SN Computer Science (2024) 5:676
https://doi.org/10.1007/s42979-024-03025-7

ORIGINAL RESEARCH

Integrated Dataset-Preparation System for ML-Based Medical Image Diagnosis with High Clinical Applicability in Various Modalities and Diagnoses

My N. Nguyen[1,2] · Kotori Harada[1] · Takahiro Yoshimoto[1] · Nam Phong Duong[1] · Yoshihiro Sowa[3] · Koji Sakai[4] · Masayuki Fukuzawa[1]

Received: 31 March 2024 / Accepted: 31 May 2024

Abstract

This study proposed an integrated dataset-preparation system for ML-based medical image diagnosis, offering high clinical applicability in various modalities and diagnostic purposes. With the proliferation of ML-based computer-aided diagnosis using medical images, massive datasets should be prepared. Lacking of a standard procedure, dataset-preparation may become ineffective. Besides, on-demand procedures are locked to a single image-modality and purpose. For these reasons, we introduced a dataset-preparation system applicable for a variety of modalities and purposes. The system consisted of a common part including incremental anonymization and cross annotation for preparing anonymized unprocessed data, followed by modality/subject-dependent parts for subsequent processes. The incremental anonymization was carried out in batch after the image acquisition. Cross annotation enabled collaborative medical specialists to co-generate annotation objects. For quick observation of dataset, thumbnail images were created. With anonymized images, preprocessing was accomplished by complementing manual operations with automatic operations. Finally, feature extraction was automatically performed to obtain data representation. Experimental results on two demonstrative systems dedicated to esthetic outcome evaluation of breast reconstruction surgery from 3D breast images and tumor detection from breast MRI images were provided. The proposed system successfully prepared the 3D breast-mesh closures and their geometric features from 3D breast images, as well as radiomics and likelihood features from breast MRI images. The system also enabled effective voxel-by-voxel prediction of tumor region from breast MRI images using random-forest and k-nearest-neighbors algorithms. The results confirmed the efficiency of the system in preparing dataset with high clinical applicability regardless of the image modality and diagnostic purpose.

Keywords Computer-aided diagnosis · Machine learning · Dataset preparation · Anonymization · Cross annotation · Feature extraction

✉ My N. Nguyen
nnmy.ag@gmail.com

✉ Masayuki Fukuzawa
fukuzawa@kit.ac.jp

1 Graduate School of Science and Technology, Kyoto Institute of Technology, Matsugasaki, Sakyo-ku, Kyoto 606-8585, Japan

2 College of Information and Communication Technology, Can Tho University, 3-2 Street, Ninh Kieu District, Can Tho, Vietnam

3 Department of Plastic Surgery, Jichi Medical University, Yakushiji, Shimotsuke-shi, Tochigi 329-0498, Japan

4 Graduate School of Medical Science, Kyoto Prefectural University of Medicine, Kajii-cho, Kawaramachi Hirokoji, Kamigyo-ku, Kyoto 602-8566, Japan

Introduction

Machine learning-based computer-aided diagnosis (ML-based CAD) is a field that involves analyzing large datasets of patient data, particularly medical images, to assist clinicians in decision-making. Numerous studies in this field have been conducted on different subjects and image modalities such as the characterization of breast tumors with MRI scans [1–3], the detection of cerebral aneurysms with CT angiographies [4, 5], and the detection of lung nodules with chest X-rays [6, 7]. Within such studies, the dataset serves as the foundation upon which ML models are trained, directly influencing performance of a CAD system. However, preparation of the dataset entailed difficulty due to the

Published online: 26 June 2024

repeat of hard-to-automate processes such as image acquisition, anonymization, annotation and preprocessing, therefore required additional time and effort from research participants. Several studies, including our earlier work, proposed comprehensive process flows to address the difficulty and efficiently prepare the dataset [8, 9]. Nevertheless, our previously proposed process flow was still rigid, as it required a complete re-implementation when dealing with another image modality and diagnostic purpose. As a result, this study improves our former study by reorganizing processes into common and modality/subject-dependent parts, therefore enabling the system to be partly reusable on different joint-research projects regardless of image modality and diagnostic purpose. To evaluate its effectiveness, we introduce two demonstrations related to breast cancer diagnosis.

Breast cancer diagnosis is a typical example of ML-based CAD applications. Advanced techniques allow early detection and treatment of the cancer with an excellent survival expectation. By providing detailed images of breast tissue through MRI scans, precise information about tumor size, location, and proximity to surrounding structures can be realized with popular ML algorithms. Furthermore, an extensive amount of quantitative data extracted as radiomics features from the tumor region can be helpful in diagnosis and prognosis since they are believed to reveal underlying mechanisms at genetic and molecular levels [10]. Meanwhile, some researchers have shifted their focus to long-term esthetic outcomes following breast reconstruction surgery due to the cancer treatment. Various developed tools have been introduced such as the Breast Cancer Conservative Treatment (BCCT.core) [11], the Breast Analyzing Tool (BAT) [12] and the kOBCS© [13] to overcome the subjectivity and provide reliability in the evaluation of esthetic outcome. Among these, BCCT.core was utilized in some recent studies [14, 15], however, it is challenging to perform collective evaluation of multiple cases due to its requirements of interactive operations in every step.

The rest of this paper is structured as follows. "Dataset Preparation in ML-Based Medical Image Diagnosis" section presents our improved procedure for dataset preparation in ML-based CAD. "Procedure of Image Acquisition, Anonymization and Annotation" section describes the implementation of the common process part for two demonstrative systems. The processing specific to each of the demonstrative systems is introduced in "Processing of 3D Breast Images for Geometric Feature Extraction" and "Processing of Breast MRI Images for Radiomic Feature Extraction" sections. Finally, "Conclusion" section offers a summary of the main contributions drawn from this study.

