data

MDPI

Data Descriptor

King Abdulaziz University Breast Cancer Mammogram Dataset (KAU-BCMD)

Asmaa S. Alsolami [1,*,†], Wafaa Shalash [1,2,†], Wafaa Alsaggaf [1], Sawsan Ashoor [3], Haneen Refaat [3] and Mohammed Elmogy [4,*]

[1] Faculty Computers and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia; wshalash@kau.edu.sa (W.S.); waalsaggaf@kau.edu.sa (W.A.)
[2] Faculty of Computers and Artificial Intelligence, Benha University, Benha 13511, Egypt
[3] Sheikh Mohammed Hussein Al-Amoudi Center of Excellence in Breast Cancer, King Abdulaziz University, Jeddah 21589, Saudi Arabia; dr.sawsanashoor@gmail.com (S.A.); Haneen.refaat@hotmail.com (H.R.)
[4] Information Technology Department, Faculty of Computers and Information, Mansoura University, Mansoura 35516, Egypt
* Correspondence: asalselami@kau.edu.sa (A.S.A.); melmogy@mans.edu.eg (M.E.)
† Both authors contributed equally to this manuscript.

Abstract: The current era is characterized by the rapidly increasing use of computer-aided diagnosis (CAD) systems in the medical field. These systems need a variety of datasets to help develop, evaluate, and compare their performances fairly. Physicians indicated that breast anatomy, especially dense ones, and the probability of breast cancer and tumor development, vary highly depending on race. Researchers reported that breast cancer risk factors are related to culture and society. Thus, there is a massive need for a local dataset representing breast cancer in our region to help develop and evaluate automatic breast cancer CAD systems. This paper presents a public mammogram dataset called King Abdulaziz University Breast Cancer Mammogram Dataset (KAU-BCMD) version 1. To our knowledge, KAU-BCMD is the first dataset in Saudi Arabia that deals with a large number of mammogram scans. The dataset was collected from the Sheikh Mohammed Hussein Al-Amoudi Center of Excellence in Breast Cancer at King Abdulaziz University. It contains 1416 cases. Each case has two views for both the right and left breasts, resulting in 5662 images based on the breast imaging reporting and data system. It also contains 205 ultrasound cases corresponding to a part of the mammogram cases, with 405 images as a total. The dataset was annotated and reviewed by three different radiologists. Our dataset is a promising dataset that contains different imaging modalities for breast cancer with different cancer grades for Saudi women.

Dataset: https://www.kaggle.com/asmaasaad/king-abdulaziz-university-mammogram-dataset.

Dataset License: CC0.

Keywords: breast cancer mammogram dataset; ultrasound breast cancer scans; BI-RADS; clinical data

check for updates

Citation: Alsolami, A.S.; Shalash, W.; Alsaggaf, W.; Ashoor, S.; Refaat, H.; Elmogy, M. King Abdulaziz University Breast Cancer Mammogram Dataset (KAU-BCMD). *Data* **2021**, *6*, 111. https://doi.org/10.3390/data6110111

Academic Editors: Munish Kumar, R. K. Sharma, Ishwar Sethi and Sameer Antani

Received: 13 April 2021
Accepted: 19 October 2021
Published: 25 October 2021

1. Summary

Breast cancer is considered a common disease and the second leading cancer among women in the world [1,2]. According to the international agency for research on cancer report, more than 2 million women were diagnosed with breast cancer [1,3]. Moreover, the Saudi ministry of health reported that one out of eight women is diagnosed with breast cancer [4]. These figures signify an urgent need for a local public dataset that utilizes modern technology to build an accurate computer-aided detection and diagnosis (CAD) system to detect and classify breast cancer. Breast screening is the only way to detect early breast cancer. Therefore, it is essential for women, especially those over 40 years, to undergo it periodically even if they have no symptoms [2,5,6].

Several methods are available for breast imaging, such as mammography, ultrasound (US), magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), and microwave imaging [7]. Breast imaging that uses low-dose x-rays to detect cancer is known as screening mammography. A mammogram is the most widely used and reliable tool for breast cancer screening, exceeding even US as a tool for breast cancer detection. Breast US is rarely used as a diagnostic method for breast cancer as it does not detect early signs of cancer, such as microcalcifications (tiny calcium deposits) [7,8].

During mammography screening, each case includes the recording of two views for each breast: the craniocaudal (CC), which is a top-to-bottom view, and the mediolateral oblique (MLO) which is a side view [9]. The breast imaging reporting and data system (BI-RADS) is a classification system for breast anomalies that was introduced in 1986 [10]. The BI-RADS enables standardized breast imaging reporting [11] by providing mammography reports, including categories for the description of breast cancer stages. The categories are numbered from 0 to 6, though the last category for the approved malignant state was recently added to this report [11,12]. BI-RADS 0 refers to an incomplete diagnosis that needs an additional image for reclassification. Table 1 shows the categories for BI-RADS in detail [13,14]. The authors follow the Al-Amoudi Center of Excellence in Breast Cancer to scale cases from 0 to 6.

Table 1. The BIRAD classification system.

BIRADS	Category	Description
0	Mammography incomplete	Needs additional image
1	Negative	Normal
2	Benign	5% changes
3	Probably benign	Follow up (6 months)
4	Suspicious malignant	Probability of malignancy
5	Malignant	Highly suggestive of malignancy (>95% probability of malignancy)
6	Proven malignant	Known biopsy

Researchers have an increasing need for datasets to develop, test, and evaluate automatic breast cancer CAD systems and build diagnostic systems [15,16]. Most mammogram datasets are private, and few datasets are public for researchers to use during the development of breast cancer tools. This situation has resulted in a lack of comparison among different classification methods. Researchers also reported that breast cancer risk factors are related to culture and society [4,17–19]. Therefore, local and public mammogram datasets are needed to help researchers detect and classify automatic breast cancer systems in women in Saudi Arabia, especially in the early stages. Some factors affect the probability of increasing breast cancer in Saudi more or less than other countries, such as health-related characteristics, menstrual history, obesity, and lack of exercise [8,12,20,21]. Early detection of breast cancer increases the probability of a cure to 92–96% [1,3].

This research's main contribution is a published local mammogram dataset based on BI-RADS categories that attempted to solve local public datasets' availability problem. This is achieved by collecting, categorizing, and annotating mammogram images from a local hospital.

The main advantage of this work it provides a new digitalized mammogram dataset for breast cancer in Saudi Arabia. Additionally, the dataset will help researchers provide reliable systems for the early detection of breast cancer, thereby supporting the medical field, especially in Saudi Arabia. It will also support the medical and educational fields by providing physicians with different diagnosed cases. The King Abdulaziz University Breast Cancer Mammogram Dataset (KAU-BCMD) contains 1416 cases, each with two types of views for both the right and left breasts, resulting in 5662 images. The dataset was collected from 2019 to 2020 from Sheikh Mohammed Hussein Al-Amoudi Center of

Excellence in Breast Cancer in King Abdul-Aziz University. Information about dataset accessibility and specifications is provided in Table 2. The KAU-BCMD is a valuable tool in developing and testing decision support systems due to its size and ground truth (GT).

Table 2. The database specifications.

Subject Area	**Breast Cancer, Mammogram**
More specific subject area	Breast cancer early detection based on BIRAD system
Modality	Mammogram, Ultrasound (US)
Type of data	DICOM, JPG
How data was acquired	Breast imaging technology from IMS Giotto as DICOM images
Data format	Raw and Annotations
Experimental factors	All the patients were subjected to breast cancer classification with one of BIRAD level
Experimental features	Provide enough data for breast cancer detection and classification using deep learning classification.
Data source location	Sheikh Mohammed Hussein Al-Amoudi Center of Excellence in Breast Cancer in King Abdul-Aziz University, Jeddah, Saudi Arabia
Data accessibility	https://www.kaggle.com/asmaasaad/king-abdulaziz-university-mammogram-dataset (accessed on 20 October 2021)

The remaining part of the paper is structured as follows. Section 2 describes some of the available mammogram datasets. Section 3 presents the dataset description. Section 4 discusses the methods used to collect and generate the dataset. Section 5 provides the discussion. Finally, Section 6 provides the conclusion and future work.

2. Related Work

In the following subsections, we describe the most famous public and private breast mammogram datasets. The main goal for discussing these datasets is to help researchers in the medical field and improve the CAD system's performance.

