MELBA Journal Machine Learning for Biomedical Imaging

The Indian Mammography DataBase (IMDB): A Versioned Open Mammography Resource from a Screening-Naive Indian Population for Artificial Intelligence Research

Om Shivom **Nagpal** [1], Varun **Holla** [1], Kushagra **Chaturvedi** [1], Sathish **R** [1], Aditi **Madame** [1], Ashish **Rastogi** [1], Vipin **Thampi** [1], Hema **Malhotra** [1], Pushp **Lochan** [2], Mayank **Bharadwaj** [1], Kshitiz **Jain** [3], Chetan **Arora** [3], Debnath **Pal** [2], Sanjay **Thulkar** [1], Smriti **Hari** [1], Tanmaya **Vyas** [2], Nishant **Chavan** [2], Amit **Gupta** [1], Krithika **Rangarajan** [1]

1 Department of Oncoradiology, All India Institute of Medical Sciences (AIIMS), New Delhi
2 Department of Computational and Data Sciences, Indian Institute of Science (IISc), Bengaluru
3 Department of Computer Science and Engineering, Indian Institute of Technology (IIT), New Delhi

Abstract

The development of robust artificial intelligence (AI) algorithms for breast imaging requires large, well-annotated datasets representative of diverse and underrepresented populations. However, publicly available mammography datasets predominantly originate from Western screening programs and provide limited representation of South Asian populations and screening-naive cohorts. We present the Indian Mammography DataBase (IMDB), an open-access mammography resource comprising two independently curated cohorts from a screening-naive Indian population undergoing diagnostic and opportunistic screening mammography. The release includes 12,674 mammography images from 3,219 patients across IMDB r1.0 and r2.0, making it one of the largest publicly available mammography datasets from India. Associated metadata include patient age, breast density, BI-RADS assessment, and histopathology-confirmed outcomes for suspicious lesions. Ground-truth labels were established using histopathology for BI-RADS 4/5 examinations and imaging / telephonic follow-up or triple expert readings for lower-risk examinations. Quality assurance procedures included image-view verification, metadata consistency checks, pathology linkage validation, and de-identification audits. As a representative AI use case, multiple convolutional neural network and YOLO-based models were successfully evaluated and trained using the dataset, demonstrating its suitability for machine-learning applications such as breast cancer classification, risk prediction, and algorithm benchmarking. The dataset is publicly available through the Indian Biological Images Archive (IBIA) and MIDAS, providing a FAIR-compliant resource for the development and evaluation of breast imaging AI systems.

Keywords

Mammography, Breast Cancer, Artificial Intelligence, Screening, India

Article informations

https://doi.org/10.59275/j.melba.2026-bd61
Volume 3, Received: 2026-06-19, Published 2026-09-21
Corresponding author: krithikarangarajan86@gmail.com
Special issue: MICCAI Open Data 2026 x MELBA
Guest editors: AT, MS et al.

1. Background

Mammography remains the cornerstone of breast cancer detection and has enabled the development of numerous artificial intelligence (AI) algorithms for screening and diagnosis Dembrower et al. (2020, 2023). The availability of large public datasets such as CBIS-DDSM, INbreast, VinDr-Mammo, EMBED, and the RSNA Screening Mammography Challenge dataset has accelerated progress in computer-aided detection and risk prediction models Lee et al. (2017b); Moreira et al. (2012); Nguyen et al. (2023); Mazurowski et al. (2022); Shen et al. (2021). More recently, foundation-scale mammography models such as Mirai and other deep learning systems have demonstrated strong performance in breast cancer risk assessment and lesion detection when trained on large, well-curated screening cohorts Yala et al. (2022). However, most publicly

available mammography datasets and landmark AI studies, including MASAI, originate from organized screening programs in North America and Europe and predominantly represent populations participating in organized screening programs Lång et al. (2023); Halling-Brown et al. (2021); Frazer et al. (2023); Lee et al. (2017a). Asian populations are particularly under-represented in these datasets despite having a significantly different clinical and imaging profile of breast cancer Bobdey et al. (2015); Tea et al. (2013).

The clinical characteristics of breast cancer in India differ substantially from those represented in existing public datasets. Breast cancer in India occurs at a younger age than in many Western populations, with the median age at diagnosis reported to be nearly a decade earlier and a substantial proportion of patients presenting before the age of 40 years Malvia et al. (2017). Also a larger cohort of women have increased breast density, which not only reduces mammographic sensitivity but also presents unique challenges for the development and validation of AI systems. As there are no established breast cancer screening programmes in India da Costa Vieira et al. (2017); Francies et al. (2020), a substantial proportion of Indian patients continue to present with locally advanced disease. As a result, mammography datasets from India frequently contain larger, more conspicuous masses that differ from the early-stage cancers commonly encountered in screening cohorts Chen et al. (2020). Indian datasets also contain a broad spectrum of benign breast lesions that are underrepresented in screening-focused datasets. Furthermore, long-term imaging follow-up, which is routinely available in many Western screening datasets and used to establish benign outcomes, is often unavailable because patients may seek care at multiple institutions or fail to return for interval surveillance. These factors create unique challenges for dataset curation while simultaneously providing an opportunity to develop and evaluate AI models in these underrepresented populations.

