

A DICOM Image De-identification Algorithm in the MIDI-B Challenge

Hongzhu Jiang[1], Sihan Xie[1], and Zhiyu Wan[1,2(✉)]

[1] ShanghaiTech University, Shanghai 201210, China
wanzhy@shanghaitech.edu.cn
[2] Vanderbilt University Medical Center, Nashville, TN 37203, USA

Abstract. Image de-identification is essential for the public sharing of medical images, particularly in the widely used Digital Imaging and Communications in Medicine (DICOM) format as required by various regulations and standards, including Health Insurance Portability and Accountability Act (HIPAA) privacy rules, the DICOM PS3.15 standard, and best practices recommended by the Cancer Imaging Archive (TCIA). The Medical Image De-Identification Benchmark (MIDI-B) Challenge at the 27th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2024) was organized to evaluate rule-based DICOM image de-identification algorithms with a large dataset of clinical DICOM images. In this report, we explore the critical challenges of de-identifying DICOM images, emphasize the importance of removing personally identifiable information (PII) to protect patient privacy while ensuring the continued utility of medical data for research, diagnostics, and treatment, and provide a comprehensive overview of the standards and regulations that govern this process. Additionally, we detail the de-identification methods we applied—such as pixel masking, date shifting, date hashing, text recognition, text replacement, and text removal—to process datasets during the test phase in strict compliance with these standards. According to the final leaderboard of the MIDI-B challenge, the latest version of our solution algorithm correctly executed 99.92% of the required actions and ranked 2nd out of 10 teams that completed the challenge (from a total of 22 registered teams). Finally, we conducted a thorough analysis of the resulting statistics and discussed the limitations of current approaches and potential avenues for future improvement.

Synapse ID: 3506984

Keywords: de-identification · DICOM · pseudonymization · patient privacy · HIPAA

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
U. Baid et al. (Eds.): MICCAI 2024, LNCS 15354, pp. 126–136, 2027.
https://doi.org/10.1007/978-3-032-31709-4_13

1 Introduction

1.1 MIDI-B De-identification

In the era of digital healthcare, the processing and analysis of medical images are critical for diagnostics, treatment planning, and research. One of the key challenges in this domain is the de-identification of medical images [1, 2], which involves removing or obfuscating personally identifiable information (PII) to protect patient privacy while preserving data utility. This process is essential for ensuring compliance with privacy regulations like General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA), and for promoting the responsible sharing of medical data.

Medical images, particularly in the widely used Digital Imaging and Communications in Medicine (DICOM) format, contain not only the image data but also sensitive metadata, such as patient names and birthdates. While this metadata is invaluable for clinical workflows, it poses significant privacy risks if not properly de-identified. Therefore, effective de-identification of DICOM images is crucial for safely managing and sharing medical imaging data.

To address these challenges, the Medical Image De-Identification Benchmark (MIDI-B) challenge[1] evaluates rule-based DICOM de-identification algorithms using a diverse set of standardized clinical images with synthetic identifiers [3]. This competition aims to advance automated de-identification methods that maintain both privacy and data utility, thereby supporting the secure sharing of medical research data and fostering innovation in privacy-preserving technologies.

1.2 De-Identification Standards

In the context of the MIDI-B challenge, de-identification (deID) refers to adherence to the US HIPAA Privacy Rule safe harbor method[2], DICOM Standard PS3.15 (Attribute Confidentiality Profile)[3], and Best Practices described in the TCIA Submission Overview Page[4].

HIPAA Privacy Rule Safe Harbor. The HIPAA Privacy Rule is a U.S. regulation designed to protect the privacy of individuals' health information. By removing or obscuring identifiable information, it ensures that medical data does not reveal personal identities, thereby supporting the lawful use and sharing of data.

The Safe Harbor Provision is part of the HIPAA Privacy Rule and provides a method for ensuring that Protected Health Information (PHI) is appropriately de-identified. This involves removing or altering personal identifiers so that data cannot be traced back to specific individuals. According to HIPAA, de-identification requires the removal of 18

[1] https://www.synapse.org/Synapse:syn53065760/wiki/625274.

[2] https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html.

[3] https://dicom.nema.org/medical/dicom/current/output/html/part15.html#chapter_E.

[4] https://wiki.cancerimagingarchive.net/display/Public/Submission+and+De-identification+Overview.

types of information, including names, addresses, birthdates, Social Security numbers, medical record numbers, and insurance policy numbers. Retaining this information could lead to the identification of specific individuals.

