Jaradat JH, Amro R, Hamamreh R, Musleh A, Abdelgalil MS. From Data to Diagnosis: Narrative Review of Open-Access Mammography Databases for Breast Cancer Detection. *High Yield Medical Reviews*. Published online June 1, 2024.
doi:10.59707/hymrPFNZ8344

Online first

From Data to Diagnosis: Narrative Review of Open-Access Mammography Databases for Breast Cancer Detection

Jaber H. Jaradat[1], Raghad Amro[1], Rawan Hamamreh[2], Ayman Musleh[3], Mahmoud Shaaban Abdelgalil[4]

[1] Mutah University, [2] Hashemite University, [3] University of Jordan, [4] Ain Shams University

Keywords: Mammograms, Open-access databases, Machine Learning, Deep Learning, Breast tumors, Breast Cancer

https://doi.org/10.59707/hymrPFNZ8344

High Yield Medical Reviews

Introduction: Breast cancer remains a significant global health challenge, necessitating advancements in screening and diagnostic methods for its early detection and treatment. This review explores the role of open-access mammography databases in facilitating research and development in the field of breast cancer detection, particularly through the integration of artificial intelligence techniques, such as machine learning and deep learning.

Methods: We conducted a comprehensive literature search to identify open-access databases related to mammography. For each database, we collected descriptive data including the number of images, types of lesions, and associated clinical metadata.

Results: A total of six databases were identified, including the Digital Database for Screening Mammography (DDSM), Curated Breast Imaging Subset of DDSM (CBIS-DDSM), Mini-DDSM, INbreast, Mammographic Image Analysis Society Dataset (MIAS), and China Mammography and Mastopathy Dataset (CMMD). A narrative synthesis of each database was pursued, and analyzed in terms of its composition, features, limitations, and contributions to breast cancer research. In addition, we highlight the importance of open-access databases in enabling collaborative research, improving algorithm development, and enhancing the accuracy and efficiency of breast cancer detection methods and computer-aided diagnosis.

Conclusion: This review highlights the significance of open-access mammography databases in the advancement of computer-aided diagnosis of breast cancer. By providing access to large and diverse datasets, these databases play a crucial role in accelerating research progress, fostering innovation, and ultimately improving outcomes in patients with breast cancer.

INTRODUCTION

Breast cancer (BC) is a formidable health challenge, and its prevalence and incidence make it the most widespread cancer among females globally.[1,2] Despite great improvements in diagnosis, screening, and therapeutic approaches, the incidence of breast cancer continues to rise, accounting for 36% of tumor cases, and is the fifth leading cause of cancer-related deaths worldwide.[1,2] Therefore, it is crucial to improve breast cancer screening programs and improve early detection, diagnosis, and treatment to reduce its aggressiveness and hinder its widespread.[2] In the early 1900s, the diagnosis of BC was mainly clinical, with consequent delayed diagnosis and poor prognosis. The development of non-invasive BC diagnostic techniques in recent years has been impressive, with mammography being the most widely used imaging modality. Mammography best detect lesions in women aged 50 and older, with average sensitivity and accuracy of 85%. If a suspicious lesion is found, further imaging modalities are indicated and with last confirmatory step being biopsy.[2] The increase in BC prevalence and incidence is attributed to increased organized efforts in breast cancer screening programs, with a consequently more favorable prognosis and a great reduction in mortality-related breast cancers.[3,4] Diversity in breast density has led to fluctuations in the accuracy of mammography screening programs and has made human contributions in classifying breast lesions susceptible to human errors. Therefore, the development of new techniques to improve breast lesion detection and classification is essential, such as contrast-enhanced mammography (CEM), artificial intelligence (AI), and radiomics, which show promise for earlier detection.[3]

Machine learning (ML) and deep learning (DL) are subsets of artificial intelligence (AI) that have shown immense potential in the field of medicine. Harnessing AI can help in diagnosis, disease detection, treatment selection, and patient monitoring, and enable more accurate and efficient health delivery, leading to improved affordability and quality of care.[5] AI requires huge datasets to be trained on, which makes medicine a great option, given the large patient databases of various file types: texts, audio, images, tables, and videos. A Convolutional Neural Network (CNN) is a class of artificial neural networks and a subclass of DL,

which master processing data that has a grid pattern, such as images designed to automatically and adaptively learn spatial hierarchies of features.[6] However, owing to their vast capability, it is essential to address two main challenges faced by CNN: small datasets and overfitting.[6] In this review, we address one of the main challenges of the CNN algorithm and highlight the immediate open-access datasets (no registration required) for mammography, thus enhancing and encouraging the integration of AI in medicine to improve patient care quality. Our research focused on analyzing open-access mammography databases, specifically determining the available datasets, and providing a descriptive analysis of these datasets.