Dataset Preparation in ML-Based Medical Image Diagnosis

In practice, ML-based CAD systems require the cooperation from different participant groups including medical institution, collaborative medical specialists, and research institution to prepare a large number of training datasets. Nonetheless, such cooperation may lead to issues in data consistency, confidentiality and interoperability unless a well-organized procedure is established. The expected procedure should not only address these issues but also be widely applicable to a variety of image modalities for different diagnostic purposes. Accordingly, we developed a novel dataset preparation process for ML-based CAD and its schematic representation is shown in Fig. 1. Briefly, it consists of a common part for generating anonymized and unprocessed exchanging data, followed by modality/subject-dependent parts for processing these data into a final dataset targeting specific diagnostic purposes.

Common Part for Image Acquisition, Anonymization and Annotation

The common part aims to generate an anonymized version of unprocessed data for exchanging among participant groups. Since implementation of processes in this part is either similar or falls under few anticipated cases, it is expected to be highly reusable across different research.

Medical institutions perform image acquisition, incremental anonymization and filtering. As part of routine tasks, it is favorable to commonly adopt a predefined configuration when acquiring clinical images from individual patients. Then, these images and sensitive information are securely stored in a confidential media. Occasionally, anonymization and filtering are carried out for multiple patients at once to generate anonymized images with less effort. Required objects/tools for sending annotation requests to medical specialists are also prepared in advance. As for the annotation form, employing an Optical Character Recognition (OCR) design will accelerate subsequent data collection and error correction at research institution.

Collaborative medical specialists from the medical institution in which the clinical images are acquired or other institutions are requested to perform the manual annotation. By observing given anonymized images, specialists reflect their diagnosis and findings as ROI bounding box, ROI mask or information filled into the annotation form. Each examination case is supposed to undergo multiple times of annotation by different specialists, ensuring a thorough evaluation.

Eventually, outcome of the common part consists of anonymized images and annotation objects like filled

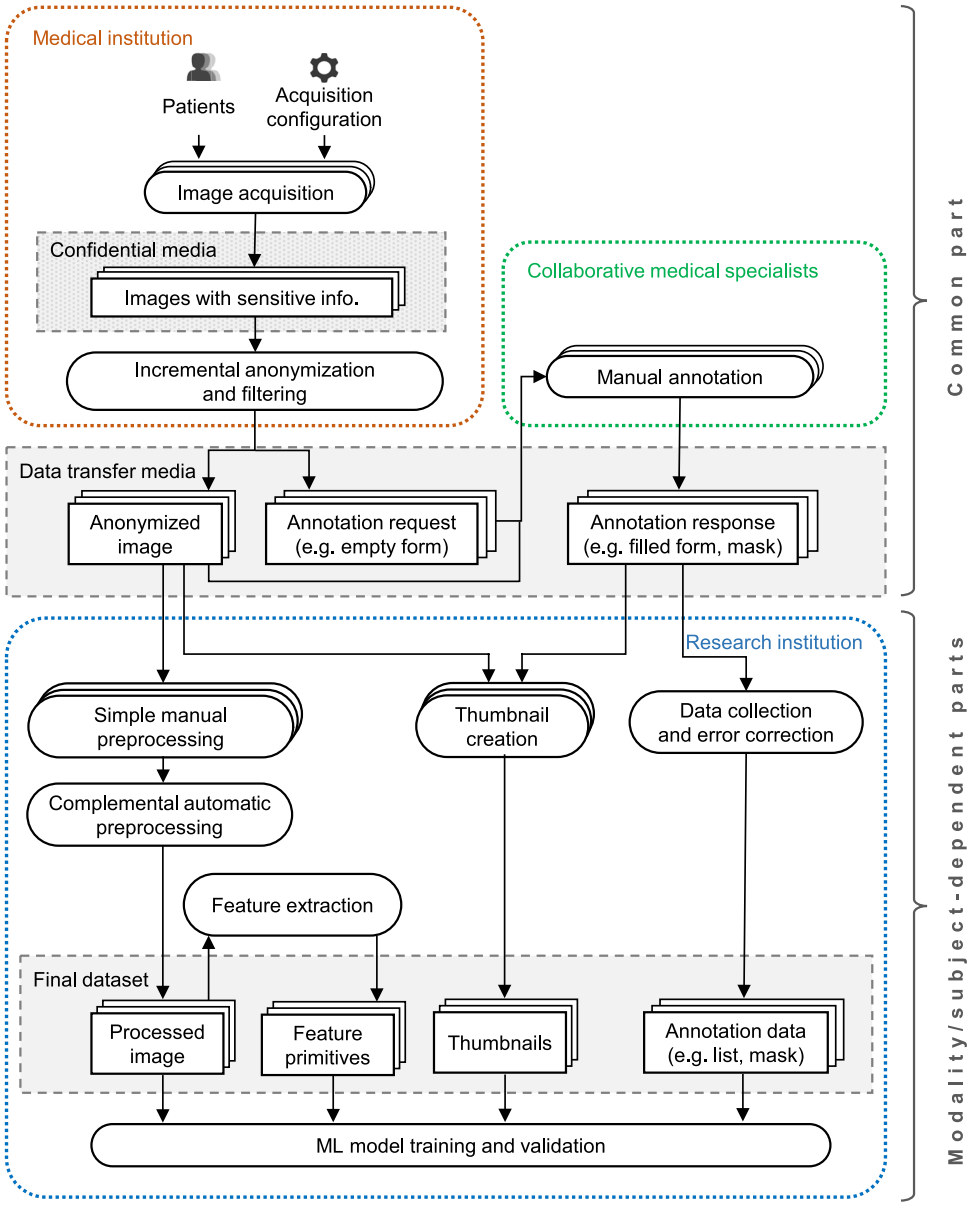


Fig. 1 Process flow of dataset preparation dedicated to ML-based CAD

forms and ROI masks, which are stored in a data transfer media. Since this media serves as a hub for the interaction among participant groups, data consistency and interoperability are guaranteed. Meanwhile, confidentiality is maintained through incremental anonymization. By widely applying the workflow of this part across studies, unnecessary variation of procedures due to difference in image modality and diagnostic purpose can be alleviated.