Only a few available mammography datasets include dedicated data from Asian populations and diagnostic or opportunistic screening settings. To address this, we present an open-access mammography dataset—the Indian Mammography DataBase (IMDB)—curated from a large tertiary care institution in India. The first release (IMDB r1.0) comprised examinations acquired during 2020 and was designed to provide pathology-anchored mammographic data from a screening-naive population. The next release (IMDB r2.0) expands the resource through inclusion of an independently curated cohort acquired between August and December 2025. At our tertiary care center, an opportunistic screening model is followed for performing mammography among women diagnosed with other cancers, especially those with a hormonal predilection like ovarian and endometrial cancers. The dataset captures the demographic, pathological, and imaging characteristics of an Indian diagnostic mammography population, including a higher prevalence of dense breasts, younger patients, diverse benign lesions, and clinically detected cancers. By making this resource publicly available, we aim to facilitate the development, benchmarking, and external validation of AI algorithms across populations that remain largely absent from existing mammography datasets.

2. Summary

The vision of IMDB is to establish a sustainable, openly accessible, and clinically grounded resource that enables the development, validation, and benchmarking of artificial intelligence systems across diverse populations while promoting reproducible and equitable breast imaging research. The dataset is intended for researchers in medical imaging, machine learning, computer vision, radiology, and breast cancer research.

IMDB is AI-ready and was curated as part of a multicenter clinical evaluation program for an AI-based mammography interpretation system. The dataset comprises deidentified mammography examinations in native DICOM format, preserving full spatial resolution, full image contrast, and acquisition metadata required for advanced image analysis and AI development. Each examination is accompanied by structured metadata, including demography, BI-RADS assessments, breast density information, pathology results where available, and imaging follow-up outcomes. Ground-truth labels were established using histopathological confirmation and longitudinal imaging follow-up whenever available, providing clinically meaningful endpoints for algorithm development and validation. The dataset supports a broad range of machine learning applications, including breast cancer detection, classification, risk prediction, multimodal learning, domain adaptation, foundation model development, and external benchmarking. Access to the resource requires standard DICOM-compatible software and appropriate computational infrastructure for image analysis.

3. Discussion

The Indian Mammography Database (IMDB) addresses an important gap in publicly available breast imaging resources by providing a large-scale mammography dataset derived from a screening-naive Indian population. A major strength of the dataset is the availability of clinically meaningful ground-truth information. Suspicious lesions were linked to histopathological outcomes whenever available, while benign examinations were supported by imaging follow-up or, in IMDB r1.0, structured telephonic follow-up when imaging follow-up records were unavailable. In addition,

the dataset includes structured radiologist assessments and breast density information, enabling a wide range of applications including classification, breast density assessment, risk prediction, multimodal learning, and foundation model development. The dataset also incorporates examinations acquired on two generations of mammography systems from the same vendor (Hologic Selenia platforms). While maintaining consistency in image acquisition philosophy, the inclusion of multiple hardware generations introduces technical variability and may improve the robustness of algorithms developed using the dataset. The initial release, IMDB r1.0, contained 3,577 mammography images from 583 patients. IMDB r2.0 substantially expands the resource to 9,097 mammography images from 2,636 patients, increasing both the scale and diversity of available examinations.

Several limitations should be acknowledged. First, the dataset remains a single-institution resource and may not fully represent variations across different regions of India. Second, data curation required manual linkage of imaging studies, pathology reports, and clinical records because automated electronic health record integration was not available during much of the study period. This increased the effort required for dataset construction and may have limited the scale of earlier releases. In addition, long-term follow-up information is not uniformly available for all patients. This reflects real-world clinical practice in India, where patients frequently seek care across multiple institutions or may not return for longitudinal surveillance. Although pathology-confirmed outcomes are available for suspicious lesions whenever feasible, incomplete follow-up remains an inherent limitation of retrospective dataset curation in this setting.

Future releases will focus on further expanding both the scale and diversity of the resource. Planned enhancements include incorporation of additional institutions participating in our multicenter clinical evaluation program, improved integration with electronic clinical records, expansion of longitudinal follow-up data, and inclusion of digital breast tomosynthesis examinations as archival workflows mature. We also plan to incorporate data from the National Cancer Institute (NCI), Jhajjar, which participates in the broader multicenter clinical trial framework associated with this project in near future. These developments will further strengthen the geographic diversity, clinical representativeness, and research utility of IMDB.

Compared with widely used mammography datasets such as CBIS-DDSM, INbreast, and VinDr-Mammo, IMDB provides complementary value through its representation of an Indian screening-naive population, pathology-linked outcomes, real-world diagnostic imaging practice, and multimodal mammographic data.

4. Resource Availability

The Indian Mammography DataBase (IMDB) is publicly available through the Indian Biological Images Archive (IBIA) under an Open Access licence, and can be downloaded without a user account or login. To download the data, visit the IBIA browse page (`https://ibdc.dbt.gov.in/ibia/studybrowse/`), enter "mammo" in the search box, and click the **Data** button (download icon) on the two IMDB study cards to obtain the DICOM images and the accompanying CSV metadata file.