Open-Access Databases (OAD) are a crucial data source for either model development or testing. Even when models are initially trained on private datasets, open databases serve as additional validation sources. OAD may be more trustworthy, and their results are more reliable than private datasets because public databases are created and made available to everyone; therefore, they are less susceptible to researchers' bias. The global shift towards full and immediate Open Access in academic publishing is gaining traction despite initial challenges and resistance. This shift not only maintains comparable performance, but also enhances accuracy, particularly in specific demographic groups. Full and immediate Open Access is gaining momentum across most developed countries despite some difficulties and resistance in the process of moving away from traditional subscription publishing.[7] Over the past decade, the number of well-cited open-access articles by numerous non-English researchers, supported by international grants for advanced fields of science, has multiplied.[8]

DIGITAL DATABASE FOR SCREENING MAMMOGRAPHY

The Digital Database for Screening Mammography (DDSM) is a comprehensive collection of digital mammograms and associated data that has been widely used for research and development of computer-aided detection and diagnostic techniques in the field of breast cancer.[9] It contains thousands of mammographic images along with detailed clinical information for each image.

DDSM was first developed as a collaborative effort among several research institutions, namely the University of Florida, Massachusetts General Hospital, and Sandia National Laboratories in the late 1990s. Since then, it has been continuously updated and expanded, and has become a widely used database. The images in the database were derived from various institutions. The aim was to make standardized and publicly available dataset for researchers and developers in the field of breast cancer imaging.[9]

The database consists of 2620 cases, with each case being a collection of images and information corresponding to one mammogram examination session (Table 1). The cases were categorized as normal cases where no further workup was required, cancer cases where a minimum of one pathology-proven cancer was found, benign cases, suspicious findings, were later determined as non-malignant by biopsy or ultrasound, or benign without callback cases, which are benign cases that did not have any additional films or biopsies taken. The distribution of the cases is as follows: 695 normal, 914 cancer, 870 benign, and 141 benign without callbacks.[9]

Despite the benefits of the DDSM database, challenges emerges as researchers use it due to its large size, prompting the development of subsets of the database aimed at improving and curating it to be more useful. These subsets include the Curated Breast Imaging Subset of DDSM (CBIS-DDSM) released by Lee et al.[10] and the Mini-DDSM released by Lekamlage et al.[11]

THE CURATED BREAST IMAGING SUBSET OF DDSM (CBIS-DDSM)

Many researchers and developers have chosen DDSM database owing to the well-documentation. Due to the large size of the database and the limited region of interest (ROI) annotations in the old version, this prompt CBIS-DDSM creators to form this as a subset of the DDSM addressing and resolving the issues with the previous database. Moreover, they curate the database, double check the mammographs' report by two separate mammographer, changed the image attribute to the Digital Imaging and Communications in Medicine (DICOM) format which is the standard format for medical imaging, updated the ROI masks, and ensured precise delineation of masses from the surrounding tissues.[10,12]

In the process of updating the DDSM, the images underwent decompression and they were processed, with some cropped to emphasize abnormalities and the region around the ROI.[10] Consequently, among the 9671 images labeled with an image description in the CBIS-DDSM, 36.9% were cropped images, 33.6% were ROI mask images, and 29.5% were full mammogram images (Table 1).

The CBIS-DDSM database encompasses 10,239 images. The percentage distribution of laterality (left or right breast), view (craniocaudal (CC) or mediolateral oblique (MLO)), and abnormality type (mass or calcification) are presented in supplementary tables 2 & 3.

It is categorized into two types of abnormalities: calcifications and masses. Each abnormality category included both testing and training sets. With 1566 unique patients and 3,069 unique images, a single mammogram image for a patient may contain no findings, one finding, or multiple findings.[13] Consequently, among the 1566 patients, 1872 had calcifications and 1696 had mass findings.