Modality/Subject-Dependent Part for Preprocessing and Feature Extraction

This part aims to transform anonymized images into feature primitive, generate thumbnail images, and collect valid annotation data. Then, its resulting final dataset is utilized for ML model training and validation. Unlike the common part, processes in this part are tailored to deal with specific image modality for particular diagnostic purpose.

Annotation objects prepared by medical specialists may not be immediately usable for calculation by the research institution because of their paper-based representation or their mismatch with desired format. In order to quickly access the data in filled annotation forms, the use of an OCR device is necessary. From the fields and data recognized by OCR, error correction is performed to confirm the data validity. As for the received ROI markups in form of mask and bounding box, further checks are required to compare their spatial dimension with corresponding anonymized image and to ensure a mutual agreement of markup locations among different annotators. Next, transforms including value

range standardization and format conversion can be executed on qualified markups. Annotation data qualified from the data collection and error correction process are then saved to the final dataset.

Since clinical images are often stored in specialized formats such as DICOM or NIfTI, they may be incompatible with popular image viewers and their previews within the dataset are not possible. The use of dedicated software, however, comes at high cost in terms of time and memory. For these reasons, thumbnail images are created as a compact representation of clinical images, enabling quick observation of multiple images in the dataset without loading them.

Regarding the preprocessing, it can be enhanced by incorporating complementary automatic tasks following hard-to-automate tasks. From the entire or local region of each processed image, feature primitives such as pixel-value statistics, geometric features, or transformed image are extracted and appended to the final dataset. At the end, a particular ML model is trained and validated by employing the prepared dataset.

The proposed process flow can be reused in other studies of ML-based medical image diagnosis with high clinical applicability regardless of image modality or diagnostic purpose. Its common part facilitates the participation in cross annotation from external collaborators while securing patient privacy through automated anonymization of clinical images. By breaking the system down into parts, new modality/subject-dependent parts can be added and customized without disrupting the existing procedure of the common part.

Procedure of Image Acquisition, Anonymization and Annotation

In this section, implementation derived from a common procedure for incremental anonymization and cross annotation of two demonstrative applications utilizing 3D breast images and breast MRI images is described. Further processing of each application will be presented in subsequent sections.

For the application of breast esthetic outcome evaluation, 3D images were prepared through the use of a handheld depth camera (Intel Realsense L515 [16]) to reflect the shape aspects of the breast. For the application of breast tumor detection, various T1-weighted MRI images were acquired by 1.5 T or 3 T scanners to reflect the tissue structures inside the breast before and after contrast enhancement.

Regarding the anonymization, personal information that reveals patient's identity was completely removed both in folder name, file name, file header, and image content from acquired images, while some non-personal information remained for certain purposes. Particularly, some examination-related information in folder name and file name can be included, such as patient ID, acquisition date, and acquisition parameters (like MRI sequence and sequence order in MRI images), for ease of study organization. As for the file header, some demographic information such as sex, age, and weight can be provided. In the case of MRI header, information dedicated to image orientation and position, slice thickness, pixel spacing were included to ensure interoperability and compatibility across different imaging systems. As for image content, patient faces must be avoided in all images. The incremental anonymization was automatically performed on a regular basis and on only newly added images, therefore it eliminated the repeated anonymization of previously processed images.

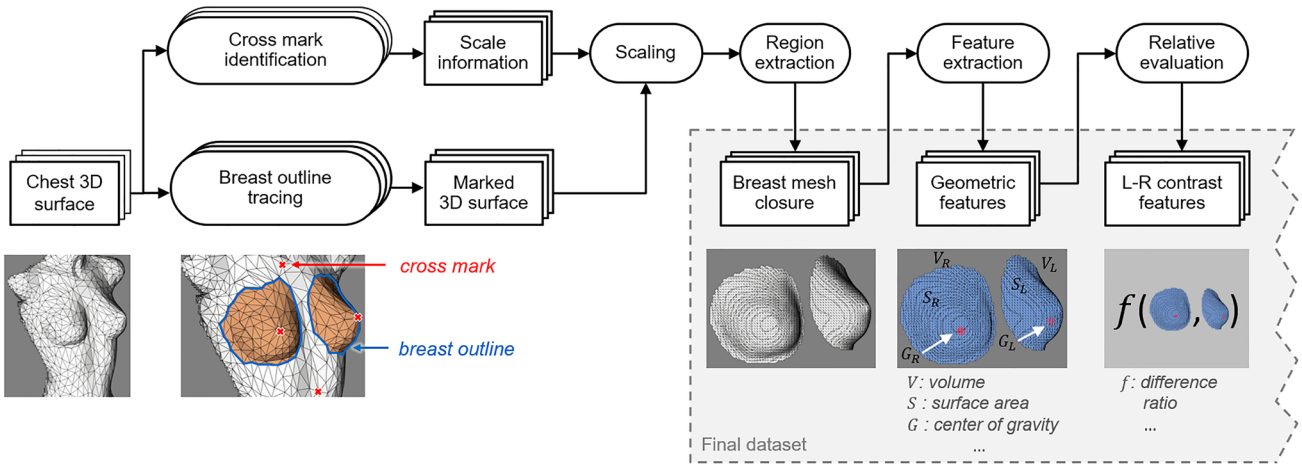
Regarding the cross annotation, anonymized 3D breast images and corresponding OCR-compatible annotation sheets were distributed to collaborative medical specialists for filling their diagnoses. In case of breast MRI images, coordinates of bounding boxes encapsulating the breast tumor were given by the dataset provider. Based on these boxes, which were drawn by eight radiologists, we also prepared the tumor mask for selected images using the Segmentation module in 3D Slicer software. Various types of annotation are expected to serve a wider spectrum of analyses.

