OBIA: An Open Biomedical Imaging Archive

Enhui Jin[1,3,4,#,a], Dongli Zhao[2,#,b], Gangao Wu[1,3,4,#,c], Junwei Zhu[1,3,d], Zhonghuang Wang[1,3,4,e], Zhiyao Wei[2,f], Sisi Zhang[1,3,g], Anke Wang[1,3,h], Bixia Tang[1,3,i], Xu Chen[1,3,j], Yanling Sun[1,3,*,k], Zhe Zhang[5,*,l], Wenming Zhao[1,3,4,*,m], Yuanguang Meng[2,5,*,n]

[1] *National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China*

[2] *Chinese People's Liberation Army (PLA) Medical School, Beijing 100853, China*

[3] *CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China*

[4] *University of Chinese Academy of Sciences, Beijing 100049, China*

[5] *Department of Obstetrics and Gynecology, Seventh Medical Center of Chinese PLA General Hospital, Beijing 100010 China*

[#]Equal contribution.

[*]Corresponding authors.

E-mail: meng6512@vip.sina.com (Meng Y), zhaowm@big.ac.cn (Zhao W), tj.zhe.zhang@gmail.com(Zhang Z), sunyanling@big.ac.cn (Sun Y).

Running title: *Jin E et al / OBIA: An Open Biomedical Imaging Archive*

[a]ORCID: 0009-0000-9916-9508.

[b]ORCID: 0000-0003-2316-5927.

[c]ORCID: 0000-0002-3036-5997.

[d]ORCID: 0000-0003-4689-3513.

[e]ORCID: 0000-0002-1268-883X.

[f]ORCID: 0000-0003-4156-3267.

[g]ORCID: 0000-0002-3852-4796.

[h]ORCID: 0000-0002-2565-2334.

[i]ORCID: 0000-0002-9357-4411.

[j]ORCID: 0000-0001-6102-1751.

[k]ORCID: 0000-0002-3175-3625.

[l]ORCID: 0009-0007-7446-7278.

[m]ORCID: 0000-0002-4396-8287.

[n]ORCID: 0000-0002-4957-3999.

Total word counts (from “Introduction” to “Discussion”): 2436

Total references: 32

Total figures: 3

Total tables: 1

Total supplementary figures: 0

Total supplementary tables: 2

Total supplementary files: 2

Total letters in article title: 34

Total letters in running title: 42

Total keywords: 5

Total word counts in Abstract: 174

Abstract

With the development of artificial intelligence (AI) technologies, biomedical imaging data play important role in scientific research and clinical application, but the available resources are limited. Here we present Open Biomedical Imaging Archive (OBIA), a repository for archiving biomedical imaging data and related clinical data. OBIA adopts five data objects (Collection, Individual, Study, Series, and Image) for data organization, accepts the submission of biomedical images of multiple modalities, organs, and diseases. In order to protect data privacy, OBIA has formulated a unified de-identification and quality control process. In addition, OBIA provides friendly and intuitive web interfaces for data submission, browsing, and retrieval. In particular, OBIA supports both metadata retrieval and image retrieval. As of September 2023, OBIA has housed data for a total of 937 individuals, 4136 studies, 24,701 series, and 1,938,309 images covering 9 modalities and 30 anatomical sites. Collectively, OBIA provides a reliable platform for biomedical imaging data management and offers free open access to all publicly available data to support research activities throughout the world. OBIA can be accessed at https://ngdc.cncb.ac.cn/obia.

KEYWORDS: Open Biomedical Imaging Archive; Database; Biomedical imaging; De-identification; Quality control

Introduction

The introduction of advanced imaging technologies has greatly facilitated the development of non-invasive diagnoses. Currently, biomedical images can show clear delineation of internal structure (anatomy), morphology, and physiological functions from the molecular scale to cellular, organ, tissue, lesion, and to the whole organism scale [1], which offers an important source of evidence for early detection of pathological changes, disease diagnosis, and treatment [2,3]. The imaging data generated during the patient visit have formed a huge accumulation. However, incomplete sharing systems make it difficult for researchers and clinicians to collaborate on using these images to obtain significant insights into health and disease [4]. Furthermore, the need for rapid diagnosis promotes the application of artificial intelligence (AI) in biomedical imaging, and the development of reliable and robust AI algorithms requires sufficiently large and representative image datasets [5,6]. Thus, the formation of high-quality biomedical imaging data storage plays an important role in promoting scientific discoveries and improving diagnostic accuracy.