The training and testing sets in CBIS-DDSM provide various types of information for each image, including the density category according to the American College of Radiology (ACR) standards, laterality, view (CC or MLO), number of abnormalities for the image, mass shape and margin (when applicable), calcification type and distribution (when applicable), Breast Imaging Reporting and Data Systems (BI-RADS) assessment, pathology status, and subtlety rating: radiologists' rating of difficulty in viewing the abnormality in the image (Supplementary table 4). BI-RADS score is as follows: 0 for incomplete examinations, 1 indi-

cating no abnormalities, 2 as benign, 3 as probably benign, 4 as suspicious, 5 as highly indicative of malignancy, and 6 as confirmed cancer through biopsy.[14]

The CBIS-DDSM dataset has been used in multiple studies, with some studies using the entirety of the dataset[15, 16] and others using only mass findings[17] or calcifications.[18] The studies involving CBIS-DDSM mainly focus on training machine learning algorithms on mammogram segmentation and detection of abnormalities,[19] with the most used algorithm being CNN.[10,20] However, the downsizing of CBIS from the original DDSM dataset led researchers to combine it with other datasets.[20,21] The dataset has proven to be helpful for many researchers because of its large size, well documentation and high-resolution images.

MINI-DDSM

A group of researchers from Sweden aimed to develop a tool for age estimation using mammography, such as the way X-ray images of the hand[22] or teeth[23] are used to estimate a patient's age. Throughout the development process, they encountered challenges owing to the scarcity of freely available public mammography datasets that included the age attribute for each mammogram image. At the time of the development of Mini-DDSM, the DDSM database was the sole dataset with attached age information. However, images in the DDSM database were compressed using lossless JPEG encoding generated by an outdated software. Even though CBIS-DDSM served as an alternative host with a superior image format, the images in CBIS-DDSM lacked the age attribute.[11]

This limitation led researchers to the only viable option of down-sampling images from DDSM, and this process resulted in the creation of what they termed the Mini-DDSM, which consisted of a total of 9684 images (Supplementary table 1). The pathological status of the mini-DDSM images was classified into 2728 (28.2%) normal images, 3360 (34.7%) benign images, and 3596 (37.1%) cancer images.[11]

The publicly available Excel sheet containing information attributed to the images in the dataset provided details for only 7808 images. These images exhibited a more balanced distribution in terms of status, with 31%, 34%, and 35% for normal, benign, and malignant cases, respectively. Images were evenly divided between the left and right sides and between CC and MLO views (Supplementary tables 2 & 3).

The Mini-DDSM dataset is not as extensively utilized in published studies as the CBIS-DDSM.[24-26] This discrepancy can be attributed to the fact that CBIS-DDSM encompasses a more diverse range of information for each mammogram image as the Mini-DDSM doesn't contain information like the abnormality type, BI-RADS assessment, subtlety score, mass shape and margin, and calcification type and distribution (Supplementary table 4).

The INbreast dataset is a public database obtained from the Breast Center in Centro Hospitalar de Sao Joao (CHSJ), Porto.[27] The image resolution varies between 3328 × 4084 and 2560 × 3328 pixels depending on the compression plate used, which was determined by the patient's breast size during acquisition.[27]

The images were stored in DICOM format.[13,27] Notably, these images displayed unique intensity profiles that were distinct from the digitized film mammograms in the CBIS-DDSM dataset.[14] In the INbreast dataset, the intensity profiles reflect the characteristics of digital imaging technology. Digital mammography produces images through electronic detectors that convert X-ray photons into electrical signals, resulting in a different intensity profile compared with traditional film-based mammography.[27]

On the other hand, the CBIS-DDSM dataset contains digitized film mammograms, which were originally captured on traditional X-ray film and later digitized for research purposes.[14] Therefore, the disparity in intensity profiles between the two datasets arises from the underlying differences in the imaging technologies (digital vs. film) and image acquisition methods. This disparity creates a valuable opportunity to assess the performance of a comprehensive image classifier when applied across diverse mammography platforms.[14]