DICOM Standard PS3.15. DICOM Standard PS3.15 (Attribute Confidentiality Profile) is a component of the DICOM standard that focuses on the confidentiality of attributes in medical imaging data [4]. This standard defines how to protect sensitive information in DICOM data to ensure patient privacy and data security [5].

It specifies requirements for safeguarding specific attributes, such as patient names and birthdates, from unauthorized access. It also outlines methods for handling and filtering sensitive attributes in DICOM images to prevent exposure during data sharing or transmission and provides de-identification procedures to remove or obscure patient identity information.

The Cancer Imaging Archive (TCIA). The TCIA Submission Overview Page outlines best practices for submitting medical imaging data to the Cancer Imaging Archive (TCIA) [6]. It provides clear guidelines to ensure data quality and privacy protection, facilitating the submission and utilization of high-quality data while meeting regulatory requirements.

In terms of de-identification, TCIA ensures that all personally identifiable information (PII) is removed from images and associated metadata, including patient names, IDs, and other demographic details. It employs standardized de-identification procedures to comply with privacy regulations such as HIPAA and thoroughly reviews data to ensure that no identifiable information remains.

These three de-identification standards effectively protect patient privacy while ensuring the research value and security of the data.

2 Methods

We implemented two categories of de-identification methods: 1) simple de-identification and 2) pseudonymization. We define simple de-identification as the removal of real patient identifiers, while pseudonymization involves replacing identifiers with a pseudonym that is unique to the individual and known within a specified context but not linked to the individual in the external world. For simple de-identification, we employed methods such as text recognition, text removal, and pixel masking. In contrast, for pseudonymization, we used techniques including date shifting, hashing, text recognition, and text replacement. Our de-identification methods are open-source and are available at https://github.com/zhywan/midi-b-challenge-2024.

2.1 Simple De-Identification

Pixels Masking. We utilized the Presidio[5], a data protection and de-identification software development kit from Microsoft to remove pixels containing sensitive information. Presidio can aid in the proper management and governance of sensitive data and offers

[5] https://microsoft.github.io/presidio/.

rapid identification and anonymization modules for processing text and images. Additionally, it consists of three major functional modules: Analyzer, Anonymizer, and Image Redactor. The Analyzer is responsible for scanning text-based data to identify sensitive information. It includes predefined recognizers and can be extended with custom recognizers. The Anonymizer is focused on desensitizing detected sensitive entities. It replaces detected PII with anonymized values using operators such as substitution, masking, and ciphering. The Image Redactor uses OCR technology to identify and desensitize sensitive information within images. Presidio allowed us to use an analyzer to detect sensitive data and identify the pixel positions of these data in the original DICOM image and cover these pixel areas with color blocks.

In our implementation, we used the Document Intelligence OCR provided by Microsoft Azure to identify text in pixel data. Additionally, we created allow lists (i.e., whitelists) and deny lists (i.e., blacklists) for this specific task, based on the validation dataset, with custom recognizers. We also fixed bugs in the original Presidio packages related to redaction colors and standardized the redaction color for consistency. We redacted the pixels for all DICOM files in the first round and then processed the metadata for each file in the second round.

An example of the final implementation of the pixel masking action is depicted in Fig. 1, where the left of the figure represents the DICOM file before the action, and the right of the figure shows the DICOM file after the action.

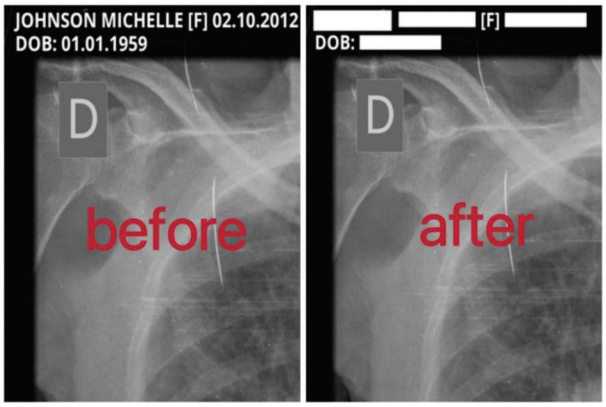

Fig. 1. An example of the final implementation of the pixel masking action

Text Removal. The text specified should be removed from tag values. To protect patient privacy, it is necessary to remove all information that may identify a patient. Depending on the action code, specific tag values can be cleared, or the portions of the content containing personally identifiable information need to be removed.

