Large-Scale Analysis of Medical Image Metadata

Elijah Chileshe
Department of Computer Science
University of Zambia
Lusaka, Zambia
elijah.chileshe@cs.unza.zm

Lighton Phiri
Department of Library and Information Science
University of Zambia
Lusaka, Zambia
lighton.phiri@unza.zm

***Abstract*—This paper offers an ongoing exploration of the systematic analysis of Medical Image Metadata encoded using the Digital Imaging and Communications in Medicine (DICOM) standard. The paper carefully looks at the organization of diverse medical image data, the complex navigation of ethical considerations involving data authorization and anonymization and the subsequent intricate processing and metadata extraction procedures. The research is part of a larger project that aims to address the pressing need for radiologists in the Republic of Zambia. With only a limited number of trained radiologists available to serve a significant population, innovative solutions are urgently required. Simultaneously, this study explores the potential of streamlined medical imaging workflows through the application of enterprise imaging techniques. The paper is dedicated to providing detailed insights into the methodologies that support the execution of large-scale medical image metadata analysis. By capturing the collection of images from multiple sources, addressing ethical concerns for patient privacy and extracting metadata from DICOM files, this ongoing study continues to provide valuable insights into the refinement of medical imaging practices and the enhancement of clinical decision-making processes.**

Keywords—DICOM, Medical Images, Medical Imaging, Metadata, Metadata Analysis

I. Introduction

Medical imaging techniques are used to show internal structures under the skin and bones, as well as to diagnose abnormalities and treat diseases [1]. Various technologies used in healthcare facilities such as X-rays, Magnetic Resonance Imaging (MRI), Computed Tomography scans (CT) and Ultrasound have generated a huge amount of medical image data. Analyzing this vast majority of data can help in the discovery of patterns and improve patient care. However, dealing with such a large amount of data presents challenges. This research endeavors to identify practical methods for conducting thorough analysis of medical images on a significant scale. The study's primary emphasis lies in developing effective strategies for processing and comprehending these images. Through the application of advanced techniques like machine learning, we aspire to extract valuable insights from these images. The outcomes of this research hold the potential to enhance medical investigations and assist healthcare professionals in making more informed decisions for their patients.

II. Related Work

Previous studies have recognised the importance of conducting large-scale analysis of medical image metadata. These studies have highlighted the significance of several vital steps that are necessary for such analyses. The organization of medical image data, ethical considerations regarding data authorization and anonymization and subsequent processing and extraction of metadata are key prerequisites for meaningful large-scale analysis. The aim of this related work section is to present a comprehensive understanding of the methods employed in previous studies and highlight their contributions to the field.

A. Healthcare Enhancements using Machine Learning and DICOM Metadata Analysis

This article explores the rising interest in leveraging Machine Learning (ML) for healthcare, driven by the potential to enhance patient care. However, the practical adoption of ML algorithms in clinical settings is impeded by a lack of necessary infrastructure, processes, and tools, despite their presence in commercial products. The authors present an automated method for identifying brain Magnetic Resonance Imaging (MRI) sequences by utilizing metadata required by the DICOM standard. This method streamlines the selection of pertinent inputs for image-related algorithms. Through testing on extensive brain MRI datasets from different institutions, the approach demonstrates notable precision and adaptability [2]. The authors propose that similar techniques could be adapted for other types of radiological imaging. The findings of this study hold relevance in the context of a comprehensive review centered around DICOM-based extensive analysis of medical image metadata.

B. DICOM Metadata Analysis for Radiology Enhancement

This paper addresses the utilization of healthcare data to enhance healthcare delivery, particularly focusing on radiology. However, challenges arise due to diverse software from various manufacturers, making data integration and patient study characterization difficult. The paper proposes utilizing DICOM metadata stored in different healthcare facilities' Picture Archiving and Communication Systems (PACS) for population characterization and patient-centered studies. The study applies this approach to chest radiographic studies across three healthcare facilities, encompassing 95,433 images from 89,980 studies involving 56,547 patients. The methodology classifies the population by age, gender, and modality, determines average studies per patient in each age group, and identifies patients with the highest studies per modality. The results highlight the value of utilizing dispersed DICOM metadata for population characterization, revealing resource usage trends and potential patient radiation over-exposure [3]. This research contributes to the comprehensive understanding of DICOM-based large-scale medical image metadata analysis.