A total of 115 cases were compiled for the INbreast dataset, with 90 cases featuring two images (MLO and CC) for each breast (Table 1). The remaining 25 cases were from women who underwent mastectomy, and two views of only one breast were included, resulting in 410 images.[27] Notably, among the 90 cases with two images per breast, eight cases included images acquired at different times, indicating follow-up observations.[28]

The dataset comprises a variety of images, including normal mammograms, and those with abnormal mammograms depicting masses, calcifications, architectural distortions, asymmetries, and images with multiple finding.[13] Calcifications are prominently represented in this database (71%), mirroring real-world trends,[29] where they constitute the most frequent observations in mammography.[27,28] Approximately half of females have benign breast calcifications on mammography.[30,31] The distribution of characteristics such as imaging view (CC or MLO) and type of abnormality are described in supplementary tables 2 and 3. Breast abnormalities were categorized using the BI-RADS classification system.

The ROIs in the dataset are delineated using contour points specified in an XML (Extensible Markup Language (XML) file. Annotations for the contour of the pectoral muscle are provided.

The dataset contains detailed information for each mammogram, including the patient's age at the time of imaging, family history, ACR breast density, and BI-RADS classification. Biopsy results are available specifically for cases categorized as BI-RADS 3, 4, 5, and 6. Instances was not subjected to biopsy are categorized as benign. Breast density is a crucial characteristic, with dense breasts presenting challenges in mammography. Each image in our database included density information measured on the ACR standard scale (Supplementary table 4).

The revised image dataset is valuable for research and practical medical applications, particularly in educational settings. Its meticulous annotations represent a significant

enhancement over existing databases; however, its small size limit this improvement.

This improvement can inspire computer vision researchers to create more precise methods for lesion characterization and enhance the effectiveness of detection and malignancy classification algorithms. Although the dataset's wide variety of images poses a challenging task, it is crucial for the development of more robust computer-aided diagnosis (CAD) systems.[27]

MAMMOGRAPHIC IMAGE ANALYSIS SOCIETY DATASET

The Mammographic Image Analysis Society (MIAS) dataset is a publicly accessible collection of digital mammography.[32] Originally introduced in 1994 by a center in the United Kingdom. MIAS includes a diverse range of mammograms, varying in size from small images of 1600 × 4320 to extra-large images of 5200 × 4320 pixels. The dataset consisted of 322 mammogram images for 161 patients, in only one view (MLO) for both breasts right and left (Table 1).

The dataset is well-annotated and comes with a CSV file reporting the type of diagnosis as either benign or malignant. It divides images based on tissue density into fatty, fatty-glandular, or dense-glandular. Additionally, abnormalities are further classified into categories such as calcifications, masses (either well-defined, spiculated, or ill-defined), architectural distortion, asymmetry, and normal findings. Additionally, the dataset offers x- and y-image coordinates pinpointing the center of abnormalities, along with the approximate radius of a circle surrounding the abnormality.

When dealing with calcifications, the focus on the center locations and radii shifts from individual spots to the entire cluster. If calcifications appear dispersed over the entire image rather than localized, the standard approach of marking center locations and radii is abandoned, as it becomes irrelevant. All images are provided in Portable Gray Map (PGM) image format. Unlike the DICOM format, which is specialized for medical imaging, the PGM format is a storage format for grayscale images that is characterized by its simplicity, as it is designed to be easily interpretable and modifiable by software. However, it lacks the functionality to manage medical data which makes it less applicable these days when more complex and data-rich formats are in demand.

The dataset has a processed form called the Mini-MIAS database[32] which has undergone various modifications, as well as clipping or padding, resulting in a standardized size of 1024 × 1024 pixels for all images.