Processing of 3D Breast Images for Geometric Feature Extraction

The processing of 3D breast images for esthetic outcome evaluation and its typical output are depicted in Fig. 2. Since they were thoroughly described in our earlier study [8], this section aims to briefly review them from the perspective of modality/subject-dependent processing part.

The evaluation of postoperative breasts was based on factors such as breast size, height of the inframammary fold, and nipple position. Our research dealt with these factors by considering the disparities between the left and right breasts from several viewpoints including shape and appearance. For the time being, 3D breast meshes constructed from triangular cells and shared edges were adopted to serve as a standard shape-based viewpoint. Figure 2a presents the process of handling these 3D breast meshes. It commenced by extracting the breast region which is a primary requirement for further calculation. It then continued with several hard-to-automate processes such as tracing the breast outline and identifying the cross marks and nipples. Then, automatic scaling and region extraction processes were carried out to collect a normalized breast mesh surface. In the next step, various geometric features including volume, surface area, and center of gravity were automatically extracted from each breast mesh closure. Eventually, L-R contrast features based on the difference and the ratio of bilateral breasts were derived by comparing the geometric feature values between breasts. Besides 3D breast

a) Process flow



b) Typical outputs

Breast mesh closure		A	B	C	D
L-R Contrast	V_R/V_L	1	1.1	2.1	13.1
	S_R/S_L	1	1.2	1.8	2.4
Esthetic score	BV*	2	1	0	0
	BS**	2	0.75	0	0

*BV: Breast Volume
* * BS: Breast Shape

Fig. 2 **a** Process flow and **b** typical outputs in processing of 3D breast images

images, their corresponding 2D images were also available as JPG files, therefore the thumbnail creation process was not needed. Processing procedure for such 2D images is not covered in this study.

As a result, four typical examples of extracted breast mesh closures and their feature primitives are presented in Fig. 2b. The mesh closures exposed a smooth polyhedral surface without any holes or splits. Based on these mesh closures, L-R contrast features such as volume ratio V_L/V_R and surface ratio S_L/S_R were calculated. Aside from that, esthetic scores averaged from evaluation scores of four specialists were also provided. The result revealed that as L-R features approached 1.0, the esthetic scores became higher, indicating a strong correlation between extracted features and esthetic outcome.

Processing of Breast MRI Images for Radiomics Feature Extraction

Aside from conventional 2D and 3D images used for assessing the external appearance of post-surgery breasts, MRI images provide a non-invasive method to explore internal breast structures and identify pertinent abnormalities. This section covers the adoption of the proposed integrated system to efficiently prepare a dataset of breast MRI images dedicated to breast tumor detection.

Material

For demonstrative purpose, we utilized the breast MRI images from the publicly available Duke-Breast-Cancer-MRI dataset [17], accessible online at the Cancer Imaging Archive website (www.cancerimagingarchive.net). This dataset consists of 922 patients with biopsy-confirmed invasive breast cancer and its preoperative MRI images were acquired by 1.5 T or 3 T scanners in the prone positions. The images were provided in DICOM format of axial plane and encompass non-fat saturated T1-weighted, fat-saturated gradient echo T1-weighted pre-contrast, and post-contrast sequences.

Patient identity information was completely removed in folder name, file name, file header, and image content from acquired images. In order to annotate the tumor position for each image, coordinates of tumor bounding boxes were provided by the dataset. Additionally, other annotation data was given in a worksheet file, for example, side of cancer (left or right) and Nottingham grade.

Creation of Thumbnail Images

Since the MRI images were in DICOM format and tumor masks were in NRRD format, the preparation of thumbnail images in a commonly-used file format, such as JPEG, will facilitate the immediate preview of multiple annotated MRI images. Figure 3 describes how a thumbnail image was generated from an MRI image and corresponding mask. The procedure begins by reading and considering each slice of the mask to select a representative slice. The selection criteria can be determined by slice-based measures such as mask area and mask diameter, or just by utilizing the middle slice of the mask. In our study, the slice with maximum mask area was chosen as the representative slice. With the chosen slice index, representative MRI and mask slices were extracted from the MRI image and its mask. Before creating the thumbnail image, some settings like mask appearance (contour, filling, opacity), output image dimension, and file format should be selected. Finally, the thumbnail image was created by subsequently overlaying three layers from background to foreground: representative MRI slice, representative mask slice, and other annotation data displayed as image caption.

Preprocessing and Feature Extraction

Anonymized MRI images were standardized by the resampling process, followed by repeated sliding of an equilateral kernel to extract local features. The process flow for these operations is illustrated in Fig. 4.

Regarding the preprocessing, it is not appropriate to perform the analysis or comparison across multiple images since different MRI images in the dataset may not share the same voxel dimension or voxel aspect ratio. Consequently, resampling was carried out on MRI images to standardize the voxel dimension. In our system, MRI images were resampled to the common voxel dimension of 1 mm × 1 mm × 1 mm. In addition, corresponding masks need to be updated to align with the resampled MRI images.