C. Enhanced Medical Imaging Analysis with DICOM Metadata

In the search to make the most of the wealth of data generated by medical imaging studies, especially the significant Digital Imaging and Communication in Medicine (DICOM) metadata that holds key insights for healthcare understanding, it becomes essential to grasp both the advantages and challenges of organizing this metadata for further analysis. This study delves into a comprehensive secondary analysis of DICOM metadata, sourced from diverse Picture Archiving and Communication Systems (PACS) across healthcare facilities, examining both advantages and challenges. Insights obtained from the research highlight the potential of aggregating and consolidating DICOM metadata to characterize healthcare provision. While efficient mechanisms were identified for acquisition and processing without disrupting PACS performance, challenges surrounding metadata quality, stakeholder identification, computational demands, information management, individual and population exposure analyses, and resource utilization were acknowledged. The findings highlight the prospect of leveraging DICOM metadata for continuous improvement in medical imaging practices, patient-centered care strategies, translational research, and multidimensional studies [4].

By conducting large-scale analysis of medical image metadata, researchers strive to make significant contributions to the broader understanding of medical imaging practices. The insights gained from these analyses have the potential to improve clinical decision making. Ultimately, the goal is to leverage the power of large-scale data analysis to unlock valuable insights that positively impact healthcare outcomes.

III. Workflow for Large-Scale Analysis of Medical Image Metadata

A. Digitization of Medical Images

Efforts are being made to digitize medical images as part of the implementation process. This involves converting physical films or analog images into a digital format. The digitization process typically includes scanning or capturing the images using specialized equipment such as digital scanners or medical imaging devices [5]. Once the images are digitized, they can be stored, processed and analyzed more efficiently using computer systems and software. Digitalization allows for easier accessibility, sharing and manipulation of medical images, enabling large-scale analysis of metadata encoded in the DICOM standard.

B. Ethical Considerations

Ethical considerations are critical in the implementation of large-scale analysis of medical image data. This involves ensuring privacy and confidentiality by obtaining informed consent, de-identifying or anonymizing data, and implementing strict access controls and encryption techniques [6]. In alignment with these ethical imperatives, the current research seeks to curate annotated medical images as the basis of its data. In adherence to ethical mandates, the study diligently seeks authorization and endorsement from respected bodies, such as the UNZA Biomedical Research Ethics Committee (UNZABREC) and the National Health Research Authority (NHRA), prior to commencing the research. Moreover, there is an expectation that the need for patient approval concerning the utilization of their medical images, specifically chest x-rays, might be exempted by the ethical oversight authorities. This exemption hinges on the fact that the data is retrospective and can be effectively anonymized. This decision reaffirms the dedication to upholding ethical norms while carrying out meaningful metadata analysis.

C. DICOM Standard

The DICOM (Digital Imaging and Communications in Medicine) standard is a widely adopted framework in the healthcare industry for encoding, exchanging, and managing medical images and associated data [7]. It ensures interoperability and consistency by providing rules and protocols for the acquisition, storage, transmission, and display of images. DICOM standardizes the format of image data and metadata, including patient demographics, study information, imaging modalities, acquisition parameters, and clinical annotations. This standardized approach enables seamless integration and analysis of medical images across different systems and facilitates comprehensive understanding and interpretation of the images within a larger healthcare ecosystem.

a) DICOM Hierarchy

The DICOM (Digital Imaging and Communications in Medicine) standard establishes a hierarchical structure for organizing medical image data, which has implications for the extraction of metadata. The DICOM hierarchy comprises various levels, including patient, study, series, and instance [8]. At the top level, the patient level, information such as patient demographics and unique identifiers is stored. The study level contains data related to a specific medical study, including imaging modalities and study-specific details. Within a study, multiple series can exist, representing different sets of images acquired during the study. Finally, each series consists of individual instances, which are the actual images captured by the imaging equipment. Understanding the DICOM hierarchy is crucial for accurately extracting metadata, as different levels contain distinct sets of information. Metadata extraction processes need to navigate this hierarchy to retrieve relevant data from each level, ensuring comprehensive and accurate analysis of medical image metadata. Figure 1. displays the four levels of DICOM hierarchy information.