2.1. The Digital Dataset for Screening Mammography (DDSM) Dataset

The DDSM dataset was developed by the University of South Florida and published in 1999 [22]. This dataset contains mammogram images accompanied by some information, such as patient age, date of the screening, abnormality type, and breast density [23]. The largest mammogram dataset contains 2620 cases with four views each and available in 43 volumes with the images categorized as normal, malignant, and benign.

2.2. The Curated Breast Imaging Subset (CBIS-DDSM) Dataset

The CBIS-DDSM dataset is an updated version of the DDSM. The main reason for this dataset is to update and enhance the image segmentation of the DDSM. The CBIS-DDSM updates the region of interest (ROI) annotation and evaluates specialist and segmentation methods. The dataset contains more than 1000 images and divides them into two types of abnormalities, calcification, and mass, for training and testing any breast cancer detection model [24,25].

2.3. The INBREAST Dataset

The INBREAST dataset was a public mammogram dataset from the breast research group. The dataset was collected from the Breast Center in CHSJ Porto Hospital of St. John (CHSJ) and was published in 2010. It had a total of 115 DICOM-formatted cases with 90 images in two views (CC, MLO) and 115 cases (410 images), of which 90 cases (4 images per case) are from women with both breasts, and 25 cases (2 images per case) are from breast surgery patients. The INBREAST dataset included mass, calcification, and normal images, according to the BI-RADS categorial. The dataset can no longer be found [26].

2.4. The Mammographic Image Analysis Society (MIAS) Dataset

The MIAS dataset is one of the oldest datasets. It is a private dataset from the UK research group. It includes a total of 161 cases and 322 images from malignant, benign, and normal mammograms. The dataset includes annotation images consisting of circles around the ROI [27].

2.5. Other Datasets

The MIRacle dataset [28] contains mammography images by radiologists and is used for computer learning. It contains 204 images from 196 cases. This dataset has two modes: classification and radiologist evaluation. The Magic 5 Italian dataset [29] was collected from several hospitals. It includes 967 cases, depending on pathology type. A dataset from Nijmegan, Netherlands, was published as a digital mammogram collection from the university hospital's radiology department, but it is no longer available [30]. The LLNL dataset [30] contains 197 images in two views saved in image cytometry standard (ICS) format. The dataset also contains patient information and biopsy results. A special dataset that integrates multiple datasets is the IRAM dataset [31], which contains a huge number of images. Table 3 shows a comparison between the different mammogram datasets. Approximately 25% of mammogram datasets are public for the research community.

Table 3. A summary of the mammogram datasets.

Dataset	**MIAS** [27]	**DDSM** [23]	**CBIS-DDSM** [24,25]	**INbreast** [26]	**MIRacle** [28]	**Magic5** [29]	**Nijmegen** [30]	**Trueta** [32,33]	**IRAM** [31]	**Malaga** [33]	**LLNL** [30]
Original	UK	USA	USA	Portugal	Greece	Italian	Netherlands	Spain	Germany	Spain	USA
Year	1994	1999	2017–2018	2010	2009	2002	1998	2008	2008	2008	2008
Number of cases	161	2620	6775	115	196	967	21	89	NA	35	50
Number of images	322	10,480	10,239	410	204	3369	40	320	10,500	NA	198
Views	MLO	MLO, CC	MLO, CC	MLO, CC	NA	MLO, CC	MLO, CC	MLO, CC	MLO, CC	MLO, CC	MLO, CC
Image type file	PGM	LJPEG	DICOM	DICOM, XML	NA	DICOM	NA	DICOM	Several	Raw	ICS
BI-RADS	NO	YES	YES	YES	YES	NO	NO	YES	YES	NA	NA
Ground truth	YES	YES	YES	NO	YES	YES	YES	YES	NO	NO	NO
Patient information	NO	YES, AGE	YES, AGE	YES	NO	YES, AGE	NA	NA	NA	NA	NO
Dataset type	Private	Public	Public	Public	Private	Private	Private	Private	Private	Private	Private

3. KAU-BCMD Data Description

The proposed mammography dataset was collected from Sheikh Mohammed Hussein Al-Amoudi Center of Excellence in Breast Cancer at King Abdulaziz University in Jeddah, Saudi Arabia, from April 2019 to March 2020. The annotation was between April and

June 2020. The device used for screening was a breast imaging technology from IMS Giotto, a GMM Group company. The device provides high-quality images with very low SNR (signal-to-noise) [34]. The dataset contains 1416 cases; all cases include images with two types of views (CC and MLO) for both breasts (right and left), making a total of 5662 mammogram images. The dataset was classified into six categories following the BI-RAD system (Table 1). The BI-RADS are verified using US scans. Three different experts verified the BIRAD system using US scans. Then, the majority voting technique is applied to determine the final BIRAD classifications.

Most of our cases fall into BIRADS 2 (48%) category, which is benign. Approximately 21% of cases fall into BIRADS 4 and 5. About a third of the cases (30%) fell into the category of BIRADS 3, as illustrated in Figure 1. The center where the cases were collected provides screening programs for the general population, which explains most of our data's negativity. Digital Imaging and Communications in Medicine (DICOM) is an international standard for transmitting, storing, and displaying medical imaging data. The images were saved in DICOM format, which is a popular format for mammograms. Figure 2 shows the steps of the preprocessing phase of the KAU-BCMD dataset, which will be discussed later. Figures 3–8 show examples from the proposed dataset for BIRADS 0 to BIRADS 5, respectively.

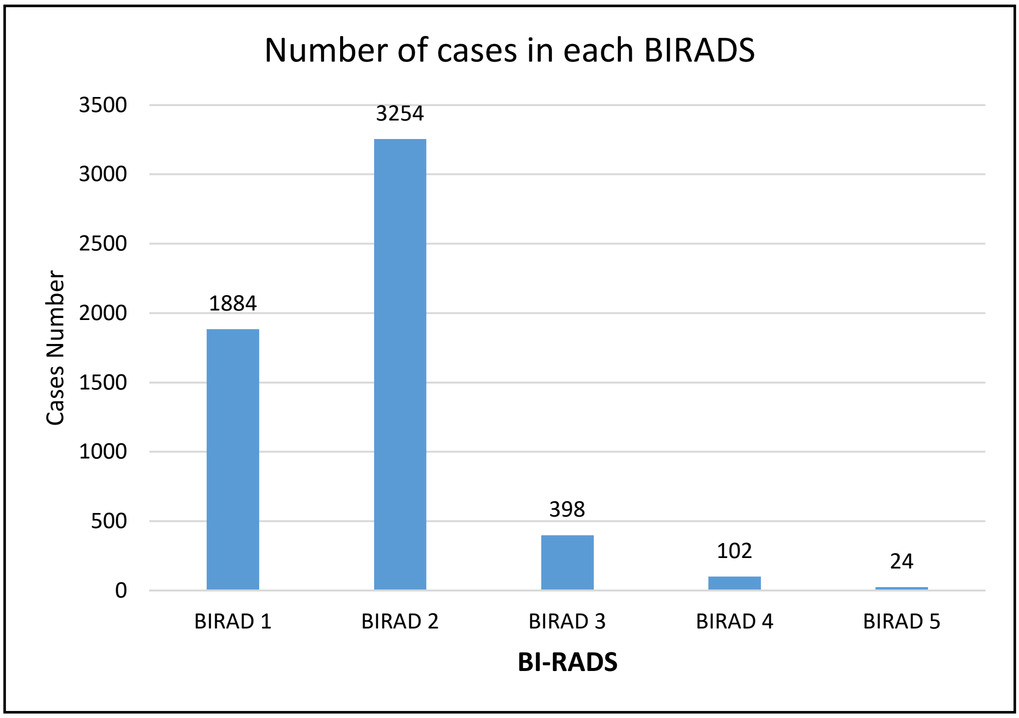


Figure 1. The BI-RADS categories in the KAU-BCMD dataset.

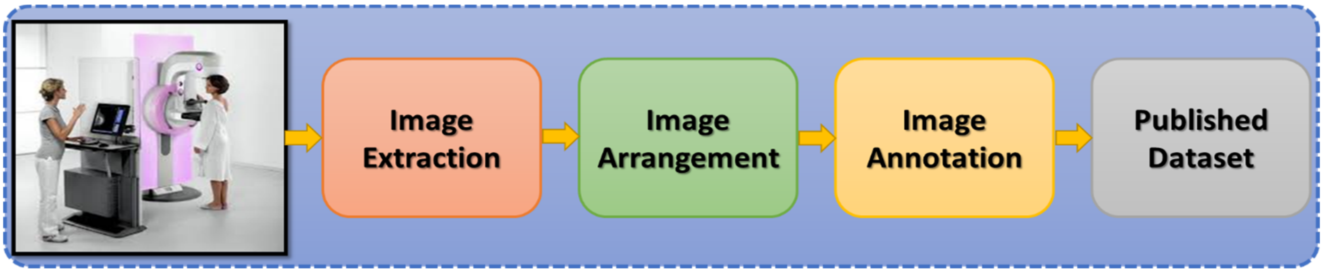


Figure 2. The steps of the preprocessing phase of the KAU-BCMD dataset.