For medical image processing, target structures are often presented within a smaller region of interest. The global

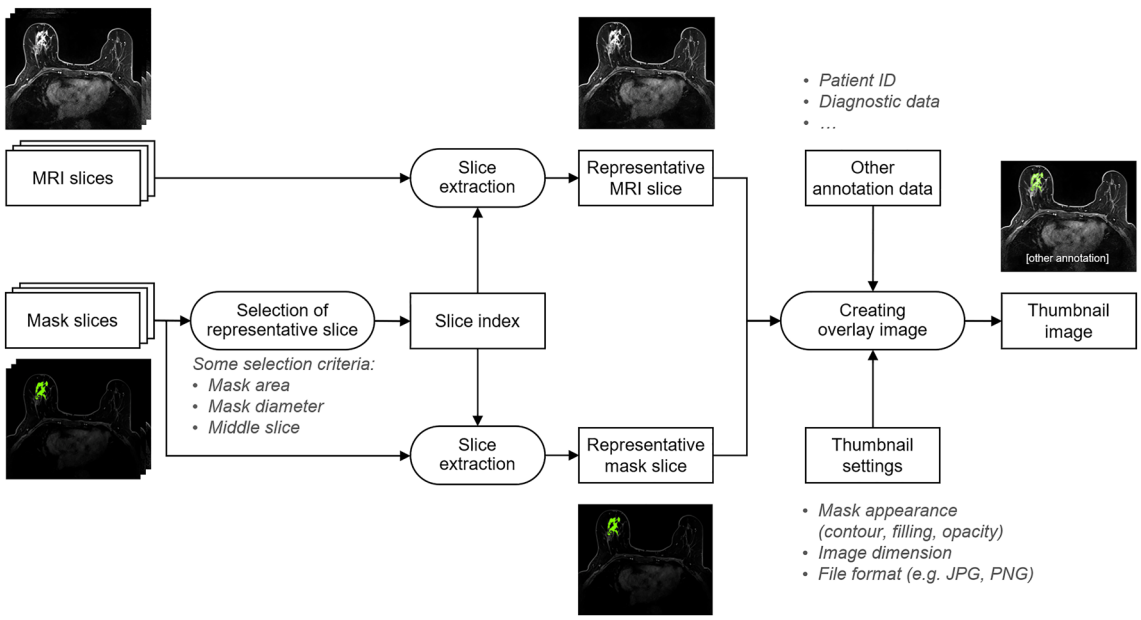


Fig. 3 Process flow of preparing thumbnail images from MRI and mask images

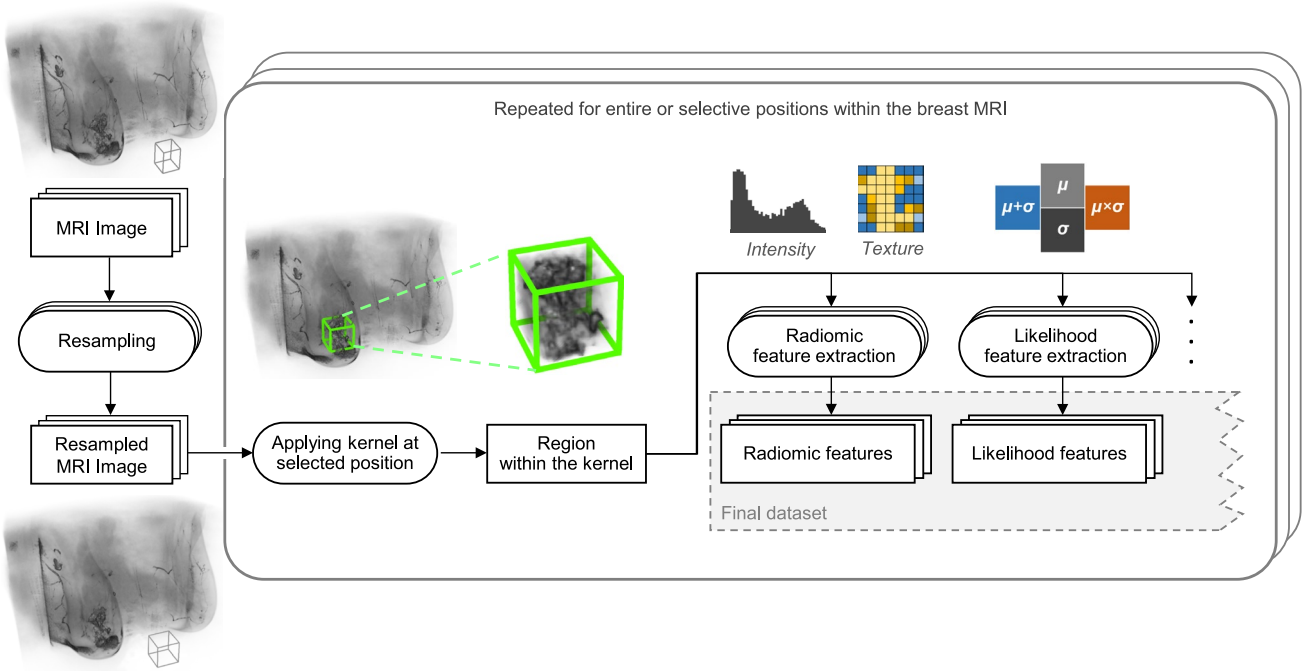


Fig. 4 Preprocessing and feature extraction flow in breast MRI images

features are not suitable in these scenarios because they capture insights from the entire image as a whole, leading to the overlook of regional structures. For these reasons, local features are advantageous due to its ability to capture the texture within a specific region. To begin, an equilateral kernel slides over the entire or selective positions within the resampled MRI image and extracts the region encapsulated by the kernel. The size of the kernel decides the field of view of local structures; therefore, it should not be significantly larger or smaller compared to the typical size of target structure. At each kernel position, the extracted local region was used to calculate two classes of features including radiomics features and likelihood features. Descriptions of these features are as follows.