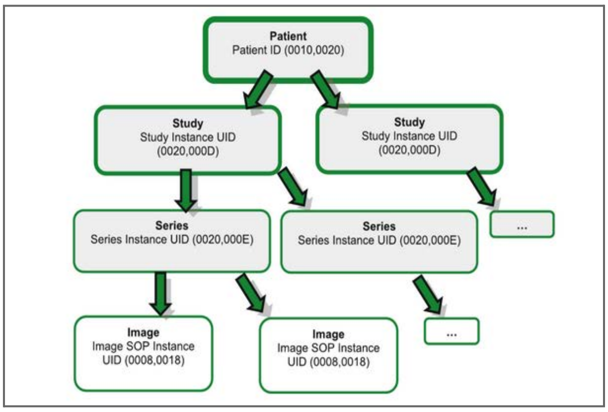
Patient
Patient ID (0010,0020)
Study
Study Instance UID (0020,000D)
Study
Study Instance UID (0020,000D)
Series
Series Instance UID (0020,000E)
Series
Series Instance UID (0020,000E)
...
Image
Image SOP Instance UID (0008,0018)
Image
Image SOP Instance UID (0008,0018)
...

Fig. 1. Four levels of DICOM information hierarchy [9]

b) DICOM Metadata

DICOM metadata encompasses the descriptive information associated with medical images. This includes a wide range of data elements that provide essential context and details about the images [10]. DICOM metadata may include information such as patient demographics (e.g., name, age, sex), imaging modality used (for example, X-ray, CT scan, MRI), imaging acquisition parameters (e.g., exposure settings, image resolution), study information (e.g., study description, study date), and clinical annotations (e.g., radiologist's observations or diagnoses). Extracting DICOM metadata involves parsing the structured data within DICOM files and retrieving specific data elements of interest. This extracted metadata provides valuable insights for large-scale analysis, enabling researchers to study patterns, trends, and associations within medical image data. Understanding the structure and content of DICOM metadata is essential for conducting meaningful analysis and utilizing the full potential of medical image datasets. Table 1 displays a table with a partial representation of DICOM metadata elements.

TABLE I. Partial Representation of DICOM Metadata Elements

Attribute Name	*Tag*	*Type*	*Attribute Description*
Modality	0008,0060	1	Device that produced the Instances in this Series
Study Description	0008,1030	3	Classification of the Study performed.
Patient Name	0010,0010	2	Patient's full name.
Patient ID	0010,0020	2	Primary identifier for the Patient.
Series Instance UID	0020,000E	1	Unique identifier of the Series.

D. Selection of Data Sources for Comprehensive Medical Image Data Collection and Analysis

The selection of appropriate data sources is a crucial aspect in the collection of medical image data for large-scale analysis. Various sources contribute to the diversity and comprehensiveness of the dataset. These sources may include hospitals, clinics, imaging centers, research institutions, and public databases. Collaboration and partnerships with healthcare providers and institutions are essential to access a wide range of imaging data. It is important to consider factors such as the availability of diverse patient populations, different imaging modalities, and a variety of medical conditions [11]. Furthermore, ensuring the data sources are reliable and representative is vital for the generalizability and validity of the analysis. Obtaining data from multiple sources increases the chances of capturing a comprehensive view of the target population, enabling more accurate and meaningful insights. Careful consideration should also be given to data sharing agreements, data ownership, and compliance with relevant regulations to ensure responsible and ethical use of the collected data.

E. Building a Robust Storage Infrastructure for Large-Scale Medical Image Metadata Analysis

When implementing machine learning models for analyzing large-scale medical image data, it is important to carefully plan and execute the process. A key element of this is setting up a robust infrastructure that can handle the storage and retrieval of the massive amounts of the medical image data effectively [12]. This infrastructure might include powerful servers, distributed computing systems, and advanced techniques for organizing and managing the data, such as dividing it into smaller portions and creating indexes for efficient retrieval. By ensuring that the storage system is reliable and scalable, researchers can efficiently manage and access the extensive collections of medical images needed for analysis.

F. Leveraging Parallel Processing for Efficient Analysis of Large-Scale Medical Image Metadata

To handle the computational demands of processing large-scale image metadata, parallel processing libraries such as Joblib [13], PySpark [14], and Dask [15] can be employed. These libraries enable the distribution of computational tasks across multiple processors or machines, allowing for faster and more efficient analysis. Joblib, for example, provides tools for parallel computing in Python, allowing tasks to be executed in parallel across multiple cores or even on remote machines. PySpark, on the other hand, is a powerful framework for distributed data processing that utilizes a cluster computing system, making it suitable for processing large datasets in a distributed manner. Dask, similar to PySpark, provides scalable parallel computing capabilities, enabling efficient processing of large-scale image data. This is particularly beneficial when dealing with large volumes of medical image data, where traditional sequential processing may be time-consuming and impractical. Parallel processing allows for the concurrent execution of tasks, effectively reducing the overall processing time and enabling more rapid analysis of the image data.