We removed the patient ID from the currently processed DICOM file by saving its value in a variable. There are 53 regular expressions designed in our algorithm to match and remove other sensitive information such as dates, phone numbers, and clinic names. Based on real-life experience, sensitive information such as abbreviations and names are identified by their relative position to prepositions such as “at” and “by” (Table 1).

Table 1. Examples of tags from which the specified text needs to be removed from their values.

Tag ID (Group, Element)	Tag Description	Action code	Action
(0008,0080)	Institution Name	X	remove
(0008,0081)	Institution Address	X	remove
(0008,0094)	Referring Physician's Telephone Numbers	X	remove
(0008,0201)	Timezone Offset From UTC	X	remove

2.2 Pseudonymization

Patient ID Replacement. It is necessary to ensure that patient ID are replaced and are consistent with patient ID mapping. The patient ID mapping file contains two columns with the header names "id_old" and "id_new". The "id_old" column contains the patient ID before de-identification and "id_new" is the patient ID after de-identification. Then we used a lookup function to replace both the patient ID and patient name with the new identifiers (i.e., "id_new"). Table 2 shows tags that need to be updated according to the TCIA standards. Table 3 shows an example of a patient ID mapping file.

Table 2. Tags that need to be updated according to the TCIA standards.

Tag ID (Group, Element)	Tag Description	Action
(0010, 0010)	Patient Name	LOOKUP(PatientID, ptid)
(0010, 0020)	Patient ID	LOOKUP(this, ptid)

Table 3. An example of the patient ID mapping file.

id_old	id_new
1059030585	0000001
1065842606	0000002
1097215536	0000003

UID Replacement. DICOM makes extensive use of universal identifiers (UID) that could be used to identify a subject. We process the UID in the DICOM file through a hashing function. First, we determine a fixed-format prefix, e.g., uid_root = '1.2.397.4.5.{patient id_new}.8.117.'. Then, we hash the UID to generate a unique hash value and keep the first 19 digits only. The final hash UID is generated by splicing the string *uid_root* in front of it as a prefix, where the "patient id_new" in the prefix is a fixed-length numeric string. This method ensures that the *id_new* of UID for different patients

are different, thus significantly reducing the possibility of *collision* (i.e., two original values are hashed into the same value) in the hashing process.

The mapping file contains two columns with the header names "id_old" (the UID before de-identification) and "id_new" (the UID after de-identification). The file provides a mapping of all UIDs pseudonymized during the participants' de-identification process to indicate the old UID and what it was transformed to. Table 4 shows an example of a UID mapping file. Table 5 shows examples of tags that the UID need to be updated according to the TCIA standards.

Table 4. An example of the UID mapping file

id_old	id_new
2.2.374.1.2.1964017.6.944.2103992807195684018	1.2.397.4.5.0000001.8.117.6881276565361048595
2.2.198.1.2.3201303.1.133.1364097273910791617	1.2.397.4.5.0000160.8.117.4832656936496443956
2.1.240.0.0.7462603.1.346.1406313923998980205	1.2.397.4.5.0000285.8.117.4677849556384631713

Table 5. An example of tags that the UID need to be updated according to the TCIA standards.

Tag ID (Group, Element)	TAG Description	Action
(0008, 0014)	Instance Creator UID	hashuid(@UIDROOT,this)
(0008, 0018)	SOP Instance UID	hashuid(@UIDROOT,this)
(0008, 1155)	Referenced SOP Instance UID	hashuid(@UIDROOT,this)
(0008, 3010)	Irradiation Event UID	hashuid(@UIDROOT,this)
(0020, 000D)	Study Instance UID	hashuid(@UIDROOT,this)

Date Shifting. To shift the date using the specified shift value, we generated a set of 322 different random numbers ranging from 1 to 365 to ensure that each patient's date offset was unique. We then subtracted the given number of offset days from the original date. This approach effectively de-identifies the date information by altering it in a way that prevents future date errors or anomalies.

There are four main formats for date values: 1) YYYYMMDD 2) YYYYMMDDHHMMSS 3) YYYYMMDDHHMMSS.FF 4) Unix timestamp. If the value includes time, we keep the time unchanged and only modify the date portion (Table 6).

Table 6. Example of tags that require the date to be shifted by a specified value.