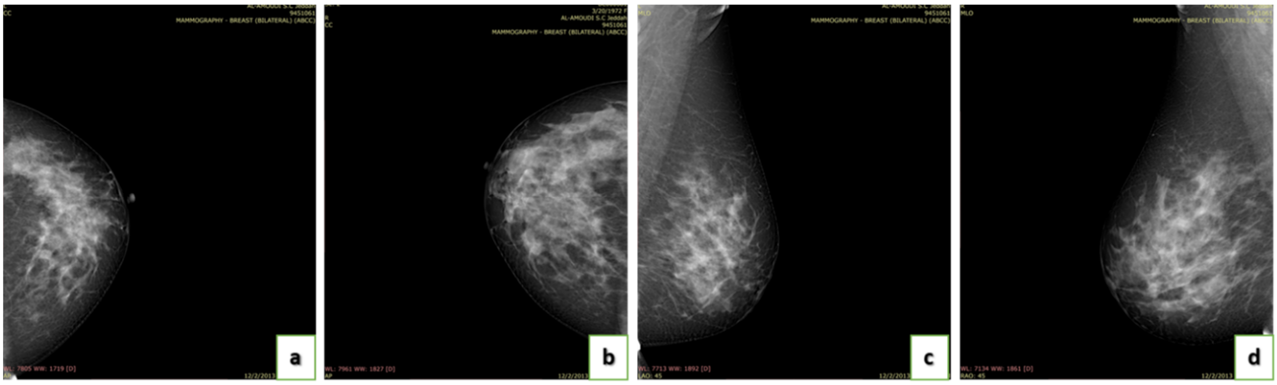


Figure 3. Some examples of BI-RADS 0: (**a**) CC view of the left breast; (**b**) CC view of the right breast; (**c**) MLO view of the left breast; and (**d**) MLO view of the right breast.

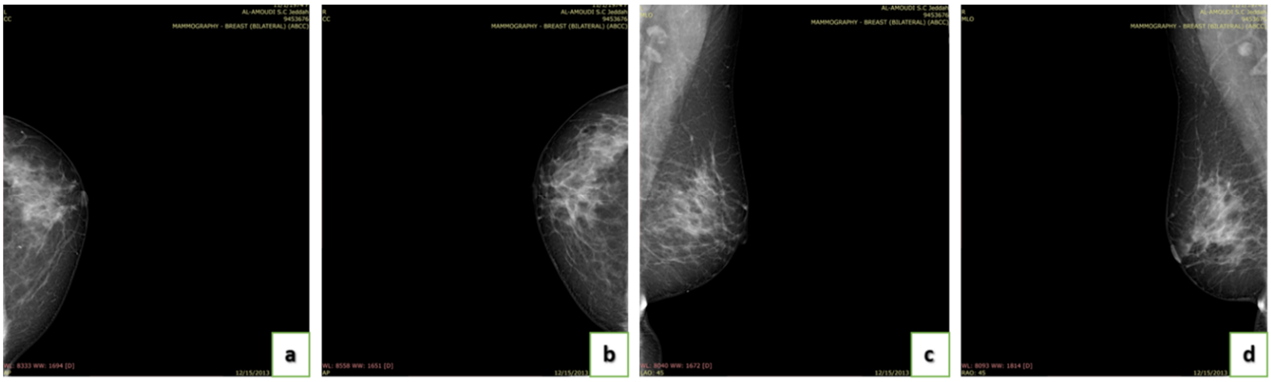


Figure 4. Some examples of BI-RADS 1: (**a**) CC view of the left breast; (**b**) CC view of the right breast; (**c**) MLO view of the left breast; and (**d**) MLO view of the right breast.

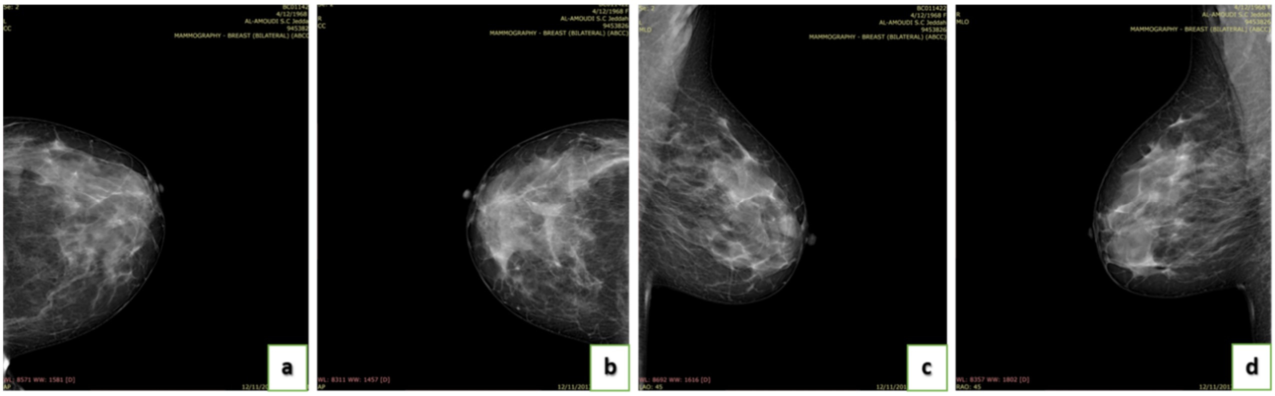


Figure 5. Some examples of BI-RADS 2: (**a**) CC view of the left breast; (**b**) CC view of the right breast; (**c**)MLO view of the left breast; and (**d**) MLO view of the right breast.

The annotation of the images was provided by three different radiologists, which are Dr. Sawsan Ashoor, Dr. Samia Alamoud, and Dr. Gawaher Al Ahadi. They are consultants at the Al Amoudi Breast Cancer Center. The final annotation is created by applying a majority voting technique. The center's system validated the collected images. They were segmented through hand-drawing on the suspicious areas.

To our knowledge, there is no published dataset for breast mammography in Saudi Arabia. Therefore, several work stages need to be accomplished to create such a dataset. Furthermore, successful attempts to construct mammographic datasets fulfilled requirements for validating a mammographic dataset. The current work met the following requirements, which were adopted from research [35–37]. Figure 2 shows a diagram of the process of creating the dataset.

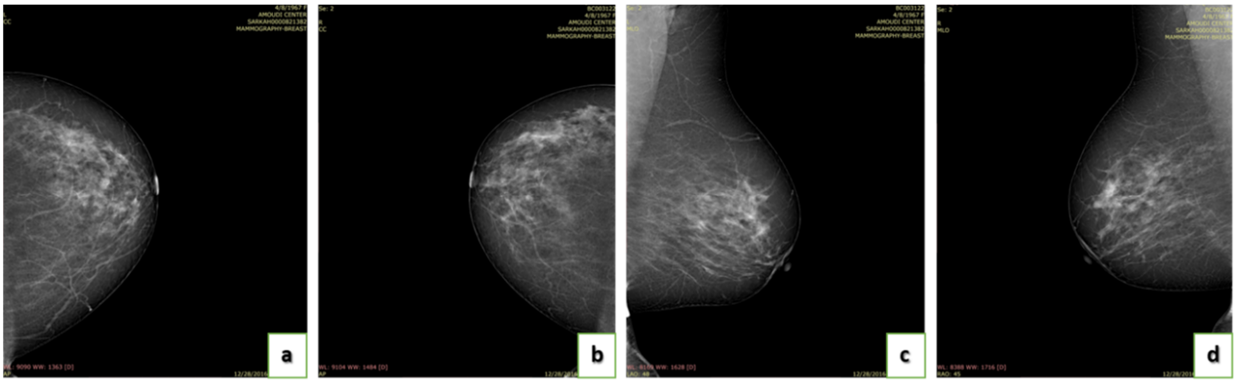


Figure 6. Some examples of BI-RADS 3: (**a**) CC view of the left breast; (**b**) CC view of the right breast; (**c**) MLO view of the left breast; and (**d**) MLO view of the right breast.