The initial release, IMDB r1.0 (Study Accession: MAMOS_1000000004), has its study page at `https://ibdc.dbt.gov.in/ibia/study_details_browse/MAMOS_1000000004/`. The expanded release, IMDB r2.0 (Study Accession: MAMOS_1000000050), has its study page at `https://ibdc.dbt.gov.in/ibia/study_details_browse/MAMOS_1000000050/`. The dataset is also available on the Medical Imaging Data Analysis System (MIDAS) platform at `https://www.midas.iisc.ac.in/fe/datasets/breast/breast-mammography`, where access is granted upon request.

The study pages linked above provide the complete study-level metadata for each release; no account creation is required to download from the IBIA repository.

The current release comprises two versions: IMDB r1.0 (583 patients; 3,577 images) and IMDB r2.0 (2,636 patients; 9,097 images). Together, these versions provide a resource containing 3,219 unique patients and 12,674 mammography images. Future releases will be assigned unique version identifiers and accompanied by release notes documenting additions, corrections, and metadata updates.

The dataset consists of de-identified Full-Field Digital Mammography (FFDM) and Synthesized Mammography (SM) examinations acquired at the AIIMS, New Delhi, India. Associated metadata include patient age, breast density, BI-RADS assessment category, follow-up status, and pathology-confirmed outcomes where available.

Access and reuse of the dataset are governed by the policies of the IBIA, the Indian Biological Data Centre (IBDC), and the Department of Biotechnology (DBT) BIOTECH-PRIDE guidelines and Framework for Exchange of Data (FeED) protocols; Department of Biotechnology, Govt. of India (2021). As confirmed in writing by the IBIA/IBDC repository, this Open Access licence category is comparable to the Creative Commons Attribution model (CC BY), permitting reuse, redistribution, and modification with attribution and without registration or access controls. The dataset is additionally hosted on MIDAS under an explicit Creative Commons Attribution 4.0 International (CC BY 4.0) licence (https://creativecommons.org/licenses/by/4.0/). The full licence terms are publicly accessible through the repository and the cited BIOTECH-PRIDE guidelines. The

authors confirm that all released data were generated at the source institution, are fully de-identified, and may be publicly disseminated, analyzed, and redistributed in accordance with the repository's Open Access policies and applicable ethical and legal requirements.

This retrospective dataset was approved by the Institutional Ethics Committee, AIIMS, New Delhi, India, under approval number AIIMSA1985/06.08.2024. The dataset does not contain direct patient identifiers, and all DICOM metadata underwent anonymization prior to release using custom Python-based de-identification procedures, hence informed consent was waived. The released resource includes adult patients ($\geq$18 years of age) and does not contain pediatric data. Data approved for public dissemination include de-identified mammography images and associated metadata described in this manuscript.

Researchers should be aware that the dataset represents a screening-naive Indian population from a tertiary-care referral center and may therefore not be fully representative of organized population-based screening programs. Potential misuse includes overestimating model generalizability to other geographic regions, healthcare systems, or demographic groups without appropriate external validation. Users are encouraged to evaluate fairness, bias, and clinical applicability before deploying models developed using this resource.

Questions regarding dataset access, licensing, ethical considerations, or future releases may be directed to Dr. Krithika Rangarajan (Department of Oncoradiology, Dr. B.R. Ambedkar Institute Rotary Cancer Hospital, AIIMS, New Delhi), at krithikarangarajan86@gmail.com.

5. Methods

5.1 Data Source and Dataset Description

The Indian Mammography Database (IMDB) was curated from mammography examinations performed at two breast imaging units within a tertiary-care academic hospital in India: the Department of Radiology, AIIMS, New Delhi and the Department of Oncoradiology, Dr. B.R. Ambedkar Institute Rotary Cancer Hospital, AIIMS, New Delhi. All examinations were acquired on two generations of Hologic Selenia mammography systems.

IMDB r1.0 contains 3,577 mammography images from 583 patients acquired between January 2020 and December 2020. IMDB r2.0 expands the resource to 9,097 mammography images from 2,636 patients acquired between August 2025 and December 2025. Together, these releases provide a large mammography resource derived from a screening-naive Indian population undergoing diagnostic or opportunistic screening mammography. Detailed demographic and clinical characteristics are summarized in Table 1. The parallel curation pipelines for both releases and their combined statistics are illustrated in Figure 1.

5.2 Dataset Curation

Women aged 18 years or older who underwent mammography during the study period and had at least one available standard mammographic view were eligible for inclusion. Patients undergoing diagnostic mammography, opportunistic screening mammography, post-treatment surveillance examinations, and patients with prior breast surgery were included.

Examinations performed exclusively for specimen imaging or mammography-guided procedures, including stereotactic biopsy and wire localization, as well as interim examinations, were excluded (total exclusions: r1.0 = 30; r2.0 = 323). Patients assigned BI-RADS 4 or BI-RADS 5 assessments without available histopathological confirmation were also excluded.