The National Institutes of Health (NIH) has funded several neuroimaging and brain imaging repositories, including Image and Data Archive (IDA) [7], NITRC Image Repository (NITRC-IR) [8], Federal Interagency Traumatic Brain Injury Research (FITBIR) [9], OpenNeuro [10], and National Institute of Mental Health Data Archive (NDA) [11]. The imaging data hosted by IDA, NITRC-IR, OpenNeuro, and NDA involve various diseases, and the data access policy is determined by its owners, while FITBIR is dedicated to storing traumatic brain injury (TBI) relevant data, all of which are controlled access. IDA, OpenNeuro, and NDA support data de-identification and quality control, and NITRC-IR provides no specific quantitative quality control processing. In terms of metadata, IDA provides more clinical data, while OpenNeuro and NDA use the Brain Imaging Data Structure (BIDS) to organize metadata. In particular, NDA use global unique identifier (GUID) to identify participants, making it possible to match participants across labs and research data

repositories. In cancer imaging, the National Cancer Institute (NCI) has funded The Cancer Imaging Archive (TCIA) [12] and Imaging Data Commons (IDC) [13]. TCIA provides cancer imaging of multiple organs and related clinical data for researchers who download image data from local resources, and IDC aims to make public TCIA collections available within the Cancer Research Data Commons (CRDC) cloud environment to support cloud-based cancer imaging research. OPTIMAM Mammography Image Database (OMI-DB) [14] (managed by Cancer Research United Kingdom) and Breast Cancer Digital Repository (BCDR) [15] (supported by University of Porto, Portugal) provide annotated breast cancer images and clinical details. To fight COVID-19, the National Institute of Biomedical Imaging and Bioengineering (NIBIB) has funded Medical Imaging and Data Resource Center (MIDRC) [16], an open-access platform for COVID-19-related medical images and associated data. In addition, some universities or institutions also provide some open source datasets, such as Open Access Series of Imaging Studies (OASIS) [17], EchoNet-Dynamic [18], Cardiac Acquisitions for Multi-structure Ultrasound Segmentation (CAMUS) project [19], Chest X-ray [20] and Structured Analysis of the Retina (STARE) [21]. Due to privacy concerns and policy restrictions, there are no publicly accessible biomedical imaging databases in China currently.

To address this issue, we built the Open Biomedical Imaging Archive (OBIA; https://ngdc.cncb.ac.cn/obia), a repository for archiving biomedical imaging data and related clinical data. As a core database resource in the National Genomics Data Center (NGDC) [22], part of the China National Center for Bioinformation (CNCB; https://ngdc.cncb.ac.cn/), OBIA accepts submissions of images from all over the world and provides free open access to all publicly available data to support global research activities. OBIA supports the de-identification, management, and quality control of imaging data, provides data services such as browsing, retrieval, and downloading to promote the reuse of existing imaging data and clinical data.

Implementation

OBIA was implemented using Spring Boot (a framework easy to create standalone Java applications; https://spring.io/projects/spring-boot) as the back-end framework. The front-end user interfaces were developed using Vue.js (an approachable, performant and versatile framework for building web user interfaces; https://vuejs.org/) and Element UI (a Vue 2.0 based component library for developers, designers and product managers; https://element.eleme.cn/). The charts on the web pages were built using ECharts (an open source JavaScript visualization library; https://echarts.apache.org). All metadata information was stored in MySQL (a free and popular relational database management system; https://www.mysql.com/).

Database content and usage

Data model

Imaging data in OBIA are organized into five objects, i.e., Collection, Individual, Study, Series, and Image (**Figure 1**). "Collections", bearing an accession number prefixed with "OBIA", provides an overall description for a complete submission. "Individual", possessing an accession number prefixed with "I", defines the characteristics of a human or non-human organism receiving, or registered to receive, healthcare services. "Study", adopting an accession number prefixed with "S", contains descriptive information about radiological examinations performed on an individual. A study may be divided into one or more "Series" according to different logics, such as body part or orientation. "Image" describes the pixel data of a single digital imaging and communications in medicine (DICOM) file, and an image is related to a single series within a single study. Based on these standardized data objects, OBIA connects the image structure defined by the DICOM standard with real research projects, realizing data sharing and exchange. Besides, each collection in OBIA is linked to BioProject [23] (https://ngdc.cncb.ac.cn/bioproject/) to provide

descriptive metadata about the research project. And the individual in OBIA can be associated with GSA-Human [24] (https://ngdc.cncb.ac.cn/gsa-human/) by individual accession number, if available, which links imaging data with genomic data for researchers to perform multi-omics analysis.

