Beyond Metadata Scrubbing: Production-Scale Pixel-Level and Metadata-Level DICOM Anonymisation for Healthcare Workflows

Nitin Kumar*, Alberto Andreotti*, Veysel Kocaman*, Yigit Gul*, and David Talby*
*John Snow Labs Inc., 16192 Coastal Highway, Lewes, DE 19958, USA
Email: veysel@johnsnowlabs.com

***Abstract*—DICOM (Digital Imaging and Communications in Medicine) files contain sensitive Protected Health Information (PHI) embedded in both pixel-level image data and metadata headers. Effective deidentification is crucial for enabling medical imaging research, data sharing, and compliance with privacy regulations such as HIPAA and GDPR. This paper presents a comprehensive methodology for DICOM deidentification using John Snow Labs Visual NLP, a specialized library designed for medical imaging that can understand both image content and associated text. A dual-level deidentification process is proposed, addressing both pixel-level PHI removal through text detection, extraction, and anonymization, and metadata-level PHI removal through tag reading, PHI detection, and anonymization. The methodology is evaluated on the MIDI-B dataset, demonstrating the effectiveness of Visual NLP pipelines for comprehensive DICOM deidentification. Experimental results show high accuracy across multiple validation metrics, including 100% success rates for tag retention, date shifting, and UID consistency, with overall text processing accuracy exceeding 99.9%. The approach maintains clinical utility for research purposes.**

***Index Terms*—DICOM, Medical Imaging, Deidentification, Visual NLP, Privacy-Preserving Medical Imaging, PHI Detection**

I. Introduction

Medical imaging data, stored in DICOM (Digital Imaging and Communications in Medicine) format, contains a wealth of information valuable for research, machine learning, and clinical decision support systems. However, DICOM files embed extensive Protected Health Information (PHI) in both their metadata headers and pixel-level image data. Metadata headers contain structured PHI such as patient names, dates of birth, medical record numbers, and imaging device information. Additionally, pixel-level data may contain burned-in annotations, visible text overlays, and other identifiers directly embedded in the image pixels.

Traditional deidentification approaches have primarily focused on metadata tag removal or anonymization, often overlooking pixel-level PHI that may be visible in the image itself. Effective deidentification must address both levels to ensure comprehensive privacy protection while enabling data sharing and maintaining regulatory compliance with standards such as HIPAA and GDPR. John Snow Labs Visual NLP addresses this gap by providing a comprehensive solution that handles both metadata and pixel-level deidentification through specialized natural language processing and computer vision techniques.

II. Related Work

DICOM deidentification has been extensively studied in the medical imaging research community, with various approaches addressing different aspects of privacy protection. The DICOM Basic Application Level Confidentiality Profile [1] provides standardized guidelines for removing identifying information from DICOM headers. A comprehensive review by Aryanto et al. [2] evaluated free DICOM deidentification tools used in clinical research, analyzing their functionality and effectiveness in protecting patient privacy. Their study highlighted significant variability in tool capabilities, with some failing to adequately handle private tags or free-text fields containing PHI.

Most existing solutions treat metadata and pixel-level deidentification as separate processes, requiring multiple tools and manual intervention. Recent comparative evaluations [3] on 70 DICOM files from the MIDI-B collection showed that Visual NLP achieved a precision of 1.0, recall of 0.714, and F1-score of 0.833, while Databricks Pixels achieved perfect precision (1.0) but significantly lower recall (0.285) and F1-score (0.444).

III. Methodology

The approach addresses deidentification at two critical levels—pixel-level and metadata-level—ensuring comprehensive PHI removal while preserving data utility. The entire deidentification workflow is implemented on top of Apache Spark, allowing all intermediate and final results to be represented as distributed DataFrames and processed in parallel across multiple worker nodes.

A. Pixel-Level Deidentification

Pixel-level deidentification addresses PHI that is directly embedded in image pixels, such as burned-in annotations, text overlays, and visible identifiers. The pixel-level deidentification workflow consists of multiple sequential stages (Fig. 1). First, images are extracted from the DICOM file. Users may configure frame extraction to process all frames or a representative subset.

To optimise performance, extracted images can be downscaled using a configurable scaling factor, with values between 50% and 75% providing an effective trade-off between resource usage and accuracy. Text regions are then detected

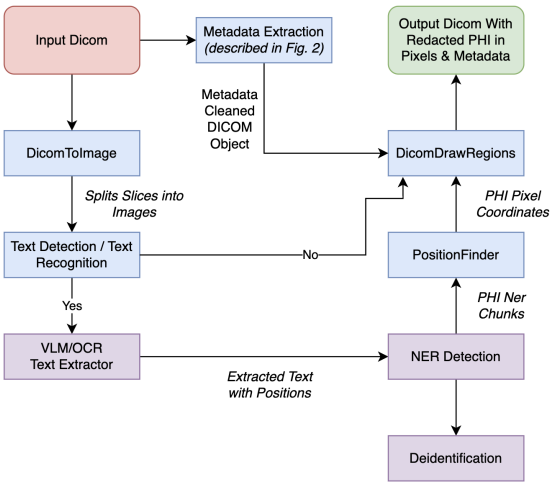

Fig. 1. Complete DICOM deidentification pipeline: dual-level approach integrating metadata extraction and pixel-level PHI detection and removal.

using advanced computer vision techniques. The framework employs proprietary text detection models based on state-of-the-art architectures such as CNN-based CRAFT [4] and transformers-based DiT [5], optimised for medical imaging contexts.