One of the main limitations of using MIAS is that it is outdated since it was provided 20 years ago,[33] along with the imbalance distribution, 209 normal, 62 benign and 51 malignant images. Furthermore, the images are only in one breast view (MLO), lacking the CC view (Supplementary tables 2 & 3), which limits its scalability and raises the risk of overlooking lesions. However, the MIAS and Mini-MIAS datasets are still being used actively in research to train and evaluate various machine learning and deep learning algorithms for detecting breast cancer.[34-37]

CHINA MAMMOGRAPHY AND MASTOPATHY DATASET

The China Mammography and Mastopathy Dataset (CMMD) is an open-access dataset of mammography images.[38] It was generated from two hospitals in China, the Sun Yat-sen University Cancer Center in Guangzhou, and the Nanhai Affiliated Hospital of Southern Medical University in Fushan, between July 2012 and January 2016, including more than 3700 mammograms from 1775 patients (Table 1). The images were stored in 8-bit grayscale in the DICOM format covering both CC and MLO views, with a resolution of (2294 × 1914) pixels.[38,39]

CMMD is divided into two subsets, CMMD1 and CMMD2. CMMD1 contain both benign and malignant images from 1026 patients (2214 images). CMMD2 include only malignant breast cancer images, but with more detailed molecular subtypes from 749 patients (1498 images). A unique feature of the CMMD dataset is that all included images were biopsy-confirmed cases.[38,40]

The image classification of the dataset included a thorough examination of each image by two experienced radiologists. The procedure began by selecting images of patients with breast lesions. These were then clinically assessed based on the MLO and CC views (Supplementary tables 2 & 3). Subsequently, biopsies were performed, and a pathological diagnosis was established. The last part involved evaluation of the surgical specimens by immunohistochemistry (IHC) to confirm the origin of the tumor and to determine the molecular subtypes for each case.

Each image carries a unique ID, structured as D1-0001 for CMMD1 and D2-0001 onward for CMMD2. Details such as the image side (right or left), the patient's age, and the nature of the abnormality (whether it is a calcification, mass, or both) were reported for all images in an excel (xlsx) file. In CMMD2, molecular subtypes (Luminal A, Luminal B, HER2-enriched and Triple-negative) were also reported. The authors of the CMMD dataset acknowledge certain limitations, such as a relatively modest sample size and the absence of marked ROIs.[38] However, the CMMD dataset is notable for its detailed pathological evaluation and immunohistochemical data. These features combined with high-resolution imaging make it a valuable resource for advancing breast cancer research and computer aided diagnosis.[39-41]

BREAST IMAGING MEASURES

The region of interest (ROI), a popular term in image datasets, refers to an area within an image, in our case a mammogram image, which contains the most important features for the analysis, diagnosis, or management. It is identified either using center coordinates with a circular radius surrounding it or by using contours that outline the ROI.[42,43] A binary mask is defined by two-pixel values; the

first is usually 0, which represents the background, whereas the second value (1 or 255) marks the ROI. This increases the accuracy of AI models and reduces the time and computational power required. Retrieval performance in large databases can be improved with the use of ROI-based feature extraction along global features of the images, which are shown to be more effective in reflecting image-(or patient-) specific interests.[42]

SUMMARY

While CBIS-DDSM dataset is well-documented, large size, with high resolution and ROI masks, it lacks information about the age. Offering these advantages made it a common choice for many researchers, thus it has been used in over 70 studies, with great accuracy, precision, and area under the curve (AUC). CBIS-DDSM is suitable for training and validating deep-learning algorithms for various tasks in breast cancer detection and diagnosis. Researchers can utilize this dataset to develop computer-aided detection systems capable of accurately identifying and classifying abnormalities, such as masses and calcifications. The availability of ROI masks makes them particularly useful for studies focusing on precise lesion segmentation and feature extraction, leading to improved diagnostic accuracy.

Mini-DDSM dataset is of small size and each image comes with age information; however, its small size may make not representative and hard to train on, making it susceptible to overfitting. However, it is well documented, it is not documented as good as CBIS-DDSM. The absence of ROI masks along with limited documentation made it not a common choice for researchers. Researchers interested in age-related analyses or age-prediction models can leverage this dataset to train and validate their algorithms. However, its small size may pose challenges for generalizability, and researchers should be cautious about overfitting when using this dataset for model training.

The use of INbreast dataset is very limited due to its small size; however, it can be used to evaluate your model performance trained on larger datasets. Researchers can use this dataset for comparative studies, comparing the performance of their models against those trained on larger datasets such as CBIS-DDSM or MIAS.

Although MIAS dataset was modestly used in the literature, it has small size and one view (MLO) only. The use of one view to train a model, which deprived it of essential features can be learnt from CC view. Furthermore, it is an old dataset generated 20 years ago, with imbalanced data. Despite its limitations, MIAS dataset can still be useful for training and testing basic machine-learning algorithms for breast cancer detection. Furthermore, it can be leveraged for educational purposes, such as dealing with imbalanced data, and comparing and observing the various techniques to deal with imbalanced data and identify the most effective method.