De-identification and quality control

Images may contain protected health information (PHI) that requires proper handling to minimize the risk of patient privacy disclosure. In order to retain as much valuable scientific information as possible while removing PHI, OBIA provides a unified de-identification and quality control mechanism (**Figure 2**) based on the rules for DICOM image de-identification defined by DICOM PS 3.15 Appendix E: Attribute Confidentiality Profile (https://www.dicomstandard.org/). The key elements and rules we adopted includes: 1) clean pixel data, 2) clean descriptors, 3) retain longitudinal temporal information modified dates, 4) retain patient characteristics, 5) retain device identity, 6) retain safe private tags.

OBIA uses Radiological Society of North America (RSNA) MIRC clinical trial processor (CTP) (https://mircwiki.rsna.org/index.php?title=MIRC_CTP) for most of the de-identification work. We constructed a CTP pipeline and developed a common base de-identification script to remove or blank certain standard tags that either contain or potentially contain PHI. This script also maps local patient IDs to OBIA individual accessions. As for private tags, since they are specific to the vendor and the scanner and can contain almost anything, we use PyDicom (https://pypi.org/project/pydicom/) to retain those attributes that are numeric only. Some research determined private element definitions by reading manufacturers' DICOM conformance statements [25]. In our practice, however, the wide variety of image sources made this work too time-consuming, and some conformance statements were not available. In addition to metadata elements, ultrasounds or those screen captures, usually add some "burned in" annotations in the pixels data to

interpret the images, which may also contain PHI. We provide a filter stage to identify these images.

Once the de-identification process is complete, OBIA will run a quality control procedure. Some problematic images will be isolated, such as images with a blank header or missing Patient ID, corrupted images, other patient images mixed in, etc. Submitters can provide relevant information to repair an image or discard it entirely. Duplicate images will be removed, with only one retained. Then we use TagSniffer (https://github.com/stlstevemoore/dicom-tag-sniffer) to generate a report for all images. All DICOM elements in the report will be carefully reviewed to ensure that they are free of PHI and that certain values (e.g., Patient ID, Study Date) have been modified as expected. In addition, we perform a visual inspection of each image series to ensure that no PHI is contained in the pixel values as well the image is viewable and uncorrupted.

Data browse and retrieval

OBIA provides user-friendly web interfaces for data query and browsing. Users can browse the data of interest by specifying non-image information (e.g., age, gender, disease) and/or the imaging data extracted from the DICOM header (e.g., modality, anatomical site). Users can also search for data by entering the accession number. OBIA allows users to browse the basic information of collection (title, description, keywords, submitter, data accessibility, etc.), individual, study (study description, study age, study date, etc.), series (modality, anatomical site, series description, etc.) and view thumbnails of image.

OBIA also provides image retrieval functionality that aims to find images similar to the query. Content-based image retrieval methods based on deep learning are gradually replacing traditional image retrieval methods that mainly rely on scale-invariant feature transform (SIFT) [26], local binary patterns (LBP) [27], histogram of oriented gradient (HOG) [28], and other methods, as deep neural networks are able to extract superior features compared to traditional machine

learning methods. Qayyum et al. [29] proposed a two-stage medical image retrieval model trained on multimodal datasets. After image classification training, they used three fully connected layers at the end of the network to extract features for retrieval. On the other hand, Fang et al. [30] designed an end-to-end model combining a triplet network with an attention mechanism in a residual network, trained on the MIMIC-CXR and Fundus-iSee datasets, and achieved promising retrieval performance. While these methods outperform traditional image retrieval approaches, they either involve a two-stage process or focus solely on single-modal data.