Following ROI detection, optical character recognition (OCR) is applied to extract machine-readable text from the detected ROIs. The OCR component uses deep learning–based models, including transformer-based architectures, fine-tuned on medical image text and related clinical document corpora. The extracted text is then analysed using transformer-based named entity recognition (NER) models [6], [7], fine-tuned on medical PHI datasets. Once PHI entities are detected, a matching step aligns the recognised entities with their corresponding spatial positions to generate pixel-level PHI coordinates. In the final stage, the original DICOM file is reloaded, and a mask corresponding to the union of detected PHI coordinates is applied to the pixel data. The photometric interpretation and pixel characteristics are preserved throughout this process.

B. Metadata-Level Deidentification

The deidentification of DICOM metadata targets protected health information (PHI) contained within DICOM header attributes using a hybrid approach that combines rule-based processing with selectively applied machine-learning methods. At the core of the system is a configurable strategy file that declaratively specifies the action for each DICOM tag based on its value representation (VR) and semantic category.

The system operates on both standard DICOM tags defined in the DICOM Data Dictionary and vendor-specific private tags, which are identified by odd group numbers. At the metadata-processing level, each tag follows exactly one of three mutually exclusive paths. First, tags explicitly marked with the cleanTag action in the strategy file are deterministically routed through a named entity recognition (NER) stage to detect and remove embedded PHI. Second, free-text or structurally ambiguous attributes that are not explicitly marked as cleanTag may optionally pass through a lightweight classification step. Third, all remaining tags are processed directly using predefined rule-based actions derived from their VR and the rule action mappings specified in the strategy file.

UID and patient identifier regeneration relies on a deterministic, patient-scoped, hash-based mapping strategy. This approach ensures that all DICOM objects associated with the same patient receive consistently regenerated identifiers, while maintaining uniqueness across different patients and minimising collision risk. Fig. 2 illustrates the complete metadata-level deidentification workflow, including the decision points governing tag processing and PHI detection.

IV. Results

The proposed methodology was evaluated on the MIDI-B dataset to assess the effectiveness of Visual NLP for comprehensive DICOM deidentification. The evaluation used the official MIDI-B validation script [8]. Table I presents the official MIDI-B validation results for both validation and test sets.

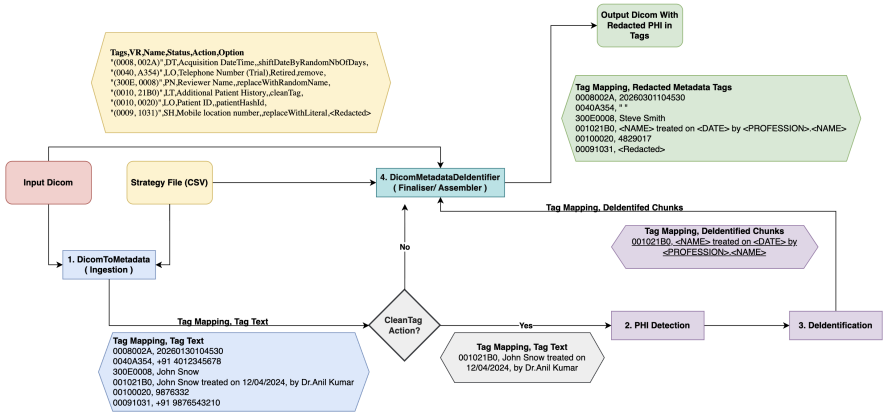

Fig. 2. Metadata-level deidentification workflow with decision point for tag cleaning: direct processing path (No) and NER-based path (Yes) for comprehensive PHI detection and removal.

TABLE I
MIDI-B VALIDATION RESULTS: ACCURACY METRICS AND TOTAL CHECKS FOR VALIDATION AND TEST SETS [9]

Action Type	Validation Set	Test Set
Tag Retained	100.0% (1,325,259)	100.0% (1,101,091)
Text Retained	99.96% (4,744,923)	100.0% (3,453,539)
Text Removed	99.88% (432,468)	99.91% (343,583)
Date Shifted	100.0% (171,930)	100.0% (139,774)
UID Changed	100.0% (280,275)	100.0% (234,418)
Pixels Retained	99.93% (29,633)	99.90% (23,886)
UID Consistent	100.0% (280,275)	100.0% (234,418)
Patient ID Consistent	100.0% (29,660)	100.0% (23,921)
Pixels Hidden	100.0% (27/27)	85.71% (30/35)

