030-41579-2_6

[19] Kabi, J. N., Maina, C. W., & Mharakurwa, E. T. (2023). *Anomaly Detection in IoT Data*. [Online]. Available: https://repository.dkut.ac.ke:8080/xmlui/bitstream/handle/123456789/8028/Anomaly_Detection_in_IoT_Data.pdf?sequence=1&isAllowed=y

[20] Khan, M. N. U., Cao, W., Tang, Z., Ullah, A., & Pan, W. (2024). *Energy-Efficient De-Duplication Mechanism for Healthcare Data Aggregation in IoT*. [Online]. Available: https://doi.org/10.3390/fi16020066

[21] Li, H., Lu, H., Jensen, C. S., Tang, B., & Cheema, M. A. (2022). *Spatial Data*

Quality in the Internet of Things: Management, Exploitation, and Prospects. [Online]. Available: https://doi.org/10.1145/3498338

[22] Long, Y., Li, H., Wan, Z., & Tian, P. (2023). *Data Redundancy Detection Algorithm based on Multidimensional Similarity*. [Online]. Available: https://doi.org/10.1109/FRSE58934.2023.00032

[23] Lu, L., & Wang, P. (2019). *Duplication Detection in News Articles Based on Big Data*. [Online]. Available: https://doi.org/10.1109/ICCCBDA.2019.8725674

[24] Ma, R., & Labeau, F. (2012). *A family of fast index and redundancy assignments for error resilient multiple description coding*. [Online]. Available: https://doi.org/10.1016/J.IMAGE.2012.01.020

[25] Mansouri, T., Reza, M., Moghadam, S., Monshizadeh, F., & Zareravasan, A. (2023). *IoT Data Quality Issues and Potential Solutions: A Literature Review*. [Online]. Available:https://doi.org/https://doi.org/10.48550/arXiv.2103.13303

[26] Müller, M., & Sauter, D. (2023). *The more the merrier? Gene duplications in the coevolution of primate lentiviruses with their hosts*. [Online]. Available: https://doi.org/10.1016/J.COVIRO.2023.101350

[27] Muralidharan, S., Yoo, B., & Ko, H. (2020). *Designing a semantic digital twin model for IoT*. [Online]. Available: https://doi.org/10.1109/ICCE46568.2020.9043088

[28] Nguyen, T. T., Huynh, T. T., Pham, M. T., Hoang, T. D., Nguyen, T. T., & Nguyen, Q. V. H. (2023). *Validating functional redundancy with mixed generative adversarial networks*. [Online]. Available: https://doi.org/10.1016/J.KNOSYS.2023.110342

[29] Patel, K. K., Patel, S. M., & Scholar, P. G. (2016). *Internet of Things-IOT: Definition, Characteristics, Architecture, Enabling Technologies, Application & Future Challenges*. [Online]. Available: http://ijesc.org/

[30] Puschmann, D., Barnaghi, P., & Tafazolli, R. (2017). *Adaptive Clustering for Dynamic IoT Data Streams*. [Online]. Available: https://doi.org/10.1109/JIOT.2016.2618909

[31] Rajamohan, K., Rangasamy, S., Pinto, N. A., Manoj, B. E., Mukherjee, D., & Shukla, J. (2023). *IoVST: Internet of vehicles and smart traffic - Architecture, applications, and challenges*. [Online]. Available: https://doi.org/10.4018/978-1-6684-8785-3.ch015

[32] Salian, D. (2023). *Usability of Open Data. In Open-Source Horizons -*

Challenges and Opportunities for Collaboration and Innovation. [Online]. Available: https://doi.org/10.5772/intechopen.1003269

[33] Searle, T., Ibrahim, Z., Teo, J., & Dobson, R. (2021). *Estimating redundancy in clinical text*. [Online]. Available: https://doi.org/10.1016/J.JBI.2021.103938

[34] Shahid, H., Angel Vázquez, M., Reynaud, L., Parzysz, F., & Shaat, M. (2024). *Open Datasets for AI-Enabled Radio Resource Control in Non-Terrestrial Networks*. [Online]. Available: https://doi.org/2404.12813

[35] Singh, A., & Mahapatra, S. (2020). *Network-based applications of multimedia big data computing in IoT environment*. [Online]. Available: https://doi.org/10.1007/978-981-13-8759-3_17

[36] Statista. (2024). *Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2023, with forecasts from 2022 to 2030*. [Online]. Available: https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/

[37] Tahir, M. U., Naqvi, M. R., Shahzad, S. K., & Iqbal, M. W. (2020). *Resolving Data De-Duplication issues on Cloud*. [Online]. Available: https://doi.org/10.1109/ICEET48479.2020.9048214

[38] Teh, H. Y., Kempa-Liehr, A. W., & Wang, K. I. K. (2020). *Sensor data quality: a systematic review*. [Online]. Available: https://doi.org/10.1186/s40537-020-0285-1

[39] Tverdal, S., Goknil, A., Nguyen, P., Husom, E. J., Sen, S., Ruh, J., & Flamigni, F. (2023). *Edge-based Data Profiling and Repair as a Service for IoT*. [Online]. Available: https://doi.org/10.1145/3627050.3627065

[40] Vijayalakshmi, K., & Jayalakshmi, V. (2021). *Analysis on data deduplication techniques of storage of big data in cloud*. [Online]. Available: https://doi.org/10.1109/ICCMC51019.2021.9418445

[41] Vongsingthong, S., & Smanchat, S. (2015). *A Review of Data Management in Internet of Things*. [Online]. Available: https://api.semanticscholar.org/CorpusID:60676220

[42] Wang, C., Huang, X., Qiao, J., Jiang, T., Rui, L., Zhang, J., Kang, R., Feinauer, J., McGrail, K. A., Wang, P., Luo, D., Yuan, J., Wang, J., & Sun, J. (2020). *Apache IoTDB: Time-series Database for Internet of Things*. [Online]. Available: https://doi.org/10.14778/3415478.3415504

[43] Widad, E., Saida, E., & Gahi, Y. (2023). *Quality Anomaly Detection Using Predictive Techniques: An Extensive Big Data Quality Framework for Reliable*

Data Analysis. [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3317354

[44] Xia, W., Jiang, H., Feng, D., & Tian, L. (2016). *DARE: A Deduplication-Aware Resemblance Detection and Elimination Scheme for Data Reduction with Low Overheads*. [Online]. Available: https://doi.org/10.1109/TC.2015.2456015

[45] Zheng, W., Li, Y., Wu, X., & Cheng, J. (2024). *Duplicate Bug Report detection using Named Entity Recognition*. [Online]. Available: https://doi.org/10.1016/J.KNOSYS.2023.111258

[46] Zubair, N., A, N., Hebbar, K., & Simmhan, Y. (2019). *Characterizing IoT Data and its Quality for Use*. [Online]. Available: http://arxiv.org/abs/1906.10497