- Radiomics features: These are quantitative features resulting from the conversion of medical images to mineable high-dimensional data. This process was driven by the belief that biomedical images reflect underlying pathophysiology, so their relationship can be revealed by extracting maximal information from care images [18, 19]. For demonstrative purposes, we extracted 15 selective radiomics features representing the intensity distribution and texture from all voxels within the local region, regardless of the tumor mask. The list of these radiomics features is included in Fig. 5 and the meaning of each feature was described at [20].
- Likelihood features: Based on the empirical observation from various breast MRI images, the region associated with the tumor tends to have different average brightness and heterogeneity compared to other regions. For this reason, likelihood features make use of voxel intensity's mean, standard deviation or their combinations like mutual sum or mutual product to characterize the probability of having tumor within the region.

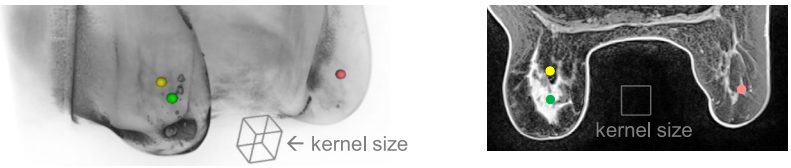
Two classes of features mentioned above were adopted to validate our proposed system. The diagnosis is not limited to only our features, but other handcrafted features or those from deep learning can also be applicable.

Experimental Result

Time Performance

For selected MRI images acquired from Duke-Breast-Cancer-MRI dataset, we referred to corresponding bounding boxes and drew the tumor mask as NRRD (Nearly Raw Raster Data) file using the Segmentation module in 3D Slicer software. As a result, the approximate duration to draw the tumor mask within the bounding box of size 41.5 mm × 38.8 mm × 12 mm from an MRI image is 6 min. These additional tumor masks enable the visualization of real tumor shape, as well as the analysis of geometric, intensity and texture inside the tumor.

To load and render one typical breast MRI image together with its tumor mask in 3D Slicer, it required roughly 1 min and 30 MB of memory. Meanwhile, it took about 7 s to



		● (inside-tumor)	● (overlapping)	● (control)	● / ●	● / ●
Intensity	img-Maximum	669	414	253	2.64	1.64
	10Percentile	21	22	31	0.68	0.71
	Entropy	3.45	3.26	2.02	1.71	1.61
	InterquartileRange	164	131	33	4.97	3.97
	Kurtosis	3.41	2.14	5.9	0.58	0.36
	Skewness	0.83	0.61	1.83	0.45	0.33
Texture	Autocorrelation	38.73	28.89	10.54	3.67	2.74
	ClusterShade	241	99	44	5.43	2.24
	Contrast	5.04	3.08	1.53	3.29	2.01
	JointEnergy	0.03	0.04	0.23	0.13	0.17
	GrayLevelNonUniformity (GLCM)	1998	2196	2837	0.7	0.77
	RunVariance	0.93	0.92	5.87	0.16	0.16
	GrayLevelNonUniformity (GLSZM)	138.55	81.52	91.74	1.51	0.89
	LargeAreaEmphasis	46839	102125	600509	0.08	0.17
	LargeAreaHighGrayLevelEmphasis	836702	1633893	2875904	0.29	0.57
Likelihood	LocMean	115.82	101.43	60.12	1.93	1.69
	LocSTD	92.6	74.76	39.32	2.36	1.9
	LocMeanxLocSTD	10725	7583	2364	4.54	3.21
	LocMean+LocSTD	208.42	176.19	99.44	2.1	1.77
Tumor rate		0.28	0.12	0		

Fig. 5 Typical output features of preprocessing flow in breast MRI images

generate a thumbnail image of less than 100 KB, which can be easily and quickly loaded at any time. The availability of thumbnail images enabled the preview through large collections of annotated MRI images.

As for preprocessing time, approximately 55 s were utilized to resample an MRI image and its tumor mask by manual manipulation with the Resample Scalar Volume module. In case the resampling was performed automatically, the expended time was only about 2.17 s.

Regarding the feature extraction, a kernel size of 32 mm × 32 mm × 32 mm was adopted to capture tumors of moderate and big sizes. The extraction time at each kernel position with our Slicer Graphic User Interface (GUI) module was about 0.035 s for 15 radiomics features and about 0.001 s for four likelihood features. In the same manner, if the features are calculated for all kernel positions within a breast MRI image of size 300 mm × 300 mm × 185 mm, it will take approximately 7 days. However, the feature extraction can take significantly less time if it is carried out without the GUI and on only selective kernel positions.

With time performance examined as above, the total duration is acceptable for preparing the dataset of various feature types. In addition, the availability of tumor masks and thumbnail images is believed to facilitate future analyses while maintaining a short generation time. Consequently, the proposed system is deemed appropriate to support medical image diagnosis with high clinical applicability.

Feature Effectiveness

Figure 5 shows the radiomics and likelihood features extracted from three typical kernel positions within a typical pre-contrast breast MRI image. These positions consisted of a position deep inside the tumor, a position where the kernel slightly overlaps with the tumor, and a position outside the tumor (control position). The kernel size was chosen to be 32 mm × 32 mm × 32 mm. The last two columns, which provide the ratio of feature value between tumor-involved positions and control position, are useful to select best features for characterizing the tumor. Based on the presented result, we selected five features including InterquartileRange, Autocorrelation and ClusterShade from radiomics group, LocMeanxLocSTD and LocSTD from likelihood group for further investigation.