Tag ID (Group, Element)	Tag Description	Action
(0008,0012)	Instance Creation Date	incrementdate(this,@DATEINC)
(0008,0020)	Study Date	incrementdate(this,@DATEINC)
(0008,0021)	Series Date	incrementdate(this,@DATEINC)
(0008,0022)	Acquisition Date	incrementdate(this,@DATEINC)
(0008,0023)	Content Date	incrementdate(this,@DATEINC)

3 Results

3.1 Dataset

The dataset we used during the validation phase contains 29,660 DICOM images with synthetic PHI and PII from 322 patients. Each patient folder contains one or more studies. Each study may have one or more series, and each series has one or more instances.

There are 813 tags in the Data Set with non-null values, except for meta information. According to the Best Practices described in the TCIA Submission Overview Page, 44 tags need to be removed, 32 tags need to keep the original values unchanged, 2 tags need a lookup function, 17 tags need to be shifted by a hash function, and 10 tags need additional processing. Figure 2 shows the distribution of these action types.

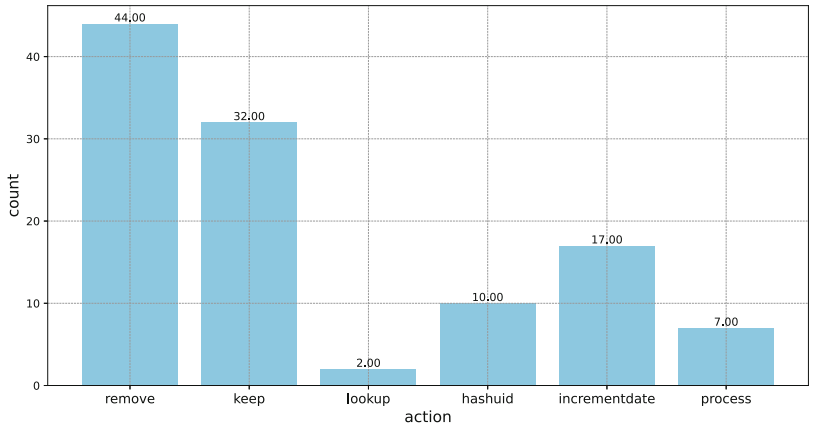

Fig. 2. Distribution of action types in the validation dataset.

3.2 Submission Results

According to the final leaderboard of the MIDI-B challenge, the latest version of our solution algorithm correctly executed 99.92% of the required actions and ranked 2nd out of 10 teams that completed the challenge (from a total of 22 registered teams), as shown in the leaderboard. The total running time is about 41 h and 41 min, and the pixel

processing takes 99% of the time (i.e., 41 h and 11 min). The running time for each DICOM image file is 5.06 s. The winning team correctly executed 99.93% of required actions. In other words, the winning team correctly executed around 58 more required actions than our team did.

Action Report. The action report provides counts of both passed and failed values for each action, as shown in Table 7.

Table 7. Action report detailing our solution's performance on the test dataset.

	Action	Fail	Pass	Total
0	<date_shifted>	2	2304	2306
1	<patid_consistent>	0	429	429
2	<pixels_hidden>	0	15	15
3	<pixels_retained>	0	29471	29471
4	<tag_retained>	8	121682	121690
5	<text_notnull>	12	85311	85323
6	<text_removed>	326	5490	5816
7	<text_retained>	131	254818	254949
8	<uid_changed>	4	40629	40633
9	<uid_consistent	4	40629	40633
Total		487	580778	581265

Category Report. The category report provides counts of both passed and failed values for each answer category, as shown in Table 8.

Table 8. Category report detailing our solution's performance on the test dataset.

	Category	Subcategory	Fail	Pass
0	dicom	DICOM-IOD-1	20	170626
1	dicom	DICOM-IOD-2	0	36367
2	dicom	DICOM-P15-BASIC-C	0	429
3	dicom	DICOM-P15-BASIC-U	4	40629
4	hipaa	HIPAA-A	0	676
5	hipaa	HIPAA-B	2	30
6	hipaa	HIPAA-C	2	2816
7	hipaa	HIPAA-D	0	20
8	hipaa	HIPAA-G	0	149
9	hipaa	HIPAA-H	1	657

(*continued*)

Table 8. (*continued*)

	Category	Subcategory	Fail	Pass
10	hipaa	HIPAA-R	4	41574
11	tcia	TCIA-P15-BASIC-D	0	198
12	tcia	TCIA-P15-BASIC-X	0	13
13	tcia	TCIA-P15-BASIC-X/Z/D	0	147
14	tcia	TCIA-P15-BASIC-Z	0	482
15	tcia	TCIA-P15-BASIC-Z/D	0	11
16	tcia	TCIA-P15-DESC-C	8	2340
17	tcia	TCIA-P15-DEV-C	0	17
18	tcia	TCIA-P15-DEV-K	0	179
19	tcia	TCIA-P15-MOD-C	0	2565
20	tcia	TCIA-P15-PAT-K	0	1113
21	tcia	TCIA-P15-PIX-K	0	29471
22	tcia	TCIA-PTKB-K	117	31670
23	tcia	TCIA-PTKB-X	257	901
24	tcia	TCIA-REV	72	217698
Total		487	580778	581265

Scoring Report. The scoring report takes the categories from the previous report and assigns them to scoring categories, as shown in Table 9. The score is calculated in terms of the ratio of fail/pass actions.