5.3 DICOM Extraction and Image Processing

All mammography examinations were retrieved directly from PACS in Digital Imaging and Communications in Medicine (DICOM) format. Both FFDM and synthesized mammography images generated during routine clinical workflow were included. Imaging metadata, including view position, laterality, acquisition parameters, and image type, were extracted directly from DICOM headers. All mammography images underwent de-identification prior to public release. Custom Python-based anonymization pipelines were used to remove protected health information from DICOM metadata while preserving acquisition-related information required for research applications. To promote transparency and reproducibility, the DICOM de-identification, image processing, and benchmarking scripts, are publicly available at `https://github.com/krithikarangarajan-coe/imdb_ibia`. Each examination was assigned a unique study identifier. Clinical metadata and imaging labels were stored separately in accompanying comma-separated value (CSV) files to facilitate downstream analysis and machine learning workflows.

5.4 Quality Assurance

A multi-stage quality assurance process was performed before dataset release. Automated metadata filtering was initially used to identify image type, view position, and acquisition characteristics. Duplicate examinations and duplicate images were identified and removed. Cases containing multiple images for the same breast view were manually reviewed. Spot magnification views, post-procedural mammograms, specimen radiographs, and other non-standard acquisitions were excluded from the final dataset. When

Table 1: Demographic and clinical characteristics of IMDB r1.0 and IMDB r2.0.

Characteristic	IMDB r1.0	IMDB r2.0	Total
Number of patients	583	2,636	3,219
Mean age ± SD (years)	50.0 ± 11.0	49.9 ± 10.3	—
Age distribution (years)			
20–29	13	18	31
30–39	85	364	449
40–49	194	1,033	1,227
50–59	172	736	908
60–69	95	360	455
70–79	20	117	137
80–89	4	8	12
Distribution of cancer			
Cancer	73	315	388
Non-cancer	510	2,311	2,821
Breast density			
Dense (C + D)	159	964	1,123
Non-dense (A + B)	731	1,672	2,403
BI-RADS distribution			
2 and below	777	2,224	3,001
3	40	146	186
4	18	69	87
5	42	83	125
6	13	110	123
Image-level data			
Number of images	3,577	9,097	12,674
MLO views	1,786	4,536	6,322
CC views	1,791	4,561	6,352

multiple standard-view images were available for the same projection, the image with superior technical quality and diagnostic completeness was retained.

Additional quality-control procedures included verification of image-view assignments, metadata consistency, correspondence between radiology reports and pathology records, and confirmation of successful anonymization.

5.5 Ground-Truth Determination

Radiological interpretation was performed using a structured multi-reader workflow. All mammograms underwent an initial independent review by two trainee radiologists, who annotated the mammographic abnormalities. Subsequently, the examination was reviewed by a consultant breast radiologist who had access to supplemental breast imaging modalities such as ultrasound or MRI when available, and integrated all available information to establish the reference standard interpretation. As an additional step, a second senior breast radiologist independently reviewed the annotations and assigned labels. In cases of discrepancy, the final annotation was established through expert adjudication by the senior breast radiologists, ensuring consistency and accuracy of the dataset labels.

Ground-truth labels were established using a composite reference standard. BI-RADS 1 and BI-RADS 2 examinations were categorized as negative or benign. BI-RADS 3 examinations were assigned outcomes using available follow-up imaging and clinical records, acknowledging that longitudinal follow-up was not available for all patients due to variable follow-up adherence in routine clinical practice. For BI-RADS 4 and BI-RADS 5 examinations, histopathological diagnosis obtained from image-guided biopsy or surgical specimens served as the reference standard. Cases without available histopathological confirmation were excluded from the final dataset. The composite ground-truth determination framework is summarized in Figure 2. Overall workflow is summarized in Figure 3.