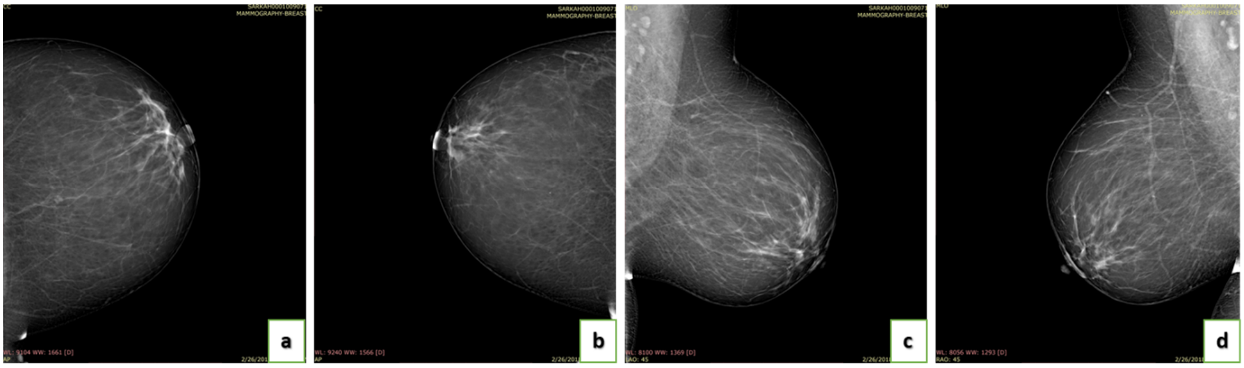


Figure 7. Some examples of BI-RADS 4: (**a**) CC view of the left breast; (**b**) CC view of the right breast; (**c**) MLO view of the left breast; and (**d**) MLO view of the right breast.

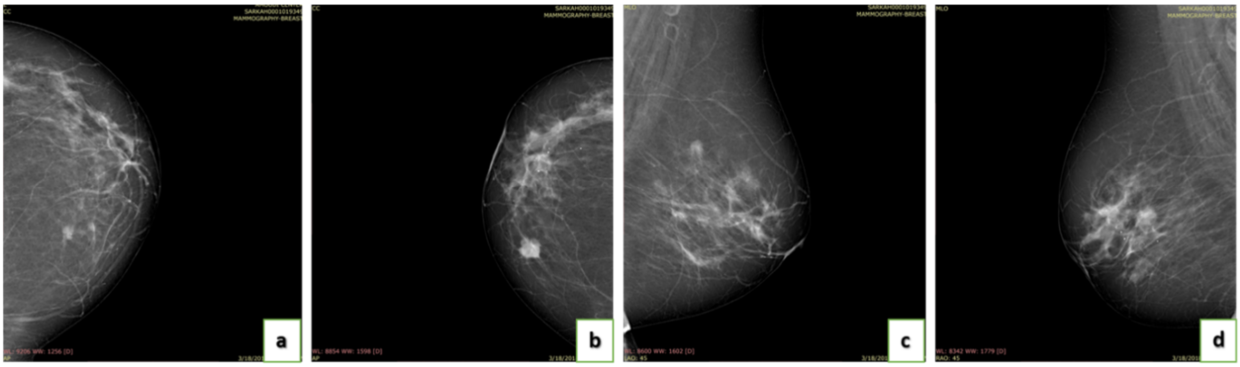
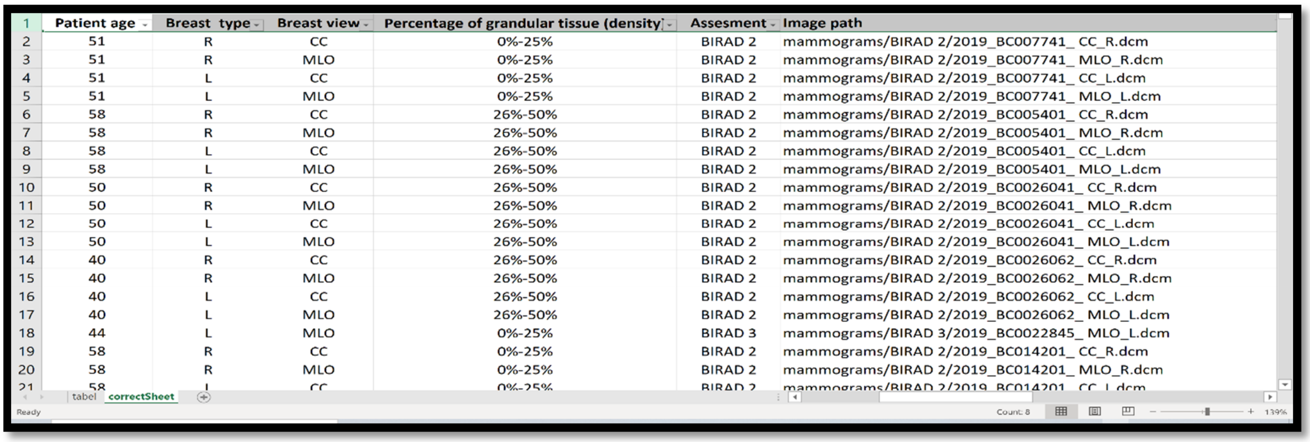


Figure 8. Some examples of BI-RADS 5: (**a**) CC view of the left breast; (**b**) CC view of the right breast; (**c**) MLO view of the left breast; and (**d**) MLO view of the right breast.

The dataset contains five folders divided based on BIRAD categories and includes DICOM and JPG image formats in separate folders. In addition, they include a tumor mask for benign and malignant in JPG formats. The dataset also contains the information in the CSV file, as shown in Figure 9. The CSV file contains the following fields:

A. Date of the scan: the study of mammogram screening.
B. Patient ID: It is a unique number to distinguish the records.
C. Patient age.

D. Breast type: left or right breast.
E. Breast view: CC or MLO.
F. Assessment: BIRAD categories level.
G. Images path: contains the scan folder.



1	Patient age	Breast type	Breast view	Percentage of grandular tissue (density)	Assesment	Image path
2	51	R	CC	0%-25%	BIRAD 2	mammograms/BIRAD 2/2019_BC007741_ CC_R.dcm
3	51	R	MLO	0%-25%	BIRAD 2	mammograms/BIRAD 2/2019_BC007741_ MLO_R.dcm
4	51	L	CC	0%-25%	BIRAD 2	mammograms/BIRAD 2/2019_BC007741_ CC_L.dcm
5	51	L	MLO	0%-25%	BIRAD 2	mammograms/BIRAD 2/2019_BC007741_ MLO_L.dcm
6	58	R	CC	26%-50%	BIRAD 2	mammograms/BIRAD 2/2019_BC005401_ CC_R.dcm
7	58	R	MLO	26%-50%	BIRAD 2	mammograms/BIRAD 2/2019_BC005401_ MLO_R.dcm
8	58	L	CC	26%-50%	BIRAD 2	mammograms/BIRAD 2/2019_BC005401_ CC_L.dcm
9	58	L	MLO	26%-50%	BIRAD 2	mammograms/BIRAD 2/2019_BC005401_ MLO_L.dcm
10	50	R	CC	26%-50%	BIRAD 2	mammograms/BIRAD 2/2019_BC0026041_ CC_R.dcm
11	50	R	MLO	26%-50%	BIRAD 2	mammograms/BIRAD 2/2019_BC0026041_ MLO_R.dcm
12	50	L	CC	26%-50%	BIRAD 2	mammograms/BIRAD 2/2019_BC0026041_ CC_L.dcm
13	50	L	MLO	26%-50%	BIRAD 2	mammograms/BIRAD 2/2019_BC0026041_ MLO_L.dcm
14	40	R	CC	26%-50%	BIRAD 2	mammograms/BIRAD 2/2019_BC0026062_ CC_R.dcm
15	40	R	MLO	26%-50%	BIRAD 2	mammograms/BIRAD 2/2019_BC0026062_ MLO_R.dcm
16	40	L	CC	26%-50%	BIRAD 2	mammograms/BIRAD 2/2019_BC0026062_ CC_L.dcm
17	40	L	MLO	26%-50%	BIRAD 2	mammograms/BIRAD 2/2019_BC0026062_ MLO_L.dcm
18	44	L	MLO	0%-25%	BIRAD 3	mammograms/BIRAD 3/2019_BC0022845_ MLO_L.dcm
19	58	R	CC	0%-25%	BIRAD 2	mammograms/BIRAD 2/2019_BC014201_ CC_R.dcm
20	58	R	MLO	0%-25%	BIRAD 2	mammograms/BIRAD 2/2019_BC014201_ MLO_R.dcm
21	58	L	CC	0%-25%	BIRAD 2	mammograms/BIRAD 2/2019_BC014201_ CC_L.dcm

Figure 9. A sample of the KAU-BCMD dataset metadata that is stored in CSV file format.