To construct a multi-modal medical image retrieval model, we utilized TCIA multi-modal cancer data. Our model employed EfficientNet [31] as the feature extractor and train the model with a triplet network with an attention module to compress the image into a discrete hash value (**Figure 3**). Faiss is a high-performance similarity search library developed by Facebook AI Research, which is mainly used to solve the nearest neighbor search problem in large-scale datasets, particularly in the field of deep learning. The hash codes of the images are stored in the Faiss library to construct the image retrieval database. We compute image similarity using the hamming distance and return the most similar images to the query. As a result, our model achieved a mean average precision (MAP) value that surpassed the performance of existing advanced image retrieval models on TCIA dataset. Furthermore, we utilized this model to construct an image retrieval system within OBIA.

Data access and download

OBIA states that the data access policy is set by the owners. There are two different types of data accessibility: open access and controlled access. Open access means that all data is public for global researchers if released, whereas controlled access means that data can be downloadable only after being authorized by the submitter. OBIA supports online data requests and reviews. Before applying, users need to register and

log into OBIA via the Beijing Institute of Genomics (BIG) Single Sign-On (SSO; https://ngdc.cncb.ac.cn/sso/) system. Applicants shall provide their own basic information, specify the scope of data usage, and promise not to try to restore the privacy information in the data. Only if the data owner approves the request, a download link will be provided to the applicant.

Data submission

OBIA accepts submissions of biomedical imaging data in DICOM format from clinical or specific research projects. To create a submission, users need to log in, fill in the basic information of the collection, and email the necessary clinical information. Image data will be transferred to OBIA offline after de-identification. Users can use the de-identification process recommended by OBIA or use their own methods to de-identify the image. All arriving images will undergo the quality control process. OBIA will assign a unique accession number to each collection, individual, and study and arrange images into a standard organization. In order to ensure the security of submitted data, a copy of backup data is stored on a physically separate disk. Finally, the metadata from the image file headers will be extracted and stored in the database to support the query of the data.

Data statistics

As of September 2023, OBIA has housed a total of 937 individuals, 4136 studies, 24,701 series, and 1,938,309 images covering 9 modalities and 30 anatomical sites. Representative imaging modalities are computed tomography (CT), magnetic resonance (MR), and digital radiography (DX) (Table S1). Anatomical sites include abdomen, breast, chest, head, liver, pelvis, and so on. (Table S2). The first batch of collections submitted to OBIA came from Chinese People's Liberation Army (PLA) General Hospital, including imaging data of the three major gynecological tumors: endometrial cancer, ovarian cancer, and cervical cancer. The data were divided into

four collections, and **Table 1** shows the number of individuals, studies, series, and images in OBIA by these four collections. In addition, we collected associated clinical metadata, such as demographic data, medical history, family history, diagnosis, pathological types, and treatment methods.

Discussion

OBIA is committed to archiving biomedical imaging data. Different from existing related databases, OBIA is characterized by publishing imaging data and clinical information from various diseases and a wide range of imaging modalities in a common DICOM format. Algorithm developers can use existing tools to convert images into any format they need, and clinicians and researchers can also combine clinical information and images for secondary analysis. To ensure the security of human data, OBIA has also formulated a unified de-identification and quality control process. As part of the NGDC resource, OBIA also links to BioProject and GSA-Human, enabling seamless integration of imaging and genomic data to facilitate multi-omics analysis. To sum up, OBIA provides a reliable platform for research-relevant clinical imaging data sharing among investigators from different institutions and also fills the gap in China's biomedical imaging database.

In the future, we will continue to upgrade the infrastructure of OBIA and increase security protection measures to realize long-term secure storage, management and access to a large number of images. At the same time, we will collect more types of biomedical imaging data and gradually increase the corresponding genomic data to expand our data resource. To facilitate data submission and ensure privacy security, we will continue to optimize the image de-identification process, trying to remove PHI from image pixels with machine learning-based optical character recognition (OCR) [32] method. Meanwhile, quality control and background automatic review processes will be improved to speed up data submission. In compliance with

applicable regulations and ethical norms, OBIA's goal is to preserve as much effective image metadata as possible to provide researchers with high-quality imaging data. In addition, we will develop more intuitive and interactive web interfaces according to users' needs, increase database functions, and integrate related online tools to help analyze biomedical images. Finally, we call for collaborators to work together to build OBIA, break down data silos, catalyze new biomedical discoveries, and provide the possibility to create personalized treatments.

Ethical statement

The collection of human imaging data was approved by the Local Ethical Committees in the First Medical Center of the Chinese PLA General Hospital (Approval No. S2022-403). The written informed consent was obtained from the participating subjects.