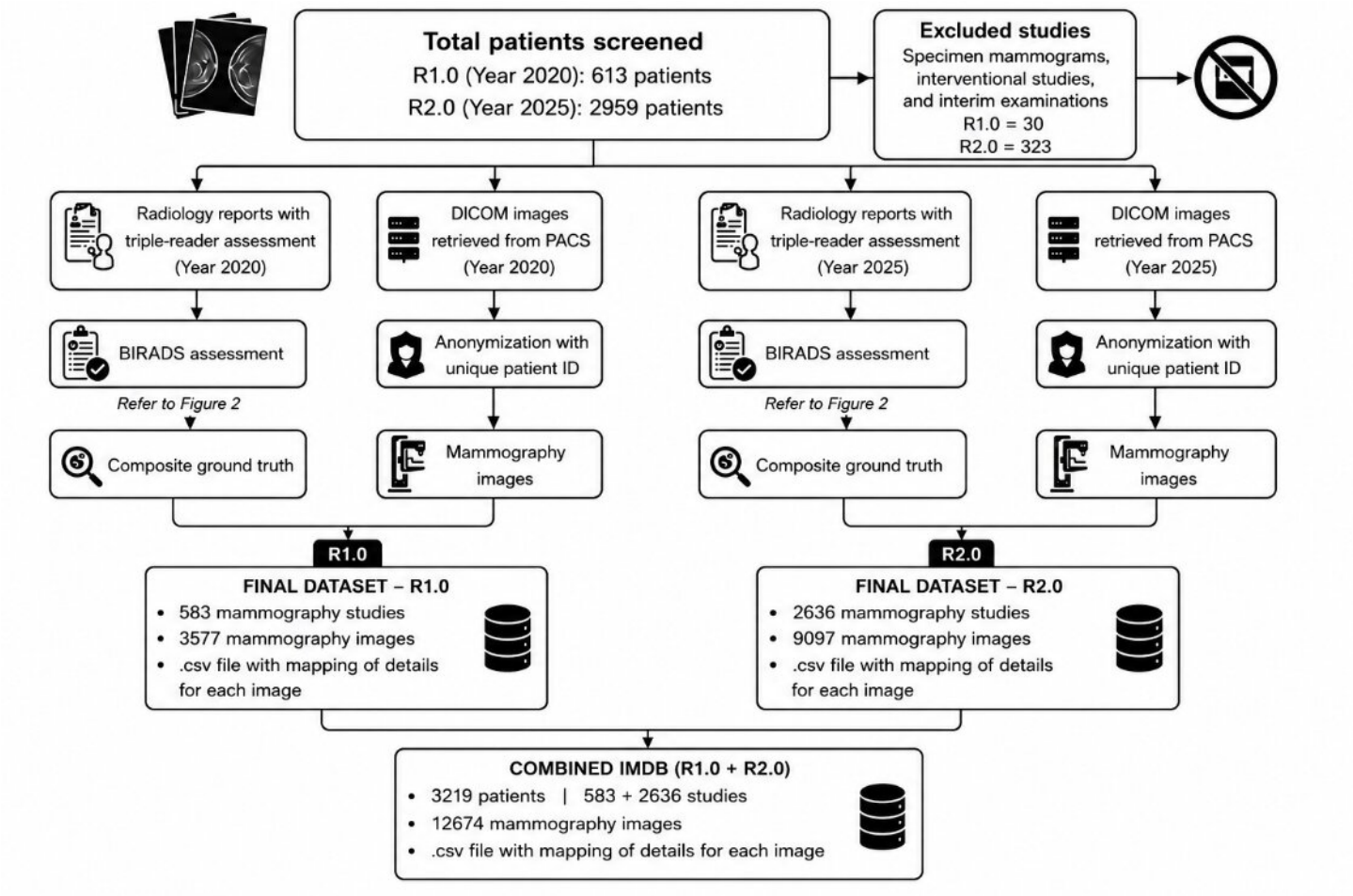

Figure 1: Dataset curation flowchart for IMDB r1.0 (year 2020) and IMDB r2.0 (year 2025), illustrating parallel processing pipelines from mammography acquisition through PACS retrieval, anonymization, and composite ground-truth assignment, along with the combined dataset statistics.

6. Validation

Dataset quality was ensured through a multi-stage curation workflow. All mammography examinations underwent automated metadata validation and manual review to verify image type, view position, laterality, and metadata consistency. Duplicate examinations, specimen radiographs, post-procedural mammograms, spot magnification views, and other non-standard acquisitions were excluded. All DICOM files were de-identified prior to release, and correspondence between imaging studies, radiology reports, pathology records, and follow-up data was verified through manual review.

To demonstrate the usability of the dataset for artificial intelligence research, benchmark experiments were performed on both IMDB r1.0 and r2.0 using multiple pre-trained CNN and YOLO architectures for binary classification of malignant and non-malignant mammography images (Table 2). Among the evaluated models, YOLOv9-s demonstrated the strongest baseline performance on both releases, confirming that the benchmarking approach transfers directly to the larger cohort. A YOLOv11-nano model, transfer-learned from the YOLOv9-s weights on class-balanced subsets of r1.0 (malignant-to-benign ratios from 1:1 to 1:7, each split 80:10:10), achieved the highest test accuracy of 89.6% at a 1:7 ratio. These results demonstrate that the dataset supports development of robust machine-learning models and can serve as a benchmark resource for future algorithm development, validation, and comparative evaluation studies.

Acknowledgments

The authors acknowledge the Department of Radiology, AIIMS New Delhi, and the Department of Oncoradiology, IRCH, AIIMS New Delhi, for their technical and clinical support. We thank Dr. Radhika Rajeev for her assistance during the early stages of dataset curation. The r1.0 was supported by Ministry of Education (T-316) and r2.0 by Translational AI for Networked Universal Healthcare (TANUH) foundation. We also acknowledge the Indian Biological Images Archive (IBIA) and the Indian Biological Data Centre