4. Methods

4.1. Ethics Statement

The authors followed the Saudi executive regulations of the system of ethics for research on living creatures. The dataset received approval from the local research ethics Committee at King Abdul Aziz University to be published with the dataset (1 February 2021).

4.2. Annotation of Images

Initially, all listed cases in the dataset were annotated and validated by three different radiologists: Dr. Sawsan Ashoor, Dr. Samia Alamoud, and Dr. Gawaher Al Ahadi. Figures 10–12 show examples of the image annotation from our proposed dataset. The breast cancer images were segmented through hand-drawing on the suspicious areas in the dataset for the BI-RADS 3, 4, and 5. Figures 13 and 14 show examples of the dataset masks for BI-RADS 3 and 4, respectively. The dataset includes RoI segmentation and bounding box images generated by the image labeler App in MATLAB. This application marks RoI labels as rectangular on the tumor area for malignant cases (BIRADS 4 and 5), as shown in Figure 15. The app then exported the images to tables containing the coordinator x, y, width, and height provided with dataset images.

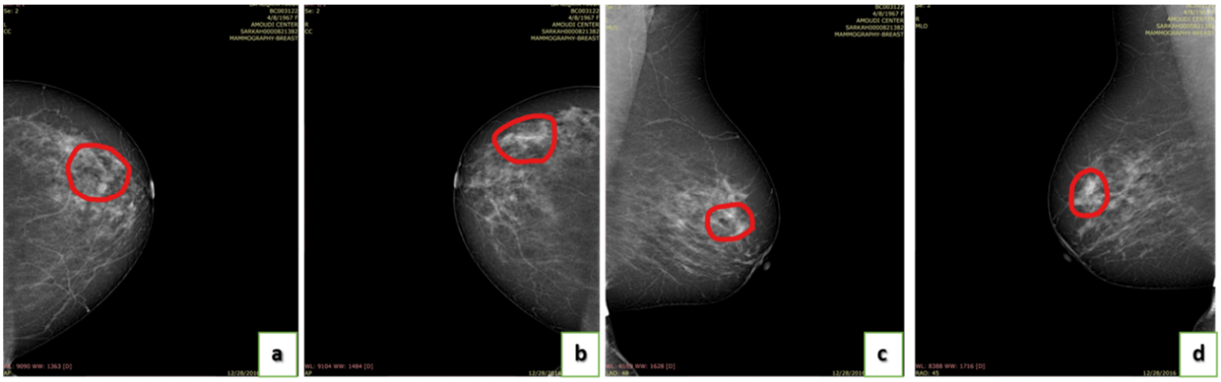


Figure 10. Examples of dataset annotation for BI-RADS 3: (**a**) CC view of the left breast; (**b**) CC view of the right breast; (**c**) MLO view of the left breast; and (**d**) MLO view of the right breast.

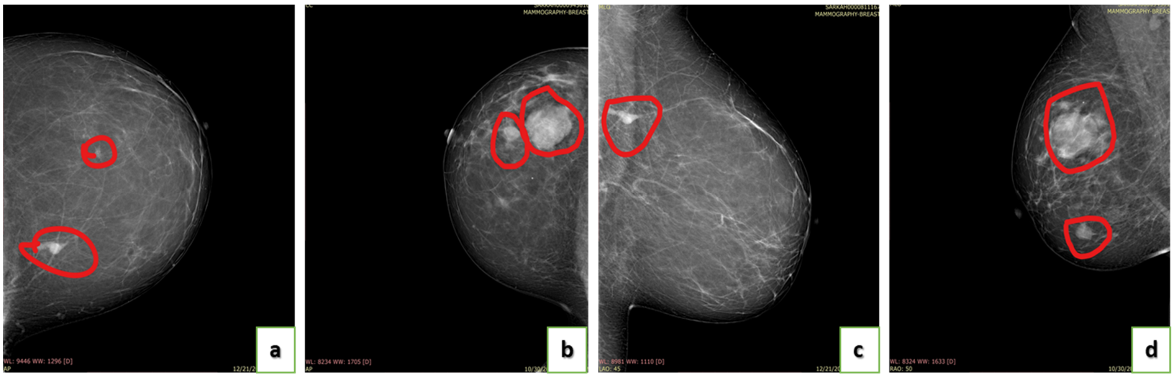


Figure 11. Examples of dataset annotation for BI-RADS 4: (**a**) CC view of the left breast; (**b**) CC view of the right breast; (**c**) MLO view of the left breast; and (**d**) MLO view of the right breast.

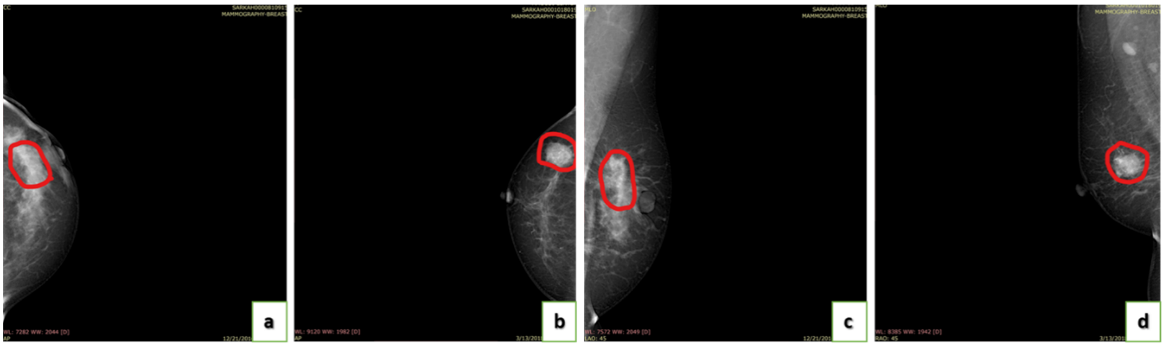


Figure 12. Examples of dataset annotation for BI-RADS 5: (**a**) CC view of the left breast; (**b**) CC view of the right breast; (**c**) MLO view of the left breast; and (**d**) MLO view of the right breast.

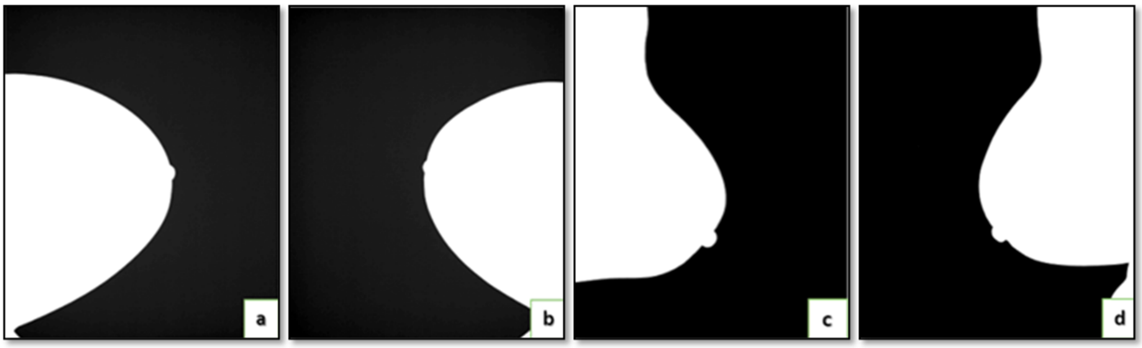


Figure 13. Example of dataset mask for BI-RADS 3: (**a**) CC view of the left breast; (**b**) CC view of the right breast; (**c**) MLO view of the left breast; and (**d**) MLO view of the right breast.

4.3. Data Acquisition

The dataset includes normal, benign, and malignant cases. In addition, it contains pathology details and patients' histories. It includes the age and previous screenings, as this may be useful for the researcher's study. BI-RADS categories were also reported, as they are considered essential information for a digital mammogram dataset. The authors provide DICOM and JPG format on the dataset. We followed the following steps, as shown in Figure 2:

A. Image preparing and collecting.
B. Image labeling.
C. Image validation by a committee of radiologists.
D. Publish the dataset.