IV. Challenges with Large-Scale Analysis of Medical Image Data

A. Challenges in Managing and Storing Large Scale Medical Image Data

The large-scale analysis of medical image data presents several challenges that must be overcome to achieve meaningful results. One primary challenge is the management and storage of the immense volume of data generated by medical imaging technologies. As hospitals move towards a filmless, paperless environment, there will be a never-ending demand for digital storage space [16]. Developing efficient storage systems capable of handling the continuous production of data while ensuring data accessibility and integrity is a significant undertaking. This may involve the use of distributed storage solutions, cloud-based storage or data archiving strategies to optimize storage capacity and data retrieval performance.

B. Computational Challenges in Processing Large Scale Medical Image Datasets

Another major challenge is the computational demand associated with processing large-scale image datasets. Processing large datasets requires substantial computational resources and can be computationally expensive and time-consuming. There needs to be access to powerful computing infrastructures equipped with high-performance GPUs [17] or even applying for High Performance Computing (HPC) [18] to accelerate processing time. Implementing techniques such as model parallelism or distributed computing frameworks can help alleviate the computational burden.

C. Ensuring Data Privacy and Security in Medical Image Analysis

Ensuring data privacy and security is an ongoing challenge when working with medical image data. Ethical guidelines encourage respecting privacy, that is, the ability to retain complete control and secrecy about one's personal information [19]. Medical images contain sensitive patient information that must not be disclosed, making it crucial to implement robust data protection measures, adhere to privacy regulations and adopt secure data transfer protocols. Encryption, access controls and anonymization techniques play a vital role in safeguarding patient privacy and maintaining data security throughout the analysis pipeline.

D. Addressing Data Quality and Veracity Challenges in Large-Scale Medical Image Analysis

Data quality and veracity are critical aspects in the analysis of large-scale medical image data. The variability of medical images, stemming from factors like acquisition protocols and imaging modalities, necessitates addressing data quality challenges for accurate analysis. Standardization techniques, preprocessing steps, and robust algorithms mitigate variability and enhance data quality. Additionally, establishing standardized imaging protocols contributes to more consistent and comparable medical images. Veracity, encompassing issues such as inconsistencies, missing data, ambiguities, deception, fraud and duplication is vital in healthcare decision-making, and managing data quality is a fundamental challenge [20]. By addressing data quality and veracity concerns, healthcare professionals can ensure reliable and trustworthy information for improved analysis and decision-making.

E. Challenges in Acquiring Expertise for Medical Image Metadata Analysis

Acquiring the expertise to analyze medical image metadata poses challenges in terms of specialized training and knowledge [21]. It requires a deep understanding of medical imaging techniques, data analysis methods, and domain-specific applications. The availability of trained professionals and access to comprehensive training resources are key obstacles in ensuring a skilled workforce capable of effectively analyzing and interpreting medical image data. The complex nature of medical imaging and the continuous advancements in technology demand ongoing professional development and specialized education. Limited access to training programs and resources further hinders the acquisition of necessary skills. Collaborative efforts among educational institutions, industry, and professional organizations are crucial for developing comprehensive curricula, promoting research and innovation and improving access to training materials. Overcoming these challenges is essential to meet the growing demand for experts in medical image analysis and advance the field for improved patient care.

Conclusion

In conclusion, this paper offers an ongoing exploration of the systematic analysis of Medical Image Metadata through the utilization of the DICOM standard. It delves into the organization of varied medical image data, navigates ethical dimensions, and investigates the extraction of metadata, all in the context of addressing Zambia's scarcity of radiologists. Through this paper, valuable insights are provided to enhance medical imaging practices. Overcoming the challenges associated with large-scale analysis of medical image metadata demands collaborative multidisciplinary efforts. By fostering partnerships among researchers, healthcare professionals, data scientists, and industry stakeholders, these obstacles can be collectively overcome.

Acknowledgment

This study was conducted as part of a much larger Enterprise Medical Imaging in Zambia project. The project is funded through generous grants from Google Research, Data Science Africa and The Directorate of Research Graduate Studies at The University of Zambia. We are very appreciative of this funding.

References

[1] M. Laal, "Innovation process in medical imaging," *Procedia Soc. Behav. Sci.*, vol. 81, pp. 60–64, Jun. 2013.