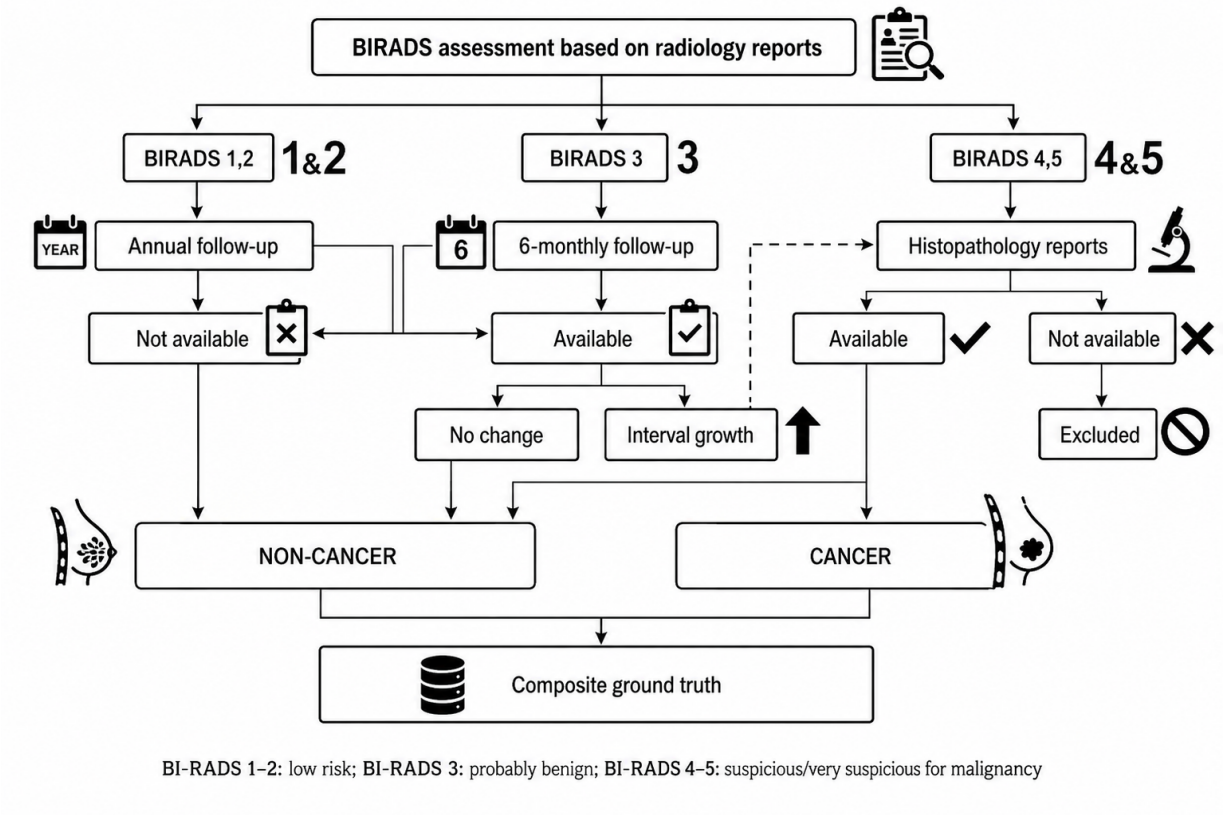

Figure 2: Ground-truth determination framework based on BI-RADS assessment category. BI-RADS 1–2 examinations are categorized as non-cancer. BI-RADS 3 examinations are assigned outcomes based on 6-monthly imaging follow-up. BI-RADS 4–5 examinations require histopathological confirmation; cases without available pathology are excluded.

Table 2: Model performance metrics for classification task on our dataset for pre-trained CNN and YOLO models. Each cell reports **r1.0 / r2.0** values. TP: true positive; FP: false positive; TN: true negative; FN: false negative. Models used to check concordance and discordance of predictions are marked in bold. r1.0 results are from Terzi et al. (2025); r2.0 results were obtained by applying the same pre-trained models to IMDB r2.0.

Model	TP	FP	TN	FN	Accuracy	Precision	Recall/Sensitivity	F1 Score	Specificity
DoubleHead R-CNN	226 / 535	731 / 2304	2546 / 5534	74 / 724	0.77 / 0.67	0.24 / 0.19	0.75 / 0.42	0.36 / 0.26	0.78 / 0.71
Faster R-CNN	234 / 583	862 / 2625	2415 / 5213	66 / 676	0.74 / 0.64	0.21 / 0.18	0.78 / 0.46	0.34 / 0.26	0.74 / 0.67
FCOS	234 / 1074	861 / 6381	2416 / 1457	66 / 185	0.74 / 0.28	0.21 / 0.14	0.78 / 0.85	0.34 / 0.25	0.74 / 0.19
RetinaNET	236 / 1066	983 / 6425	2294 / 1413	64 / 193	0.71 / 0.27	0.19 / 0.14	0.79 / 0.85	0.31 / 0.24	0.70 / 0.18
DETR	287 / 1079	2585 / 6516	692 / 1322	13 / 180	0.27 / 0.26	0.10 / 0.14	0.96 / 0.86	0.18 / 0.24	0.21 / 0.17
Dynamic R-CNN	285 / 582	2584 / 2660	693 / 5178	15 / 677	0.27 / 0.63	0.10 / 0.18	0.95 / 0.46	0.18 / 0.26	0.21 / 0.66
ATSS	289 / 1083	2618 / 6665	659 / 1173	11 / 176	0.26 / 0.25	0.10 / 0.14	0.96 / 0.86	0.18 / 0.24	0.20 / 0.15
VarifocalNET	283 / 1242	2621 / 7731	656 / 107	17 / 17	0.26 / 0.15	0.10 / 0.14	0.94 / 0.99	0.18 / 0.24	0.20 / 0.01
YOLOv3	299 / 620	3236 / 2817	41 / 5021	1 / 639	0.10 / 0.62	0.08 / 0.18	0.99 / 0.49	0.16 / 0.26	0.01 / 0.64
YOLOv9-c	**188 / 617**	**402 / 1381**	**2875 / 6457**	**112 / 642**	**0.86 / 0.78**	**0.32 / 0.31**	**0.63 / 0.49**	**0.42 / 0.38**	**0.88 / 0.82**
YOLOv9-m	**186 / 631**	**443 / 1515**	**2834 / 6323**	**114 / 628**	**0.84 / 0.76**	**0.30 / 0.29**	**0.62 / 0.50**	**0.40 / 0.37**	**0.86 / 0.81**
YOLOv9-s	**193 / 610**	**388 / 1335**	**2889 / 6503**	**107 / 649**	**0.86 / 0.78**	**0.33 / 0.31**	**0.64 / 0.48**	**0.44 / 0.38**	**0.88 / 0.83**

