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

Bulletin of Electrical Engineering and Informatics • Article • 2025 •
DOI: 10.11591/eei.v14i5.10063

Md Isa, Siti Noor Basirah [a] ; Emran, Nurul A. [a] ; Harum, Norharyati [a] ;
Machap, Logenthiran [b] ; Nordin, Azlin [c]

[a] Department of Software Engineering, Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia

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.

Author keywords

Data duplication; Duplicates detection; Healthcare; Internet of things data; Visualization

Funding details

Details about financial support for research, including funding sources and grant numbers as provided in academic publications.

Funding sponsor	Funding number	Acronym
Pusat Pengurusan Penyelidikan dan Inovasi, Universiti Utara Malaysia See opportunities		
Universiti Teknikal Malaysia Melaka See opportunities by UTeM		UTeM
Malaysian Technical University Network	MTUN/2024/UTEM-FTMK/CRG/MS0008	

Funding text

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