Data availability

OBIA is publicly available at https://ngdc.cncb.ac.cn/obia.

CRediT author statement

Enhui Jin: Investigation, Methodology, Software, Writing – Original Draft. **Dongli Zhao:** Resources. **Gangao Wu:** Investigation, Methodology, Software, Writing – Original Draft. **Junwei Zhu:** Software. **Zhonghuang Wang:** Methodology. **Zhiyao Wei:** Resources. **Sisi Zhang:** Methodology. **Anke Wang:** Software, Writing – Review & Editing. **Bixia Tang:** Resources. **Xu Chen:** Resources. **Yanling Sun:** Investigation, Methodology, Writing – Review & Editing, Project administration. **Zhe Zhang:** Investigation, Methodology, Writing – Review & Editing. **Wenming Zhao:**

Conceptualization, Methodology, Writing – Review & Editing, Supervision, Funding acquisition. **Yuanguang Meng:** Conceptualization, Methodology, Writing – Review & Editing. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work was supported by grants from the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB38050300); Genomics Data Center Operation and Maintenance of Chinese Academy of Sciences (Grant No. CAS-WX2022SDC-XK05); the Key Technology Talent Program of the Chinese Academy of Sciences (awarded to YW); the Key Technology Talent Program of the Chinese Academy of Sciences (awarded to BT).

ORCID

0000-0002-4957-3999 (Yuanguang Meng)
0000-0002-4396-8287 (Wenming Zhao)
0009-0007-7446-7278 (Zhe Zhang)
0000-0002-3175-3625 (Yanling Sun)
0009-0000-9916-9508 (Enhui Jin)
0000-0003-2316-5927 (Dongli Zhao)
0000-0002-3036-5997 (Gangao Wu)
0000-0003-4689-3513 (Junwei Zhu)
0000-0002-1268-883X (Zhonghuang Wang)

0000-0003-4156-3267 (Zhiyao Wei)

0000-0002-3852-4796 (Sisi Zhang)

0000-0002-2565-2334 (Anke Wang)

0000-0002-9357-4411 (Bixia Tang)

0000-0001-6102-1751 (Xu Chen)

References

[1] Wallyn J, Anton N, Akram S, Vandamme TF. Biomedical imaging: principles, technologies, clinical aspects, contrast agents, limitations and future trends in nanomedicines. Pharm Res 2019;36:78.

[2] Li MK, Guo R, Zhang K, Lin ZC, Yang F, Xu SH, et al. Machine learning in electromagnetics with applications to biomedical imaging: a review. IEEE Antennas Propag Mag 2021;63:39-51.

[3] Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical image analysis using convolutional neural networks: a review. J Med Syst 2018;42:226.

[4] Moody A. Perspective: the big picture. Nature 2013;502:S95.

[5] Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing medical imaging data for machine learning. Radiology 2020;295:4−15.

[6] Sun C, Shrivastava A, Singh S, Gupta A. Revisiting unreasonable effectiveness of data in deep learning era. IEEE Int Conf Comput Vis 2017:843−52.

[7] Crawford KL, Neu SC, Toga AW. The image and data archive at the laboratory of neuro imaging. Neuroimage 2016;124:1080−3.

[8] Kennedy DN, Haselgrove C, Riehl J, Preuss N, Buccigrossi R. The NITRC image repository. Neuroimage 2016;124:1069−73.

[9] Thompson HJ, Vavilala MS, Rivara FP. Common data elements and federal interagency traumatic brain injury research informatics system for TBI research. Annu Rev Nurs Res 2015;33:1−11.

[10] Markiewicz CJ, Gorgolewski KJ, Feingold F, Blair R, Halchenko YO, Miller E, et al. The OpenNeuro resource for sharing of neuroscience data. Elife 2021;10:e71774.

[11] Lee SM, Majumder MA. National institutes of mental health data archive: privacy, consent, and diversity considerations and options for improvement. AJOB Neurosci 2022;13:3−9.

[12] Prior FW, Clark K, Commean P, Freymann J, Jaffe C, Kirby J, et al. TCIA: an information resource to enable open science. Annu Int Conf IEEE Eng Med Biol Soc 2013;2013:1282−5.

[13] Fedorov A, Longabaugh WJR, Pot D, Clunie DA, Pieper S, Aerts H, et al. NCI imaging data commons. Cancer Res 2021;81:4188−93.