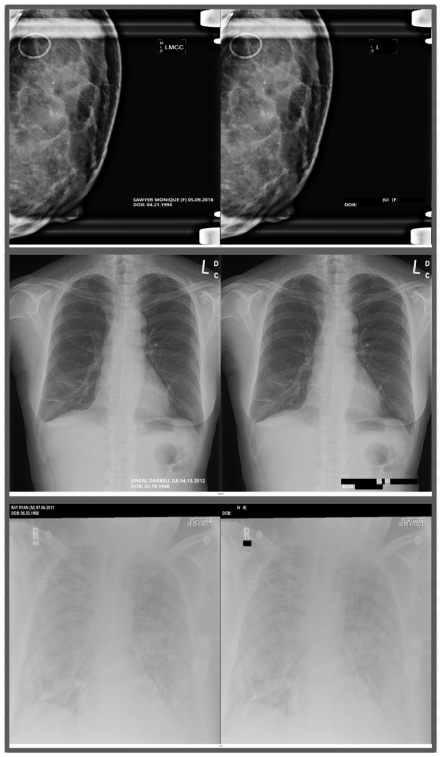

Fig. 3. Before and after comparison of pixel-level deidentification: (left) original DICOM image with visible patient identifiers, (right) deidentified image with PHI removed while preserving diagnostic quality.

According to the MIDI-B validation results, tag retention, date shifting, and UID changes achieved 100% accuracy on both sets, confirming proper temporal anonymization and referential integrity maintenance. Pixel-level PHI hiding achieved 100% accuracy on the validation set (27/27 cases) and 85.71% on the test set (30/35 cases). All missed instances were associated with short machine-generated overlay abbreviations, typically consisting of fewer than four characters. These instances resulted in low OCR confidence scores and were subsequently removed during threshold-based filtering, leading to their exclusion from the extracted text corpus. Fig. 3 demonstrates the effectiveness of pixel-level deidentification.

A. Scalability and Computational Performance

Table II presents the performance metrics on 100 DICOM files (1,000 total frames) across different cluster configurations, demonstrating near-linear scalability with a 2.78x speedup when increasing workers from 2 to 6.

TABLE II
SCALABILITY AND COMPUTATIONAL PERFORMANCE: METRICS ACROSS DIFFERENT CLUSTER CONFIGURATIONS

Workers	Time (s)	Cost (DBU/h)
2	4,582	16.66
4	2,380	22.36
6	1,647	28.06

V. DISCUSSION AND CONCLUSION

This paper presents a comprehensive methodology for DICOM deidentification using John Snow Labs Visual NLP, addressing both pixel-level and metadata-level Protected Health Information (PHI) removal. The dual-level approach ensures comprehensive privacy protection that traditional metadata-only methods may miss, particularly for burned-in annotations and text overlays embedded in image pixels. The methodology addresses key requirements of major privacy regulations, removing all 18 HIPAA identifiers through comprehensive pixel and metadata analysis [10]. The accuracy of pixel-level deidentification depends on OCR performance, which may vary with image quality, contrast, and text characteristics. An important limitation of this work is the reliance on a proprietary, closed-source library, which presents challenges for reproducibility, transparency, and scientific scrutiny. Evaluation on the MIDI-B dataset confirms the effectiveness of the approach, demonstrating high accuracy in PHI detection and removal while maintaining data utility.

REFERENCES

[1] DICOM Standards Committee, "DICOM Basic Application Level Confidentiality Profile," *DICOM Standard Part 15*, 2024.

[2] L. Aryanto, M. Oudkerk, and P. M. A. van Ooijen, "Free DICOM deidentification tools in clinical research: functioning and safety of patient privacy," *European Radiology*, vol. 25, no. 12, pp. 3685–3695, 2015.

[3] John Snow Labs, "DICOM De-identification Dataset - Pixels Platform Comparison," GitHub repository, 2024.

[4] J. Baek *et al.*, "Character Region Awareness for Text Detection," in *Proc. IEEE/CVF CVPR*, 2019, pp. 9365–9374.

[5] J. Li *et al.*, "DiT: Self-Supervised Pre-training for Document Image Transformer," in *Proc. ACM Multimedia*, 2022.

[6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[7] E. Alsentzer *et al.*, "Publicly Available Clinical BERT Embeddings," in *Proc. 2nd Clinical Natural Language Processing Workshop*, 2019, pp. 72–78.

[8] Center for Biomedical Informatics and Information Technology (CBIIT), "MIDI Validation Script," GitHub repository, 2024.

[9] MIDI Task Group, "Medical Imaging De-Identification Benchmark (MIDI-B)," MIDI Challenge Dataset, 2024.

[10] K. Benitez and B. Malin, "Evaluating re-identification risks with respect to the HIPAA privacy rule," *J. Amer. Med. Inform. Assoc.*, vol. 17, no. 2, pp. 169–177, 2010.