CMMD dataset comes with a high-quality images, biopsy confirmed lesions, and along with molecular subtyping; however, it lacks the ROI masks. CMMD is well suited for research focusing on personalized medicine and subtype-specific analyses. Researchers interested in developing subtype-specific diagnostic models or investigating the molecular characteristics of breast cancer could benefit from this dataset. Despite the absence of ROI masks, the comprehensive clinical information provided by each image enhances its value for translational research and clinical decision making.

In conclusion, open-access mammography databases play a crucial role in advancing the research and improving breast cancer detection and diagnosis with the aid of artificial intelligence. The availability of diverse datasets, such as CBIS-DDSM, Mini-DDSM, INbreast, MIAS, and CMMD, facilitates the training and validation of AI algorithms, enabling the development of more accurate and efficient computer-aided diagnosis systems. Although each dataset offers unique advantages and challenges, their collective contribution to breast cancer research is undeniable. Continued efforts to expand and enhance open-access databases, coupled with advancements in AI technologies, hold great promise for improving breast cancer screening, diagnosis, and ultimately patient outcomes. Collaboration among researchers, healthcare providers, and database curators is essential for harnessing the full potential of open-access mammography databases in the fight against breast cancer. This review included all open-access mammography datasets available without prerequisite registry, like OMI-DB dataset which are out of the scope of this study.

CONFLICT OF INTEREST

Authors declare no conflict of interest.

FUNDING

None.

ACKNOWLEDGEMENTS

None.

ETHICAL STATEMENT

This study is built on a publicly available datasets not linked to individuals (e.g., Personal data had been de-identified or anonymized). According to the definition of the GDPR (the European General Data Protection Regulation), the dataset is not considered as a personal information belonging to any individual. Therefore, there shall be no ethical issues.

Submitted: February 14, 2024 AST, Accepted: March 31, 2024 AST

REFERENCES

1. Łukasiewicz S, Czeczelewski M, Forma A, Baj J, Sitarz R, Stanisławek A. Breast Cancer—Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies—An Updated Review. *Cancers*. 2021;13(17):4287. doi:10.3390/cancers13174287

2. Smolarz B, Nowak AZ, Romanowicz H. Breast Cancer—Epidemiology, Classification, Pathogenesis and Treatment (Review of Literature). *Cancers*. 2022;14(10):2569. doi:10.3390/cancers14102569

3. Nicosia L, Gnocchi G, Gorini I, et al. History of Mammography: Analysis of Breast Imaging Diagnostic Achievements over the Last Century. *Healthcare*. 2023;11(11):1596. doi:10.3390/healthcare11111596

4. Budh DP, Sapra A. Breast Cancer Screening. In: *StatPearls [Internet]*. StatPearls Publishing; 2023. Accessed January 9, 2024. https://www.ncbi.nlm.nih.gov/books/NBK556050/

5. Poalelungi DG, Musat CL, Fulga A, et al. Advancing Patient Care: How Artificial Intelligence Is Transforming Healthcare. *J Pers Med*. 2023;13(8):1214. doi:10.3390/jpm13081214

6. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging*. 2018;9(4):611-629. doi:10.1007/s13244-018-0639-9

7. Edwards A. Perspective: Science is still too closed. *Nature*. 2016;533(7602):S70. doi:10.1038/533s70a

8. Gasparyan AY, Yessirkepov M, Voronov AA, Koroleva AM, Kitas GD. Comprehensive Approach to Open Access Publishing: Platforms and Tools. *J Korean Med Sci*. 2019;34(27):e184. doi:10.3346/jkms.2019.34.e184

9. Heath MD, Bowyer K, Kopans D, Moore RH. THE DIGITAL DATABASE FOR SCREENING MAMMOGRAPHY. In: ; 2007. Accessed January 30, 2024. https://www.semanticscholar.org/paper/THE-DIGITAL-DATABASE-FOR-SCREENING-MAMMOGRAPHY-Heath-Bowyer/ff2218b349f89026ffaaccdf807228fa497c04bd

10. Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL. A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci Data*. 2017;4(1):170177. doi:10.1038/sdata.2017.177

11. Lekamlage CD, Afzal F, Westerberg E, Cheddad A. Mini-DDSM: Mammography-based Automatic Age Estimation. In: *2020 3rd International Conference on Digital Medicine and Image Processing*. ACM; 2020:1-6. doi:10.1145/3441369.3441370

12. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J Digit Imaging*. 2013;26(6):1045-1057. doi:10.1007/s10278-013-9622-7

13. Mračko A, Vanovčanová L, Cimrák I. Mammography Datasets for Neural Networks—Survey. *J Imaging*. 2023;9(5):95. doi:10.3390/jimaging9050095

14. Shen L, Margolies LR, Rothstein JH, Fluder E, McBride R, Sieh W. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Sci Rep*. 2019;9(1):12495. doi:10.1038/s41598-019-48995-4

15. Ribeiro RF, Torres HR, Oliveira B, Morais P, Vilaça JL. Comparative analysis of deep learning methods for lesion detection on full screening mammography. In: *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE; 2023:1-4. doi:10.1109/embc40787.2023.10340501

16. Masood A, Naseem U, Kim J. Multi-Level Swin Transformer Enabled Automatic Segmentation and Classification of Breast Metastases. In: *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE; 2023:1-4. doi:10.1109/embc40787.2023.10340831

17. Baccouche A, Garcia-Zapirain B, Elmaghraby AS. An integrated framework for breast mass classification and diagnosis using stacked ensemble of residual neural networks. *Sci Rep*. 2022;12(1):12259. doi:10.1038/s41598-022-15632-6

18. Chen JL, Cheng LH, Wang J, et al. A YOLO-based AI system for classifying calcifications on spot magnification mammograms. *BioMed Eng OnLine*. 2023;22(1):54. doi:10.1186/s12938-023-01115-w

19. Logan J, Kennedy PJ, Catchpoole D. A review of the machine learning datasets in mammography, their adherence to the FAIR principles and the outlook for the future. *Sci Data*. 2023;10(1):595. doi:10.1038/s41597-023-02430-6

20. Tsochatzidis L, Koutla P, Costaridou L, Pratikakis I. Integrating segmentation information into CNN for breast cancer diagnosis of mammographic masses. *Computer Methods and Programs in Biomedicine*. 2021;200:105913. doi:10.1016/j.cmpb.2020.105913

21. Khan MHM, Boodoo-Jahangeer N, Dullull W, et al. Multi- class classification of breast cancer abnormalities using Deep Convolutional Neural Network (CNN). *PLOS ONE*. 2021;16(8):e0256500. doi:10.1371/journal.pone.0256500

22. Manzoor Mughal A, Hassan N, Ahmed A. Bone Age Assessment Methods: A Critical Review. *Pak J Med Sci*. 1969;30(1):211-215. doi:10.12669/pjms.301.4295

23. Limdiwala PG, Shah JS. Age estimation by using dental radiographs. *J Forensic Dent Sci*. 2013;5(2):118-122. doi:10.4103/0975-1475.119778

24. Cruz-Ramos C, García-Avila O, Almaraz-Damian JA, Ponomaryov V, Reyes-Reyes R, Sadovnychiy S. Benign and Malignant Breast Tumor Classification in Ultrasound and Mammography Images via Fusion of Deep Learning and Handcraft Features. *Entropy*. 2023;25(7):991. doi:10.3390/e25070991

25. Mohapatra S, Muduly S, Mohanty S, Moharana SK. Evaluation of Deep Learning Models for Detecting Breast Cancer Using Mammograms. In: Mohanty MN, Das S, Ray M, Patra B, eds. *Meta Heuristic Techniques in Software Engineering and Its Applications*. Springer International Publishing; 2022:104-112. doi:10.1007/978-3-031-11713-8_11

26. Sahu A, Das PK, Meher S. High accuracy hybrid CNN classifiers for breast cancer detection using mammogram and ultrasound datasets. *Biomedical Signal Processing and Control*. 2023;80:104292. doi:10.1016/j.bspc.2022.104292

27. Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. INbreast: toward a full-field digital mammographic database. *Acad Radiol*. 2012;19(2):236-248. doi:10.1016/j.acra.2011.09.014

28. Muduli D, Dash R, Majhi B. Automated diagnosis of breast cancer using multi-modal datasets: A deep convolution neural network based approach. *Biomedical Signal Processing and Control*. 2022;71:102825. doi:10.1016/j.bspc.2021.102825

29. Bell BM, Gossweiler M. Benign Breast Calcifications. In: *StatPearls [Internet]*. StatPearls Publishing; 2023. Accessed March 17, 2024. https://www.ncbi.nlm.nih.gov/books/NBK557567/

30. Breast Calcification: Types, Causes, Tests & Treatment. Cleveland Clinic. Accessed February 10, 2024. https://my.clevelandclinic.org/health/diseases/17802-breast-calcifications

31. Panta P, Benjankar RB, Sharma P. Estimation of Abnormal Findings in Screening Mammography: A Crosssectional Descriptive Study. *Civil Med J*. 2023;1(1):12-14. doi:10.59338/cmj.3

32. Suckling J, Parker J, Dance D, et al. Mammographic Image Analysis Society (MIAS) database v1.21. Published online August 28, 2015. Accessed January 30, 2024. https://www.repository.cam.ac.uk/handle/1810/250394

33. Mustra M, Grgic M, Delac K. Feature selection for automatic breast density classification. In: *Proceedings ELMAR-2010*. ; 2010:9-16. Accessed February 13, 2024. https://ieeexplore.ieee.org/document/5606077

34. Thirumalaisamy S, Thangavilou K, Rajadurai H, et al. Breast Cancer Classification Using Synthesized Deep Learning Model with Metaheuristic Optimization Algorithm. *Diagnostics*. 2023;13(18):2925. doi:10.3390/diagnostics13182925

35. Harris C, Okorie U, Makrogiannis S. Spatially localized sparse approximations of deep features for breast mass characterization. *MBE*. 2023;20(9):15859-15882. doi:10.3934/mbe.2023706

36. Agnes SA, Anitha J, Pandian SIA, Peter JD. Classification of Mammogram Images Using Multiscale all Convolutional Neural Network (MA-CNN). *J Med Syst*. 2019;44(1):30. doi:10.1007/s10916-019-1494-z

37. Yang SC. A robust approach for subject segmentation of medical Images: Illustration with mammograms and breast magnetic resonance images. *Computers & Electrical Engineering*. 2017;62:151-165. doi:10.1016/j.compeleceng.2016.12.022

38. Cai H, Wang J, Dan T, et al. An Online Mammography Database with Biopsy Confirmed Types. *Sci Data*. 2023;10(1):123. doi:10.1038/s41597-023-02025-1

39. Sait ARW, Nagaraj R. An Enhanced LightGBM-Based Breast Cancer Detection Technique Using Mammography Images. *Diagnostics*. 2024;14(2):227. doi:10.3390/diagnostics14020227

40. Bobowicz M, Rygusik M, Buler J, et al. Attention-Based Deep Learning System for Classification of Breast Lesions—Multimodal, Weakly Supervised Approach. *Cancers*. 2023;15(10):2704. doi:10.3390/cancers15102704

41. Nguyen TH, Kha QH, Ngoc Toan Truong T, et al. Towards Robust Natural-Looking Mammography Lesion Synthesis on Ipsilateral Dual-Views Breast Cancer Analysis. In: *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE; 2023:2556-2565. doi:10.1109/iccvw60793.2023.00270

42. Jan M, Zainal N, Jamaludin S. Region of interest-based image retrieval techniques: a review. *IJ-AI*. 2020;9(3):520. doi:10.11591/ijai.v9.i3.pp520-528

43. Elkorany AS, Elsharkawy ZF. Efficient breast cancer mammograms diagnosis using three deep neural networks and term variance. *Sci Rep*. 2023;13(1):2663. doi:10.1038/s41598-023-29875-4

SUPPLEMENTARY MATERIALS

Supplementary material

Download: https://hymr.scholasticahq.com/article/116137-from-data-to-diagnosis-narrative-review-of-open-access-mammography-databases-for-breast-cancer-detection/attachment/222785.docx?auth_token=X5cvx0NwqmtiSpedh8wf