Table 9. Scoring report for our solution on the test dataset.

Category	Fail	Pass	Total	Score
All	487	580,778	581,265	99.92%

4 Discussion

In conclusion, the evaluation results of our solution algorithm demonstrate the effectiveness of DICOM image de-identification in protecting patient privacy while preserving the data's utility in an automated manner. However, both our solution and the challenge have limitations that need to be addressed and improved upon in the future.

4.1 Limitations of Our Algorithm

Our proposed method performs well on specific datasets, but it has certain limitations in terms of generalizability. This means that when being applied to other datasets, it may not achieve the same results and might even fail to meet the expected standards [7]. Therefore, the applicability of the method could be restricted, requiring further optimization and adjustment to ensure its effectiveness across a broader range of datasets [8].

While we have achieved some degree of effectiveness in removing sensitive information from text, there is still room for improvement in terms of accuracy. The current algorithm may either miss certain crucial sensitive data or incorrectly remove non-sensitive data. This can result in suboptimal outcomes that do not fully meet the expected security standards. Therefore, further refinement of our techniques for removing sensitive information, with a focus on enhancing accuracy, remains a key area of focus to ensure the reliability and precision of the results.

4.2 Limitations of the MIDI-B Challenge

Dataset. The dataset selected for the MIDI-B De-identification challenge is both novel and unique compared to other medical imaging datasets, such as those found in TCIA. It provides robust support for addressing critical issues in the de-identification of medical data, particularly in terms of patient privacy protection and data sharing [9, 10].

We propose the following considerations to enhance its utility: The dataset should include detailed metadata and comprehensive documentation to enable researchers to understand and replicate the de-identification process. Additionally, the dataset's size and diversity should be increased to better support the training and validation of machine learning models, facilitating broader application in various de-identification scenarios.

Privacy Protection Requirement. The MIDI-B challenge currently requires only de-identification and pseudonymization, without incorporating anonymization considering an adversarial model. It is recommended that anonymization be included in the privacy protection requirements, particularly when handling sensitive data. Furthermore, the competition can introduce a re-identification attack model [11], considering the potential for attackers to re-identify de-identified and pseudonymized data through external data sources or cross-referencing.

References

1. Chevrier, R., Foufi, V., Gaudet-Blavignac, C., Robert, A., Lovis, C.: Use and understanding of anonymization and de-identification in the biomedical literature: scoping review. J. Med. Internet Res. **21**, e13484 (2019)
2. Moore, S.M., et al.: De-identification of medical images with retention of scientific research value. Radiographics **35**, 727–735 (2012)
3. Kushida, C.A., et al.: Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. Med. Care **50**, S82-101 (2012)
4. DICOM. In PS3.15 2016a - Security and System Management Profiles, NEMA, Rosslyn, VA (2016)

5. Tanabe, K.: Pareto's 80/20 rule and the Gaussian distribution. Physica A **510**, 635–640 (2018)
6. Clark, K., et al.: The cancer imaging archive (TCIA): maintaining and operating a public information repository. J. Digit. Imaging **26**, 1045–1057 (2013)
7. Robinson, J.D.: Beyond the DICOM header: additional issues in deidentification. Am. J. Roentgenol. **203**, W658–W664 (2014)
8. Bennett, W., Smith, K., Jarosz, Q., Nolan, T., Bosch, W.: Reengineering workflow for curation of DICOM datasets. J. Digit. Imaging **31**, 783–791 (2018)
9. Rutherford, M., et al.: Dataset from medical imaging de-identification initiative (MIDI). Cancer Imaging Arch. (2021). https://doi.org/10.7937/s17z-r072
10. Rutherford, M., et al.: A DICOM dataset for evaluation of medical image de-identification. Sci. Data **8**, 183 (2021)
11. Wan, Z., et al.: A game theoretic framework for analyzing re-identification risk. PLoS ONE **10**, e0120592 (2015)