After selecting features, their data were extracted on various kernel positions from MRI images of multiple patients and utilized to train several machine learning models for detecting whether the kernel center is inside a

Table 1 Training experiments of tumor detection models

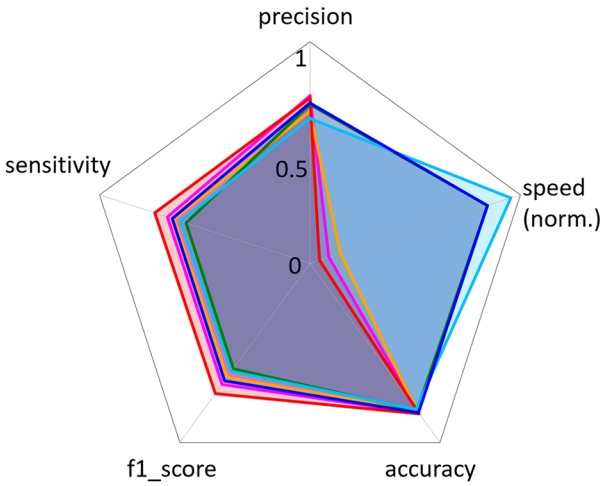
Dataset	Feature set	Algorithm	Evaluation
Pre-contrast images: 10 Data samples: 2514 (+) inside-tumor: 318 (−) outside-tumor: 2196	Radiomics: 15 Likelihood: 4 Combined: 19	RF KNN	Cross validation: tenfold Metrics: averages of precision, sensitivity, f1-score, accuracy, and speed (logarithmic inverse of training time) after cross validation

tumor or not. The purpose of these models is to confirm the feature effectiveness rather than to perform a complete tumor segmentation problem, which is beyond the scope of this paper. The model training experiment was described in Table 1. Particularly, 2514 data samples extracted from 10 pre-contrast MRI images were organized into three distinct feature sets: radiomics with 15 features, likelihood with four features, and their combination with 19 features. Six models were trained from these feature sets and two algorithms including Random Forest (RF) and K-Nearest Neighbors (KNN). RF models were trained with 100 trees (`n_estimators` = 100), each tree used 1000 samples randomly drawn from the dataset (`max_samples` = 1000), and the number of features considered for splitting at each leaf node was set to the square root of the total number of features (`max_features` = $\sqrt{N_f}$). KNN models made predictions by evaluating cosine distance from 10 nearest neighbors. The cosine distance was chosen because it is less affected by high-dimensional sparsity that some other distance metrics, like Euclidean distance, experience. The models were then assessed over tenfold cross validation and averages of precision, sensitivity, f1-score, accuracy, and speed (logarithmic inverse of training time) were treated as performance metrics. Precision assesses the correctness of positive predictions and is measured by the proportion of true positives from all predicted positives. Sensitivity (recall) evaluates the model's ability to detect true positives, indicating the proportion of true positives among all actual positives. F1-score represents a balanced measure by taking the harmonic mean of precision and recall. Accuracy reflects the overall correctness of a model and is calculated as the proportion of correct predictions out of all predictions.

The results of model assessment are shown in Fig. 6. RF model for combined feature set achieved the best score in term of sensitivity (0.804), f1 (0.795), and accuracy (0.897), whereas RF model for radiomics achieved the best precision (0.821). As for KNN models, they were superior to RF models only in term of speed and combined feature set also attained the best performance. These results confirmed the effectiveness of radiomics and likelihood features and their potentials in breast tumor diagnosis with MRI images.

Conclusion

In this article, we have proposed an integrated dataset-preparation system dedicated to ML-based medical image diagnosis with high clinical applicability, targeting any modality and diagnostic purpose. Processes in the proposed system were arranged into common part and modality/subject-dependent parts, with the common part encompassing general functions such as incremental anonymization and cross annotation, while the modality/subject-dependent parts accommodating functions tailored to specific modality and



	Model	Precision	Sensitivity	F1_score	Accuracy	Speed (normalized)
RF	Radiomics	**0.821**	0.75	0.749	0.893	0.22
	Likelihood	0.757	0.721	0.713	0.883	0.27
	Combined	0.812	**0.804**	**0.795**	**0.897**	0.18
KNN	Radiomics	0.786	0.67	0.67	0.881	0.9
	Likelihood	0.729	0.693	0.683	0.875	**1**
	Combined	0.792	0.729	0.731	0.892	0.9

Fig. 6 Performance comparison of tumor detection models

purpose such as thumbnail creation, preprocessing and feature extraction. The incremental anonymization and filtering streamlined batch processing for acquired images, thereby reducing the workload of medical specialists. Then, cross annotation on anonymized images enables the privacy-ensured and robust collaboration between different specialists. Depending on the format of clinical images, thumbnail images can be generated to provide a quick observation across the dataset. To accelerate the preprocessing and feature extraction, a process flow that combined manual and complementary automatic operations was also designed. Our system offered advantages in terms of procedure reusability, scalability, and confidentiality for joint projects with various modalities and purposes.

Two demonstrative systems were successfully employed to prove the effectiveness and high applicability of the developed procedure. Although each of them was associated with different image modality and different diagnostic purpose, they can share similar implementation on the common process part. Subsequent processing, especially the feature extraction, was customized to each demonstrative system. System dedicated to plastic surgery evaluation from 3D breast images competently prepared the datasets of 3D breast-mesh closures and their corresponding L-R contrast features. Preliminary result exhibited a strong correlation between extracted features and esthetic outcome assessed by medical specialists. Regarding the system dedicated to tumor detection from breast T1-weighted MRI images, it successfully generated a dataset of local radiomics and likelihood features from resampled images.

The effectiveness of these features was also confirmed by training several machine learning models to predict tumor region voxel by voxel. Trained with 2514 data samples, the RF model with combined radiomics and likelihood features achieved the best assessment in terms of sensitivity, f1, and accuracy, while the RF model with only radiomics features had the highest precision. In order to further validate the system's capabilities and broaden its applicability, it is necessary to consider additional modalities and diagnostic purposes in future investigations.

Acknowledgements This work was supported by JSPS Core-to-Core Program (grant number: JPJSCCB20230005).