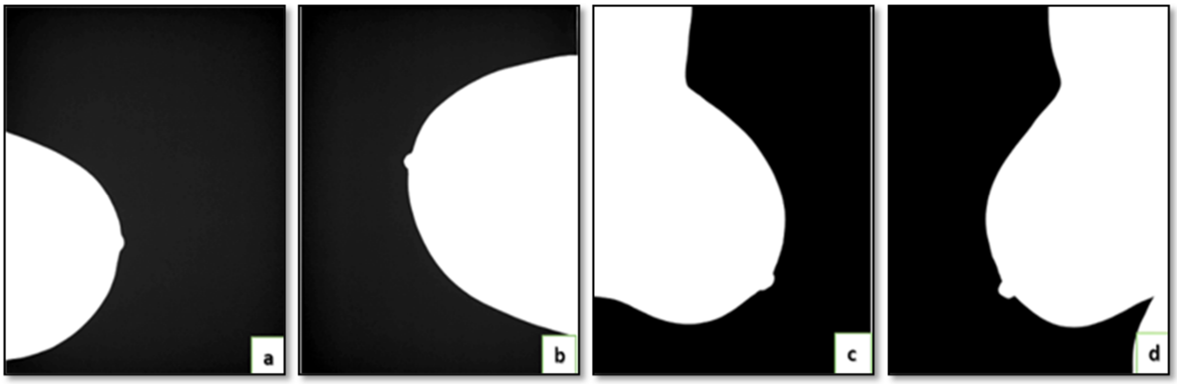


Figure 14. Example of dataset mask for BI-RADS 4: (**a**) CC view of the left breast; (**b**) CC view of the right breast; (**c**) MLO view of the left breast; and (**d**) MLO view of the right breast.

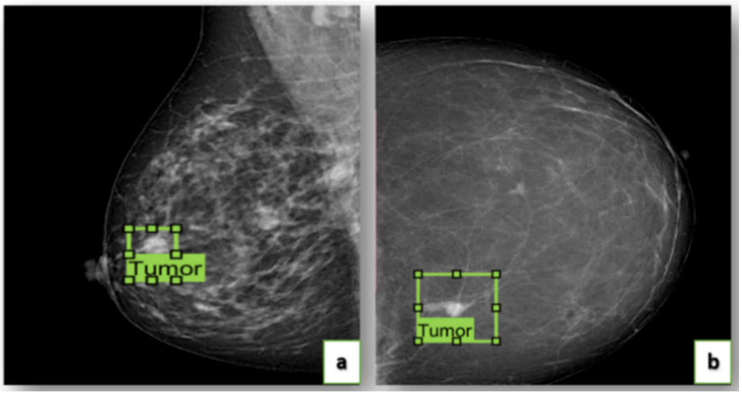


Figure 15. Example of image dataset using image labeler app in MATLAB for BIRADS 4 and 5, respectively.

US Images

The proposed dataset contains a subset of US images for 205 cases that need more investigation after mammogram screening. The total number of images is 405 different images for the left or right sides per case. The US images were obtained using the iU22 xMATRIX device. They have a size of 2816 by 3584 pixels and are stored in DICOM and JPG format. The importance of US comes after a mammogram, as a mammogram scan can detect early stages efficiently while ultrasound can detect further stages. Some of the US diagnoses were concurrent with the mammogram diagnosis, while most of the data images diagnosed in ultrasound were diagnoses as BI-RADS 0 from the mammogram results. The US images are raw, i.e., not annotated. Figures 16 and 17 show the detailed categorization of the US image data according to the BI-RADS system. Figure 18 shows a sample of the US images. The US data, with the mammogram data, open the path to more investigation and classification using multimodal data to increase the accuracy of the automatic classification system.

4.4. Breast Density

Mammographic density is considered a decisive risk factor for breast cancer. The risk of women with high breast density is 4–6 fold compared with women with low density [38,39]. Breast density refers to the volume of fibrous and glandular tissue in a woman's breasts compared to the amount of fatty tissue in the breasts. Therefore, the probability of having breast cancer increases as the women's breasts density increases. The denser breasts are, the higher the risk of breast cancer, but there is no apparent cause.

Several methods are available for measuring breast density, but it is unclear which method is the best predictor of breast cancer risk. BI-RADS is considered the most widely used method in clinics to estimate breast density. It uses a density score. BI-RADS has several limitations based on subjective visual assessment and is time-consuming [38–40].

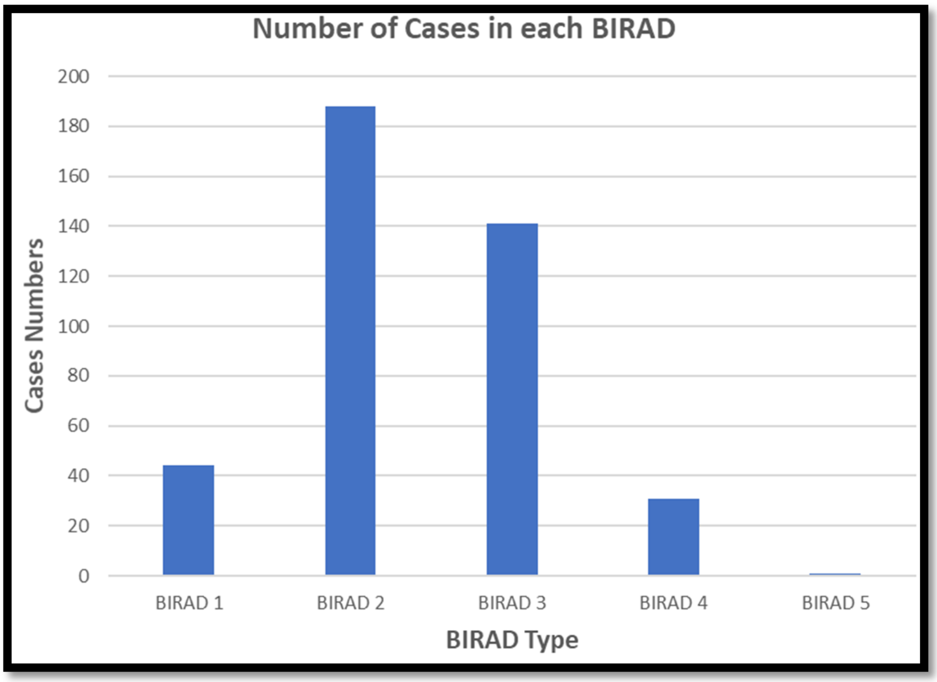


Figure 16. The BI-RADS categories of ultrasound images dataset in the KAU-BCMD dataset.

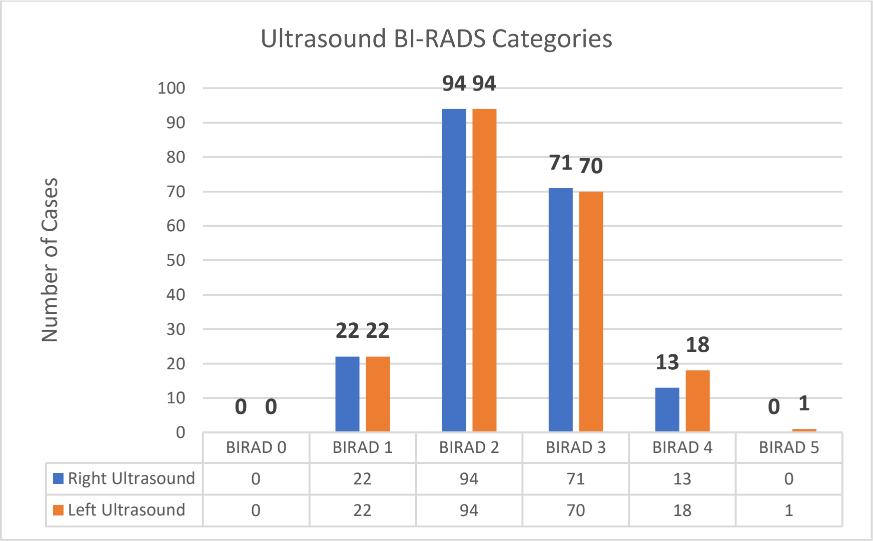


Figure 17. The BI-RADS categories of US images dataset in the KAU-BCMD.

In our dataset, breast density is estimated by the radiologist who examines the mammogram to estimate the ratio of non-dense tissue to dense tissue and assigns a level of breast density. The breast density levels are defined using the BI-RADS reporting system. The levels of density are:

- A (0–25%): Almost entirely fatty indicates that the breasts are almost entirely composed of fat. One out of ten women has this result.
- B (25–50%): Scattered areas of fibroglandular density indicate some scattered areas of density, but most of the breast tissue is non-dense. Four out of ten women have this result.