[14] Halling-Brown MD, Warren LM, Ward D, Lewis E, Mackenzie A, Wallis MG, et al. OPTIMAM mammography image database: a large-scale resource of mammography images and clinical data. Radiol Artif Intell 2021;3:e200103.

[15] Moura DC, Guevara Lopez MA. An evaluation of image descriptors combined with clinical data for breast cancer diagnosis. Int J Comput Assist Radiol Surg 2013;8:561−74.

[16] Baughan N, Whitney H, Drukker K, Sahiner B, Hu TT, Hyun KJG, et al. Sequestration of imaging studies in MIDRC: a multi-institutional data commons. Proc SPIE 2022;12035:91-98.

[17] Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL. Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J Cogn Neurosci 2007;19:1498−507.

[18] Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP, et al. Video-based AI for beat-to-beat assessment of cardiac function. Nature 2020;580:252−6.

[19] Leclerc S, Smistad E, Pedrosa J, Ostvik A, Cervenansky F, Espinosa F, et al. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE Trans Med Imaging 2019;38:2198−210.

[20] Wang XS, Peng YF, Lu L, Lu ZY, Bagheri M, Summers RM. ChestX-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proc IEEE Conf Comput Vis Pattern Recognit 2017;2017:3462−71.

[21] Guo S. DPN: detail-preserving network with high resolution representation for efficient segmentation of retinal vessels. J Ambient Intell Humaniz Comput 2023;14:5689−702.

[22] Members C-N, Partners. Database resources of the National Genomics Data Center, China National Center for Bioinformation in 2023. Nucleic Acids Res 2023;51:D18−D28.

[23] Wang Y, Song F, Zhu J, Zhang S, Yang Y, Chen T, et al. GSA: genome sequence archive. Genomics Proteomics Bioinformatics 2017;15:14−8.

[24] Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, et al. The genome sequence archive family: toward explosive data growth and diverse data types. Genomics Proteomics Bioinformatics 2021;19:578−83.

[25] Moore SM, Maffitt DR, Smith KE, Kirby JS, Clark KW, Freymann JB, et al. De-identification of medical images with retention of scientific research value. Radiographics 2015;35:727−35.

[26] Lindeberg T. Scale invariant feature transform. Scholarpedia J 2012;7:10491.

[27] Pietikäinen MJS. Local binary patterns. Scholarpedia J 2010;5:9775.

[28] Dalal N, Triggs B. Histograms of oriented gradients for human detection. IEEE Comput Soc Conf Comput Vis Pattern Recognit 2005;1:886−93.

[29] Qayyum A, Anwar SM, Awais M, Majid M. Medical image retrieval using deep convolutional neural network. Neurocomputing 2017;266:8−20.

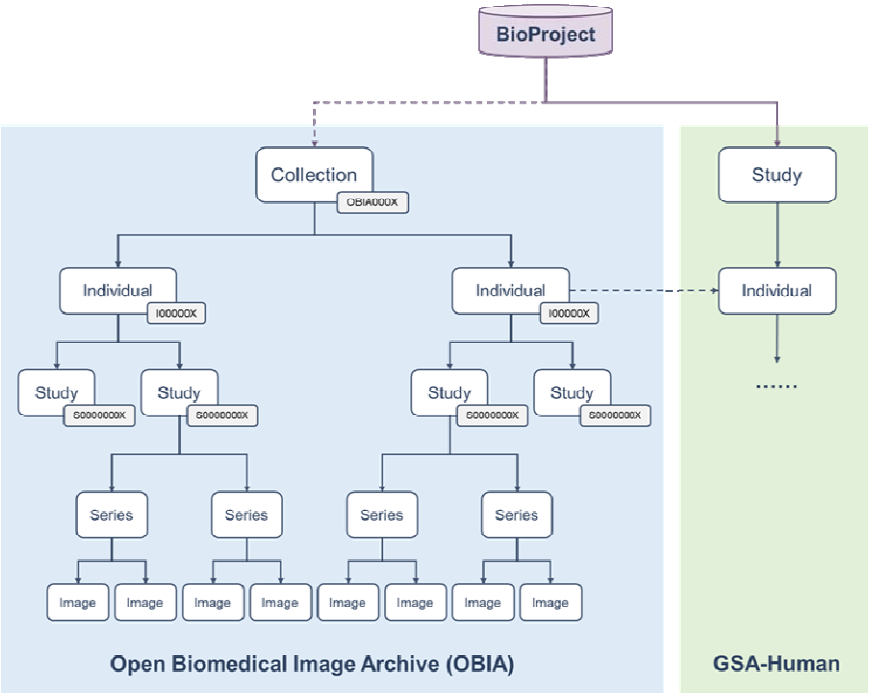
[30] Fang J, Fu H, Liu J. Deep triplet hashing network for case-based medical image retrieval. Med Image Anal 2021;69:101981.