Author Contributions The manuscript was written by My N. Nguyen and revised by all other co-authors. The main process flow was formerly designed by Kotori Harada, Takahiro Yoshimoto, and Nam Phong Duong and improved by My N. Nguyen to accommodate its applicability in various modalities and diagnostic purposes. Dr. Yoshihiro Sowa was in charge of preparing 3D breast images and giving advices on medical aspects of esthetic outcome evaluation. Investigation of 3D breast images was performed by Kotori Harada, Takahiro Yoshimoto, and Nam Phong Duong. Dr. Koji Sakai gave advices on medical aspects of breast cancer and radiomics feature extraction. Investigation of MRI breast images was performed by My N. Nguyen. Dr. Masayuki Fukuzawa initialized, supervised and ensured the integrity of the study.

Funding This work was supported by JSPS Core-to-Core Program (grant number: JPJSCCB20230005).

Data Availability The dataset of 3D breast images is not open to public due to institutional protocol. As for the dataset of MRI breast images, it was obtained from a public source at https://doi.org/10.7937/TCIA.e3sv-re93.

Declarations

Conflict of Interest The authors declare no conflicts of interest.

Ethics Approval System development of this study is an observational study that does not require ethical approval and all the clinical data was obtained after anonymization. The 3D breast image was acquired with opt-out consent at the previous affiliated medical institutions (Kyoto Prefectural University of Medicine and Kyoto University) of the author (Sowa) under the approval of their ethics committee.

References

1. Mann RM, Cho N, Moy L. Breast MRI: state of the art. Radiology. 2019;292(3):520–36. https://doi.org/10.1148/radiol.2019182947.
2. Herent P, Schmauch B, Jehanno P, et al. Detection and characterization of MRI breast lesions using deep learning. Diagn Interv Imaging. 2019;100(4):219–25. https://doi.org/10.1016/j.diii.2019.02.008.
3. Mokni R, Gargouri N, Damak A, Sellami D, Feki W, Mnif Z. An automatic computer-aided diagnosis system based on the multimodal fusion of breast cancer (MF-CAD). Biomed Signal Process Control. 2012;69: 102914. https://doi.org/10.1016/j.bspc.2021.102914.
4. Dai X, Huang L, Qian Y, et al. Deep learning for automated cerebral aneurysm detection on computed tomography images. Int J Comput Assist Radiol Surg. 2020;15:715–23. https://doi.org/10.1007/s11548-020-02121-2.
5. Ueda D, Yamamoto A, Nishimori M, et al. Deep learning for MR angiography: automated detection of cerebral aneurysms. Radiology. 2019;290(1):187–94. https://doi.org/10.1148/radiol.2018180901.
6. Li X, Shen L, Xie X, et al. Multi-resolution convolutional networks for chest X-ray radiograph based lung nodule detection. Artif Intell Med. 2020;103: 101744. https://doi.org/10.1016/j.artmed.2019.101744.
7. Juan J, Monsó E, Lozano C, et al. Computer-assisted diagnosis for an early identification of lung cancer in chest X rays. Sci Rep. 2023;13(1):7720. https://doi.org/10.1038/s41598-023-34835-z.
8. Harada K, Yoshimoto T, Duong NP, Nguyen MN, Sowa Y, Fukuzawa M (2024) A new integrated medical-image processing system with high clinical applicability for effective dataset

preparation in ML-based diagnosis. In: Thai-Nghe N, Do TN, Haddawy P (eds) Intelligent systems and data science. ISDS 2023. Communications in computer and information science, vol 1950. Springer, Singapore, pp 41–50
9. Willemink MJ, Koszek WA, Hardell C, et al. Preparing medical imaging data for machine learning. Radiology. 2020;295(1):4–15. https://doi.org/10.1148/radiol.2020192224.
10. Tagliafico AS, Piana M, Schenone D, Lai R, Massone AM, Houssami N. Overview of radiomics in breast cancer diagnosis and prognostication. The Breast. 2020;49:74–80. https://doi.org/10.1016/j.breast.2019.10.018.
11. Cardoso MJ, Cardoso J, Amaral N, et al. Turning subjective into objective: the BCCT.core software for evaluation of cosmetic results in breast cancer conservative treatment. The Breast 2007;16(5):456–461. https://doi.org/10.1016/j.breast.2007.05.002.
12. Krois W, Romar AK, Wild T, et al. Objective breast symmetry analysis with the breast analyzing tool (BAT): improved tool for clinical trials. Breast Cancer Res Treat. 2017;164:421–7. https://doi.org/10.1007/s10549-017-4255-z.
13. Soror T, Lancellotta V, Kovács G, et al. kOBCS©: a novel software calculator program of the objective breast cosmesis scale (OBCS). Breast Cancer. 2020;27:179–85. https://doi.org/10.1007/s12282-019-01006-w.
14. Kurt S, İlgün AS, Özkurt E, et al. Outcomes of reconstructive techniques in breast cancer using BCCT.core software. World Journal of Surgical Oncology. 2024;22(1):82. https://doi.org/10.1186/s12957-024-03343-3.
15. Trakis S, Lord H, Graham P, Fernandez R. Reliability of the BCCT.core software in evaluation of breast cosmesis—a systematic review. Journal of Medical Imaging and Radiation Oncology 2021;65(6):817–825. https://doi.org/10.1111/1754-9485.13190.
16. Intel Realsense L515, https://www.intelrealsense.com/lidar-camera-l515. Last accessed 25 Mar 2024.
17. Saha A, Harowicz MR, Grimm LJ, et al. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. Br J Cancer. 2018;119(4):508–16. https://doi.org/10.1038/s41416-018-0185-8.
18. Kumar V, Gu Y, Basu S, et al. Radiomics: the process and the challenges. Magn Reson Imaging. 2021;30(9):1234–48. https://doi.org/10.1016/j.mri.2012.06.010.
19. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2016;278(2):563–77. https://doi.org/10.1148/radiol.2015151169.
20. Van Griethuysen JJ, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. Can Res. 2017;77(21):e104–7. https://doi.org/10.1158/0008-5472.CAN-17-0339.