*Models used to check concordance and discordance of predictions are marked in bold.

(IBDC), MIDAS (supported by the Indian Council for Medical Research (Sanction No. BMI/ICMR Imagebank/2021)) for hosting and disseminating the dataset.

Ethical Standards

This dataset was approved by the Institutional Ethics Committee of AIIMS, New Delhi, India under approval number AIIMSA1985/06.08.2024.

Conflicts of Interest

The authors declare no conflicts of interest.

Data availability

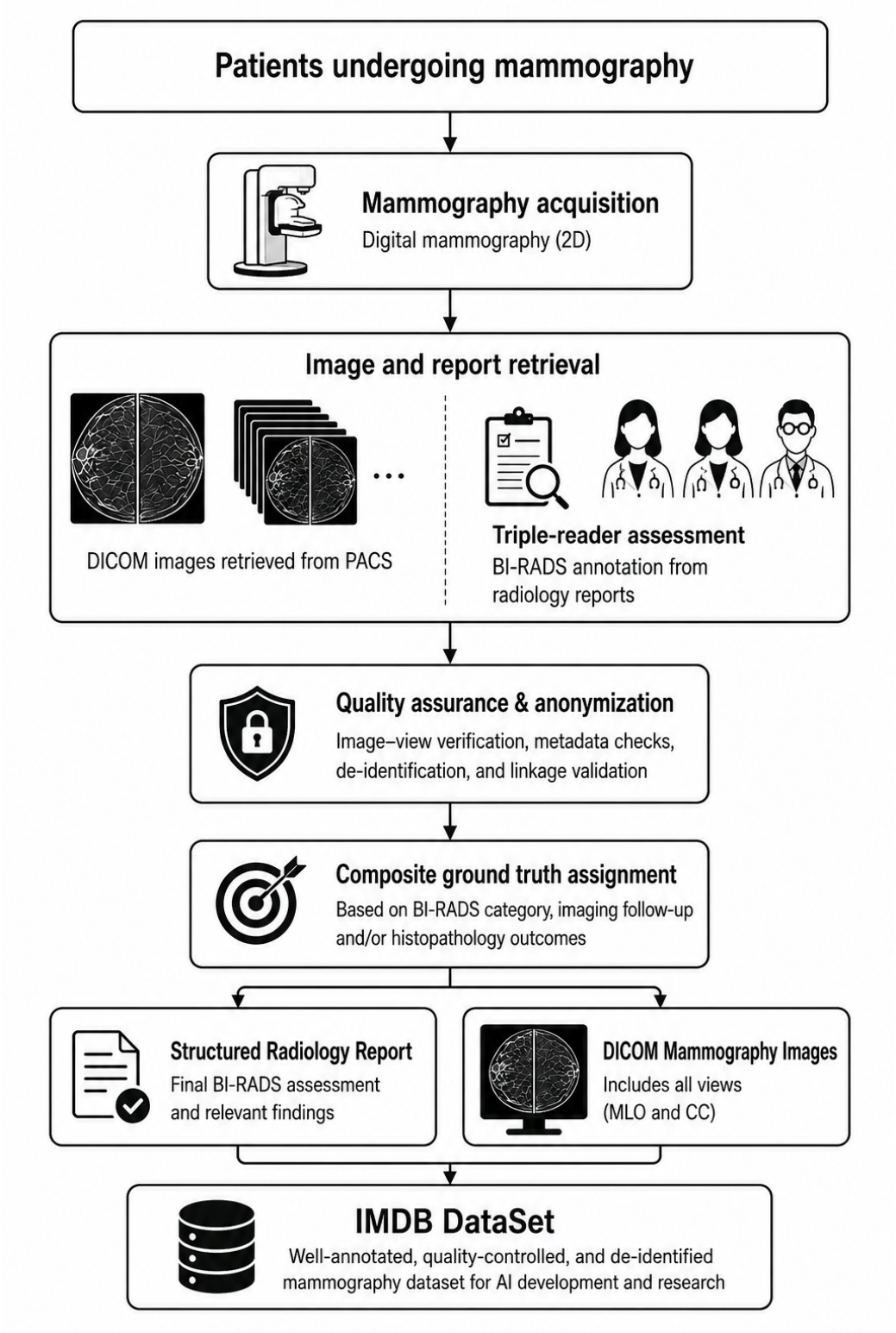

Figure 3: Overview of the mammography workflow in the radiology department and data collection for the dataset.