- C (50–75%): Heterogeneously dense indicates that there are some areas of non-dense tissue but that most of the breast tissue is dense. Four out of ten women have this result.
- D (75–100%): Extremely dense indicates that nearly all breast tissue is dense. One out of each women has this result.

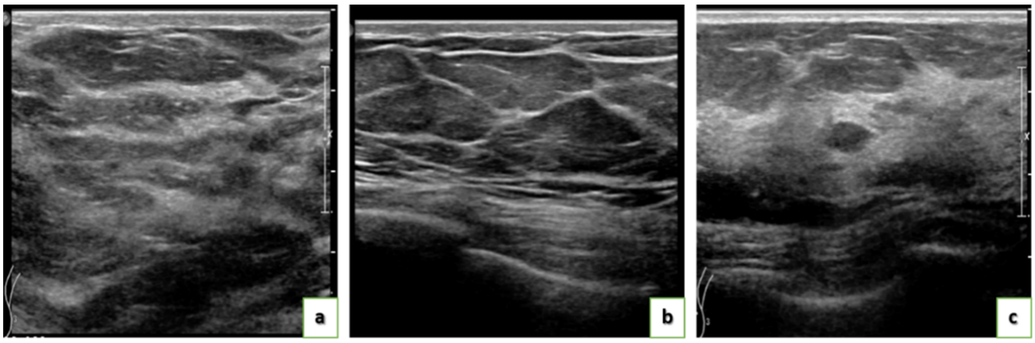


Figure 18. Examples of Ultrasound images in the KAU-BCMD dataset: (**a**) BIRADS 1 for one left breast, (**b**) BI-RADS 3 for both breasts, and (**c**) BI-RADS 4 for both breasts.

In the current work, the breast density was estimated numerically according to the BI-RADS fourth edition based on percentages [40]. It was estimated as 25% for almost entirely fat, 50% for scattered fibroglandular densities, 75% for heterogeneously dense, and finally, 100% for extremely dense. The estimation was performed manually by Prof. Sawsan Ashour (author), Dr. Samia Alamoud, and Dr. Gawaher Al Ahadi. They have more than 20 years of mammogram consulting experience.

5. Discussion

The amount and quality of datasets used to design machine learning-based CAD systems directly related to the system's final accuracy. There is a lack of standard evaluation data in mammography. Most CAD algorithms are evaluated on private datasets as most mammographic databases are not publicly available. This poses a challenge to compare the performance of different methods or to replicating prior results.

Deep learning has recently emerged as a promising medical image classification solution, but it requires many images to learn. Most of the available mammogram datasets provide an inappropriate number of samples for deep learning, which is considered a big challenge. The current work provides a dataset that satisfies public availability and a large sample size. It is the first to be collected and publicly available in the region, as far as we know. The only drawback of the presented dataset is the imbalanced size of the different classes, as shown in Figure 1. Overall, our digital mammogram dataset can be considered the first such dataset in Saudi Arabia. In the future, we aim to increase the number of cases in the BIRADs 3, 4, and 5 classes to make the dataset more balanced and thus more suitable for research purposes.

On the other hand, in deep learning-based CAD systems, the dataset's size could be increased using data augmentation techniques to overcome the imbalanced classes size. This is achieved by adding noise with different percentages or applying various transformations to the dataset and a different rotation and translation level. Moreover, transfer learning techniques are expected to work efficiently with the current dataset size as it is. Additionally, we can measure the machine learning-based CAD systems' performance on unbalanced datasets by using various performance metrics, such as sensitivity, specificity, false-positive rate, false-negative rate, geometrical mean, positive likelihood, and diagnostic odds ratio (DOR), discriminant power (DP), and YI.

The dataset includes a set of US images associated with 205 cases out of 1416 total mammogram cases. The US images were captured for most mammogram BI-RADS 0 classified images when the consultants could not decide for the case. Although the number of US is not large, it could be instrumental in designing a multimodal breast cancer classification system based on mammograms and US images to increase classification accuracy.

Finally, the proposed dataset satisfied most of the ideal medical image dataset criteria described in [36,37,41]. It has adequate data volume, curation, annotation, ground truth, reusability, and generalizability. Each medical imaging data object has metadata and an identifier.

6. Conclusions

In this research, we provide a public mammogram dataset considered a stander of a breast cancer images dataset to help a researcher work on the dataset to produce a CAD system. The proposed work has the potential to be the first digital mammogram dataset in Saudi Arabia. Additionally, the GT is provided with related information. The dataset also contains a subset of many ultrasounds' images corresponding to mammogram cases. The 405 images of ultrasound could be combined with its corresponding mammogram to develop a multimodal CAD breast cancer system. We aim to increase the number of medical images in the dataset to help researchers in breast cancer detection systems. We will develop a second version of the dataset by increasing the number of images to balance and improve their annotation.

Author Contributions: Conceptualization, W.S. and W.A.; methodology, W.S., W.A. and A.S.A.; software, A.S.A.; validation, A.S.A., H.R. and S.A.; formal analysis, A.S.A. and W.S.; investigation, A.S.A.; resources, W.A.; data curation, A.S.A.; writing—original draft preparation, A.S.A.; writing—review and editing, W.S. and M.E.; visualization, A.S.A. and M.E.; supervision, W.A.; project administration, W.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and waived by the Ethics Committee of King Abdul Aziz University (protocol code 08694 and date of approval 1 February 2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study and approved by the Ethics Committee of King Abdul Aziz University.

Data Availability Statement: This dataset is available for download at https://www.kaggle.com/asmaasaad/king-abdulaziz-university-mammogram-dataset.

Acknowledgments: We would like to thank Sheikh Mohammed Hussien Al-Amoudi Center of Excellence in Breast Cancer at King Abdul-Aziz University and the Faculty of Computer and Information Technology. Additionally, we would like to thank Samia Alamoud and Gawaher Al Ahadi for their support.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Observatory, G.C. World Health Organization. 2021. Available online: http://gco.iarc.fr/ (accessed on 20 October 2021).
2. Ahmad, A. Breast cancer statistics: Recent trends. In *Breast Cancer Metastasis and Drug Resistance*; Springer: Berlin/Heidelberg, Germany, 2019; Volume 1152, pp. 1–7.
3. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef]
4. Moh.gov. Women's Health—National Breast Cancer Early Detection Campaign. 2021. Available online: https://www.moh.gov.sa/en/HealthAwareness/EducationalContent/wh/Pages/005.aspx (accessed on 20 October 2021).
5. Krishnamoorthy, Y.; Ganesh, K.; Sakthivel, M. Prevalence and determinants of breast and cervical cancer screening among women aged between 30 and 49 years in India: Secondary data analysis of National Family Health Survey–4. *Indian J. Cancer* 2021. [CrossRef]
6. Van der Meer, D.J.; Kramer, I.; van Maaren, M.C.; van Diest, P.J.; Linn, S.; Maduro, J.H.; Strobbe, L.; Siesling, S.; Schmidt, M.K.; Voogd, A.C. Comprehensive trends in incidence, treatment, survival and mortality of first primary invasive breast cancer stratified by age, stage and receptor subtype in the Netherlands between 1989 and 2017. *Int. J. Cancer* **2021**, *148*, 2289–2303. [CrossRef]
7. Debelee, T.G.; Schwenker, F.; Ibenthal, A.; Yohannes, D. *Survey of Deep Learning in Breast Cancer Image Analysis*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 143–163.