[2] R. Gauriau *et al.*, "Using DICOM Metadata for Radiological Image Series Categorization: a Feasibility Study on Large Clinical Brain MRI Datasets," *J. Digit. Imaging*, vol. 33, no. 3, pp. 747–762, Jun. 2020.

[3] M. Santos, L. Bastião, A. Silva, and N. Rocha, "DICOM Metadata Analysis for Population Characterization: A Feasibility Study," *Procedia Comput. Sci.*, vol. 100, pp. 355–361, Jan. 2016.

[4] "Outcomes from Indexing Initiatives of Medical Imaging DICOM Metadata Repositories. A Secondary Analysis," *Procedia Comput. Sci.*, vol. 138, pp. 203–208, Jan. 2018.

[5] S. Venkataraman, "Digitization in Radiology: How Digital Tools are Converting Challenges into Opportunities," *CARRE4*, Dec. 13, 2020. https://medium.com/carre4/digitization-in-radiology-how-digital-tools-are-converting-challenges-into-opportunities-8b59ca2e248b (accessed Jun. 20, 2023).

[6] S. T. Padmapriya and S. Parthasarathy, "Ethical data collection for medical image analysis: A structured approach," *Asian Bioeth. Rev.*, Apr. 2023, doi: 10.1007/s41649-023-00250-9.

[7] M. Aiello, G. Esposito, G. Pagliari, P. Borrelli, V. Brancato, and M. Salvatore, "How does DICOM support big data management? Investigating its use in medical imaging community," *Insights Imaging*, vol. 12, no. 1, p. 164, Nov. 2021.

[8] O. S. Pianykh, *Digital Imaging and Communications in Medicine (DICOM)*. Springer Berlin Heidelberg.

[9] O. S. Pianykh, *Digital Imaging and Communications in Medicine (DICOM): A Practical Introduction and Survival Guide*. Springer Science & Business Media, 2009.

[10] "AAPM Reports - The Measurement, Reporting, and Management of Radiation Dose in CT." https://doi.org/10.37206/97 (accessed Jun. 20,

2023).

[11] A. Gosain and S. Sardana, “Handling class imbalance problem using oversampling techniques: A review,” in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE, Sep. 2017. doi: 10.1109/icacci.2017.8125820.

[12] M. A. Núñez-Gaona, R. Marcelín-Jiménez, J. Gutiérrez-Martínez, H. Aguirre-Meneses, and J. L. Gonzalez-Compean, “A Dependable Massive Storage Service for Medical Imaging,” *J. Digit. Imaging*, vol. 31, no. 5, pp. 628–639, Oct. 2018.

[13] “Joblib: running Python functions as pipeline jobs — joblib 1.2.0 documentation.” https://joblib.readthedocs.io/en/stable/ (accessed Jun. 19, 2023).

[14] “PySpark Overview — PySpark 3.4.0 documentation.” https://spark.apache.org/docs/latest/api/python/ (accessed Jun. 19, 2023).

[15] M. Rocklin, “Dask: Parallel Computation with Blocked algorithms and Task Scheduling,” in *Proceedings of the 14th Python in Science Conference*, SciPy, 2015. doi: 10.25080/majora-7b98e3ed-013.

[16] M. M. Frost Jr, J. C. Honeyman, and E. V. Staab, “Image archival technologies,” *Radiographics*, vol. 12, no. 2, pp. 339–343, Mar. 1992.

[17] A. Bizzego *et al.*, “Evaluating reproducibility of AI algorithms in digital pathology with DAPPER,” *PLoS Comput. Biol.*, vol. 15, no. 3, p. e1006269, Mar. 2019.

[18] J. J. Alnasir, “Fifteen quick tips for success with HPC, i.e., responsibly BASHing that Linux cluster,” *PLoS Comput. Biol.*, vol. 17, no. 8, p. e1009207, Aug. 2021.

[19] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, “Secure, privacy-preserving and federated machine learning in medical imaging,” *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, Jun. 2020.

[20] M. Aiello, C. Cavaliere, A. D’Albore, and M. Salvatore, “The Challenges of Diagnostic Imaging in the Era of Big Data,” *J. Clin. Med. Res.*, vol. 8, no. 3, Mar. 2019, doi: 10.3390/jcm8030316.

[21] B. O. Botwe, W. K. Antwi, S. Arkoh, and T. N. Akudjedu, “Radiographers’ perspectives on the emerging integration of artificial intelligence into diagnostic imaging: The Ghana study,” *J Med Radiat Sci*, vol. 68, no. 3, pp. 260–268, Sep. 2021.