The dataset is publicly available through the Indian Biological Images Archive (IBIA) at `https://ibdc.dbt.gov.in/ibia/study_details_browse/MAMOS_1000000004/` (IMDB r1.0) and `https://ibdc.dbt.gov.in/ibia/study_details_browse/MAMOS_1000000050/` (IMDB r2.0), and through the MIDAS platform at `https://www.midas.iisc.ac.in/fe/datasets/breast/breast-mammography`.

References

Saurabh Bobdey, Asha Jain, and Govind Balasubramanium. Epidemiological review of laryngeal cancer: An Indian perspective. *Indian Journal of Medical and Paediatric Oncology*, 36(3):154–160, 2015. .

Chih-Hao Chen, Yen-Shen Lu, Ann-Lii Cheng, Chi-Sheng Huang, Wen-Hung Kuo, Mei-Yin Wang, et al. Disparity in tumor immune microenvironment of breast cancer and prognostic impact: Asian versus Western populations. *The Oncologist*, 25(1):e16–e23, 2020. .

René Aloisio da Costa Vieira, Gustavo Biller, Gilberto Uemura, Carlos Alberto Ruiz, and Maria Paula Curado. Breast cancer screening in developing countries. *Clinics*, 72(4):244–253, 2017. .

Kristina Dembrower, Emma Wåhlin, Yue Liu, Mattie Salim, Kevin Smith, Peter Lindholm, et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. *Lancet Digital Health*, 2(9):e468–e474, 2020. .

Kristina Dembrower, Alessio Crippa, Elin Colón, Martin Eklund, Fredrik Strand, and ScreenTrustCAD Trial Consortium. Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study. *Lancet Digital Health*, 5(10):e703–e711, 2023. .

Department of Biotechnology, Govt. of India. Biotechpride (promotion of research and innovation through data exchange) guidelines. Technical report, Department of Biotechnology, Ministry of Science and Technology, Government of India, New Delhi, July 2021. URL `https://dbt.gov.in/storage/media/rulesregulation/Biotech%20PRIDE%20Guidelines.pdf`.

Flavia Z. Francies, Rodney Hull, Richard Khanyile, and Zodwa Dlamini. Breast cancer in low-middle income countries: abnormality in splicing and lack of targeted treatment options. *American Journal of Cancer Research*, 10(5):1568–1591, 2020. PMID: 32411462.

Hannah M. L. Frazer, Jessica S. N. Tang, Matthew S. Elliott, Kathryn M. Kunicki, Benjamin Hill, Rakshith Karthik, et al. ADMANI: Annotated digital mammograms and associated non-image datasets. *Radiology: Artificial Intelligence*, 5(2):e220072, 2023. .

Mark D. Halling-Brown, Lucy M. Warren, Dominic Ward, Emma Lewis, Alasdair Mackenzie, Matthew G. Wallis, et al. OPTIMAM mammography image database: A large-scale resource of mammography images and clinical data. *Radiology: Artificial Intelligence*, 3(1):e200103, 2021. .

Kristina Lång, Magnus Dustler, Viktoria Dahlblom, et al. Artificial intelligence-supported screen reading versus standard double reading in the mammography screening with artificial intelligence (MASAI) trial: a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncology*, 24(8):936–944, 2023. .

Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Maayan Gorovoy, and Daniel L. Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data*, 4:170177, 2017a. .

Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, and Daniel L. Rubin. Curated breast imaging subset of DDSM. *Scientific Data*, 4:170177, 2017b. .

Srishti Malvia, Sunil A. Bagadi, Usha S. Dubey, and Sunita Saxena. Epidemiology of breast cancer in Indian women. *Asia-Pacific Journal of Clinical Oncology*, 13(4):289–295, 2017. .

Maciej A. Mazurowski, Mateusz Buda, Ashirbani Saha, and Mustafa R. Bashir. EMBED: A large-scale public database of screening mammography and clinical outcomes. *Radiology: Artificial Intelligence*, 4(6):e220047, 2022. .

Inês C. Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria João Cardoso, and Jaime S. Cardoso. INbreast: Toward a full-field digital mammographic database. *Academic Radiology*, 19(2):236–248, 2012. .

Hieu Trung Nguyen, Ha Quy Nguyen, Hieu Huy Pham, et al. VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data*, 10(1):277, 2023. .

Yu Shen, Nan Wu, Jason Phang, et al. An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Medical Image Analysis*, 68:101908, 2021. .

Mun Kun M. Tea, Lu Fan, James W. Delancey, Christian Staudigl, Stefan Steurer, Christine Lang, et al. Is breast cancer in young Asian women more aggressive than in Caucasians? A cross-sectional analysis. *Tumour Biology*, 34(4):2379–2382, 2013. .

Ramazan Terzi, Ahmet Enes Kılıç, Gökhan Karaahmetoğlu, and Okan Bilge Özdemir. The digital eye for mammography: deep transfer learning and model ensemble based

open-source toolkit for mass detection and classification. *Signal, Image and Video Processing*, 19(1):170, 2025. .

Adam Yala, Tal Schuster, Randy Miles, et al. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*, 303(1):60–71, 2022. .