[31] Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks. Proc Mach Learn Res 2019;97:6105−14.

[32] Monteiro E, Costa C, Oliveira JL. A de-identification pipeline for ultrasound medical images in dicom format. J Med Syst 2017;41:89.

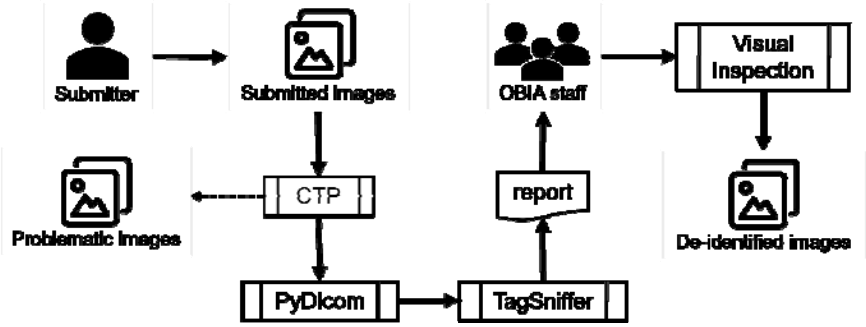
Figure legends

Figure 1 OBIA data model



The collection and individual in OBIA can be linked to BioProject and GSA-Human respectively. The accession numbers for data objects, including Collection, Individual, and Study, are indicated in the grey box. Collection accession numbers have a prefix of "OBIA" followed by four consecutive digits, Individual accession numbers have a prefix of "I" followed by six consecutive digits, and Study accession numbers have a prefix of "S" followed by eight consecutive digits.

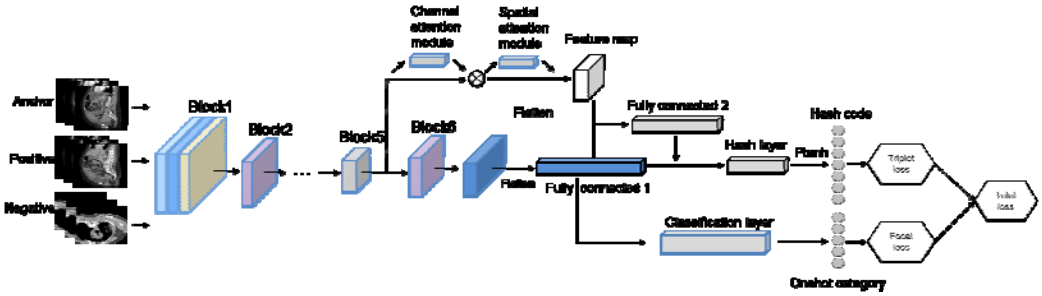
Figure 2 OBIA de-identification and quality control mechanism



Flowchart shows image submission, de-identification, and quality control steps. The de-identification steps include using CTP to process standard tags and PyDicom to handle private tags, and the problematic images will be isolated. The quality control steps include review of reports generated by TagSniffer and visual inspection of image pixels by OBIA staff. CTP, clinical trial processor.

Figure 3 Deep triplet hashing based on attention and layer fusion module

The model uses EfficientNet-B6 as the backbone network and utilizes the CBAM attention module in Block5 to obtain feature maps. Layer fusion is employed in the fully connected layers, and focal loss and triplet loss are used to generate hash code and class embedding. CBAM, convolutional block attention module.



Tables

Table 1 Number of Individual, Study, Series, and Image of each Collection

Accession	Disease	Individual	Study	Series	Image
OBIA0001	Endometrial Cancer	316	1587	8941	779,171
OBIA0002	Ovarian Cancer	202	1548	7151	548,174
OBIA0003	Cervical Cancer	117	675	3943	307,101
OBIA0004	Endometrial Cancer	302	326	4666	303,863

Note: The data statistics are up to September 2023.