8. Sheppard, V.B.; Sutton, A.L.; Hurtado-de-Mendoza, A.; He, J.; Dahman, B.; Edmonds, M.C.; Hackney, M.H.; Tadesse, M.G. Race and Patient-reported Symptoms in Adherence to Adjuvant Endocrine Therapy: A Report from the Women's Hormonal Initiation and Persistence Study. *Cancer Epidemiol. Prev. Biomark.* **2021**, *30*, 699–709. [CrossRef] [PubMed]
9. Tan, M.; Al-Shabi, M.; Chan, W.Y.; Thomas, L.; Rahmat, K.; Ng, K.H. Comparison of two-dimensional synthesized mammograms versus original digital mammograms: A quantitative assessment. *Med. Biol. Eng. Comput.* **2021**, *59*, 355–367. [CrossRef]
10. The Radiology Assistant. 2021. Available online: https://radiologyassistant.nl/breast/bi-rads/bi-rads-for-mammography-and-ultrasound-2013 (accessed on 20 October 2021).
11. Magny, S.J.; Shikhman, R.; Keppke, A.L. *Breast, Imaging, Reporting and Data System (BI-RADS)*; StatPearls Publishing: Treasure Island, FL, USA, 2020.
12. Menezes, G.L.; Winter-Warnars, G.A.; Koekenbier, E.L.; Groen, E.J.; Verkooijen, H.M.; Pijnappel, R.M. Simplifying Breast Imaging Reporting and Data System classification of mammograms with pure suspicious calcifications. *J. Med. Screen.* **2017**, *25*, 82–87. [CrossRef] [PubMed]
13. De Margerie-Mellon, C.; Debry, J.B.; Dupont, A.; Cuvier, C.; Giacchetti, S.; Teixeira, L.; Espié, M.; de Bazelaire, C. Nonpalpable breast lesions: Impact of a second-opinion review at a breast unit on BI-RADS classification. *Eur. Radiol.* **2021**, *31*, 5913–5923. [CrossRef]
14. Davis, J.; Liang, J.; Roh, A.; Kittrell, L.; Petterson, M.; Winton, L.; Connell, M.; Viscusi, R.; Komenaka, I.; Jamshidi, R. Use of breast imaging-reporting and data system (BI-RADS) ultrasound classification in pediatric and adolescent patients overestimates likelihood of malignancy. *J. Pediatr. Surg.* **2021**, *56*, 1000–1003. [CrossRef] [PubMed]
15. Jagadesh, B.; Kumari, L.K. A GLCM based Feature Extraction in Mammogram Images using Machine Learning Algorithms. *Int. J. Curr. Res. Rev.* **2021**, *13*, 145–149. [CrossRef]
16. Shaikh, K.; Krishnan, S.; Thanki, R. Deep Learning Model for Classification of Breast Cancer. In *Artificial Intelligence in Breast Cancer Early Detection and Diagnosis*; Springer: Cham, Switzerland, 2021; pp. 93–100.
17. Sharma, R. Global, regional, national burden of breast cancer in 185 countries: Evidence from GLOBOCAN 2018. *Breast Cancer Res. Treat.* **2021**, *187*, 557–567. [CrossRef]
18. Turbow, S.D.; White, M.C.; Breslau, E.S.; Sabatino, S.A. Mammography use and breast cancer incidence among older U.S. women. *Breast Cancer Res. Treat.* **2021**, *188*, 307–316. [CrossRef]
19. Alsheik, N.; Blount, L.; Qiong, Q.; Talley, M.; Pohlman, S.; Troeger, K.; Abbey, G.; Mango, V.L.; Pollack, E.; Chong, A.; et al. Outcomes by Race in Breast Cancer Screening With Digital Breast Tomosynthesis Versus Digital Mammography. *J. Am. Coll. Radiol.* **2021**, *18*, 906–918. [CrossRef]
20. Alsolami, F.J.; Azzeh, F.S.; Ghafouri, K.J.; Ghaith, M.M.; Almaimani, R.A.; Almasmoum, H.A.; Abdulal, R.H.; Abdulaal, W.H.; Jazar, A.S.; Tashtoush, S.H. Determinants of breast cancer in Saudi women from Makkah region: A case-control study (breast cancer risk factors among Saudi women). *BMC Public Health* **2019**, *19*, 1554. [CrossRef] [PubMed]
21. Alshahrani, M.; Alhammam, S.Y.M.; Al Munyif, H.A.S.; AlWadei, A.M.A.; AlWadei, A.M.A.; Alzamanan, S.S.M.; Aljohani, N.S.M. Knowledge, Attitudes, and Practices of Breast Cancer Screening Methods Among Female Patients in Primary Healthcare Centers in Najran, Saudi Arabia. *J. Cancer Educ.* **2018**, *34*, 1167–1172. [CrossRef] [PubMed]
22. USF Digital Mammography Home. 2021. Available online: http://marathon.csee.usf.edu/Mammography/Database.html (accessed on 20 October 2021).
23. University of South Florida Digital Mammography Home Page. 2021. Available online: http://www.eng.usf.edu/cvprg/Mammography/Database.html (accessed on 20 October 2021).
24. Lee, R.S.; Gimenez, F.; Hoogi, A.; Miyake, K.K.; Gorovoy, M.; Rubin, D. A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci. Data* **2017**, *4*, 170177. [CrossRef]
25. CBIS-DDSM. 2021. Available online: https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM (accessed on 20 October 2021).
26. Moreira, I.C.; Amaral, I.; Domingues, I.; Cardoso, A.; Cardoso, M.J.; Cardoso, J. INbreast: Toward a Full-field Digital Mammographic Database. *Acad. Radiol.* **2012**, *19*, 236–248. [CrossRef] [PubMed]
27. The Mini-MIAS Database of Mammograms. UK Research Groups. 2021. Available online: http://peipa.essex.ac.uk/info/mias.html (accessed on 20 October 2021).
28. Antoniou, Z.C.; Giannakopoulou, G.P.; Andreadis, I.I.; Nikita, K.S.; Ligomenides, P.A.; Spyrou, G.M. A web-accessible mammographic image database dedicated to combined training and evaluation of radiologists and machines. In Proceedings of the Information Technology and Applications in Biomedicine, Larnaka, Cyprus, 4–7 November 2009.
29. Tangaro, S.; Bellotti, R.; De Carlo, F.; Gargano, G.; Lattanzio, E.; Monno, P.; Massafra, R.; Delogu, P.; Fantacci, M.E.; Retico, A.; et al. MAGIC-5: An Italian mammographic database of digitised images for research. *La Radiol. Med.* **2008**, *113*, 477–485. [CrossRef]
30. Karssemeijer, N.; Thijssen, M.; Hendriks, J.; van Erning, L. (Eds.) *Digital Mammography: Nijmegen*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1998; Volume 13.
31. Oliveira, J.E.; Gueld, M.O.; Araújo, A.D.A.; Ott, B.; Deserno, T.M. Toward a standard reference database for computer-aided mammography. In *Medical Imaging 2008: Computer-Aided Diagnosis*; International Society for Optics and Photonics: Bellingham, WA, USA, 2008; Volume 6915, p. 69151Y.
32. Trueta Database. 2021. Available online: http://eia.udg.edu/aoliver/publications/tesi/node137.html (accessed on 20 October 2021).

33. Oliver, A.; Lladó, X.; Pérez, E.; Pont, J.; Denton, E.R.E.; Freixenet, J.; Martí, J. A statistical approach for breast density segmentation. *J. Digit. Imaging* **2010**, *23*, 527–537. [CrossRef]
34. Zimmermann, D. IMS Giotto—GMM Group—Giotto Class. 2021. Available online: https://healthcare-in-europe.com/en/radbook/mammography/731-ims-giotto-gmm-group-giotto-class.html (accessed on 20 October 2021).
35. Nishikawa. *Development of a Common Database for Digital Mammography Research*; University of Chicago: Chicago, IL, USA, 1996.
36. Kohli, M.D.; Summers, R.M.; Geis, J.R. Medical Image Data and Datasets in the Era of Machine Learning—Whitepaper from the 2016 C-MIMI Meeting Dataset Session. *J. Digit. Imaging* **2017**, *30*, 392–399. [CrossRef]
37. Harvey, H.; Glocker, B. A Standardised Approach for Preparing Imaging Data for Machine Learning Tasks in Radiology. In *Artificial Intelligence in Medical Imaging*; Ranschaert, E., Morozov, S., Algra, P., Eds.; Springer: Cham, Switzerland, 2019. [CrossRef]
38. Vilmun, B.M.; Vejborg, I.; Lynge, E.; Lillholm, M.; Nielsen, M.; Nielsen, M.B.; Carlsen, J.F. Impact of adding breast density to breast cancer risk models: A systematic review. *Eur. J. Radiol.* **2020**, *127*, 109019. [CrossRef]
39. Alonzo-Proulx, O.; Mawdsley, G.; Patrie, J.T.; Yaffe, M.J.; Harvey, J.A. Reliability of Automated Breast Density Measurements. *Radiol.* **2015**, *275*, 366–376. [CrossRef] [PubMed]
40. DSpak, D.; Plaxco, J.; Santiago, L.; Dryden, M.; Dogan, B. BI-RADS ® fifth edition: A summary of changes. *Diagn. Interv. Imaging* **2017**, *98*, 179–190. [CrossRef]
41. Chugh, G.; Kumar, S.; Singh, N. Survey on Machine Learning and Deep Learning Applications in Breast Cancer Diagnosis. *Cogn. Comput.* **2021**, 1–20. [CrossRef]