The Medical City Journal © 2020. Published Online (Ahead of Print) on December 2020

Original Article

A Pilot Study on Artificial Intelligence-Powered Medical Image Analysis for Breast Cancer Detection

Beatrice J. Tiangco,[1]* Reza Koa-Sales,[1] Maria Isabel Saludares,[2]* Mario Domingo[3]

[1]Augusto P. Sarmiento Cancer Institute, The Medical City, Ortigas Avenue, Pasig City, Philippines, [2]University of the Philippines- Diliman, Quezon City, Philippines, [3]Digenomix Inc., Ortigas Avenue, Pasig City, Philippines

* Contact Details: bjtiangco@themedicalcity.com/ isabel.saludares@neuralmechanics.net

ABSTRACT: This study explored the creation of a deep learning model capable of accurately determining whether a lesion on a mammogram is benign or malignant. Data was collected from the screening mammogram of women who were also biopsied at the Breast Center of a large tertiary hospital. Nine hundred breast images taken from 215 mammograms were used to train and build a deep learning-based model capable of accurately predicting whether the lesions were benign or malignant in nature. When compared with the gold standard of cancer diagnosis (i.e., surgical pathology), the sensitivity of the created model was 95.7% and its specificity was 87%, with an overall accuracy of 90.7% and an AUROC of 0.76. The sensitivity of the radiologists' reading in this cohort of patients was 86%, with a specificity of 46% and an overall accuracy of 79%. The deep learning-based model significantly increased the diagnostic accuracy by increasing specificity and reducing false positives readings. The model can also provide a conclusive reading of mammograms categorized as BI-RADS 0 in the radiologists' reports, thereby mitigating the need for further imaging studies prior to rendering the final diagnosis.

Keywords: *breast cancer; computer-assisted diagnosis; deep learning; mammogram reading*

INTRODUCTION

Over nine thousand screening mammograms are done at The Medical City (TMC) Pasig every year (annual census, unpublished data). Screening mammography is used to detect and diagnose breast cancer among women even in the absence of signs and symptoms and studies have shown that 0.4% to 0.5% of them are eventually found to have breast cancer.[1]

Although there is good concordance between the image interpretation of the radiologists and the histopathologic diagnosis of a breast lesion seen on a mammogram (unpublished data), the current set-up of diagnostic breast imaging interpretation is susceptible to human error brought about by difficulty in discerning tumors due to inherently dense breasts, reader fatigue, reader bias, and other technical factors.[2-4] Hence, the clinical value of computer-aided detection in the diagnosis of a breast lesion found on mammography needs to be explored.

With the use of image analysis and machine intelligence such as a computer-assisted diagnostic (CAD) system which can isolate possible malignant lesions in radiological images, specialists can be assisted in breast cancer diagnosis. With CAD, image filtering can reduce noise to properly prepare the scan for subsequent feature extraction, segmentation and classification. These techniques ensure the speed and quality of cancer pre-diagnosis, thus contributing significantly to the improvement of the diagnosis and treatment planning of breast cancer patients.

The use of an artificial intelligence technique through computer-aided detection in mammography screening for breast cancer has steadily increased in various parts of the world.[5] Computer-assisted methods have increased diagnostic accuracy by reducing false positives.[4,6,7] Different deep learning strategies have been used in many studies for image analysis in detection of breast cancers.[2,7-12] Artificial Neural Network (ANN) discriminating mass texture,[13] Convolutional Neural Networks (CNNs) detecting microcalcifications,[7] and CAD systems spotting architectural distortion in mammograms have been reported.[3]

Based on a review that use ANN on breast cancer detection via mammograms, the best performing method was Dheeba's particle swarm optimized wavelet neural network (PSOWNN) which had 94.167% sensitivity and 92.105% specificity.[14]

Jiang et al. proposed an automatic classification of breast mass lesions in mammographic images using transfer learning on GoogLeNet (AUC=0.88) and AlexNet (AUC=0.83).[15] The dataset used for the study was from the Breast Cancer Digital Repository (http://bcdr.inegi.up.pt) which is composed of 736 film mammographic images with biopsy-proven mass lesions (426 benign and 310 malignant). Samala et al. on the other hand used multi-task transfer learning DCNN on 2242 digitized screen-film mammograms (SFMs) and digital mammograms (DMs) with 2454 masses (1,057 malignant, 1,397 benign) and reached an AUC of 0.82.[16] Transfer learning in breast mammogram abnormalities classification using MobileNet and NASnet were also explored and achieved accuracies of 78.4% and 74.3% respectively.[17]

This study explores the use of AI-powered interpretation of abnormal lesions in mammography and its accuracy in detecting breast cancer compared with the current strategy of double-reading of radiologists at TMC. The imaging software utilizes deep learning technologies to automatically detect patterns from medical images. We hypothesize that this AI-powered tool can improve diagnostic interpretation, decrease the overall reading time by focusing mainly on the abnormal studies, and subsequently help the breast surgeons when making decisions related to definitive management.

METHODS

This is a retrospective cohort study conducted in a manner consistent with ICH–GCP guidelines for the conduct of clinical research study with human participants. The study protocol was approved t by the TMC Institutional Review Board in March 2017 (IRB #). The overall methodology is divided into the following parts: data collection, data anonymization, model-building for the classifier, and implementation (data de-anonymization and hospital intervention) as shown in Figure 1.

Data Collection

There are four categories of data for the study as shown in Figure 2: *pathology reading* (the gold standard), *patient information* (e.g. name, age), *radiologist reading* (observations, BI-RADs score), and *radiologic images* (mammograms). There are two groups of subjects in this study based on the gold standard, the pathology report – 'malignant' and 'benign' groups. The malignant-group's list of participants was taken from the Cancer Registry of TMC in 2017 and 2018. This list bearing the participants' name and personal identification numbers (PIN) was crosschecked with a list of mammograms conducted over the same time period. For patients with multiple mammograms, only those mammograms dated a few days or weeks (maximum of four weeks) prior to the biopsy procedure were the ones identified. The benign group consisted of those who also underwent biopsy based on the surgeon's clinical judgment, not listed in the malignant group. After identifying the list, the biopsy records, mammogram images, and the corresponding radiologist readings were downloaded from different hospital data sources and unified into a single dataset.

Data Tagging

Radiology BI-RADS (**Appendix 1A**) readings were categorized into benign or malignant and were matched with the gold standard.[11] For data labeling, BI-RADS 1, 2 and 3 readings were considered benign, and BI-RADS 4 and 5 as malignant. In this study, BI-RADS category 3 was categorized as a benign reading, and the corresponding recommendation for this category is not to proceed with biopsy but do a short-interval follow-up mammogram instead. Even though BI-RADS 4A and 4B have only >2-10% and 11-50% chance of malignancy, respectively, these BI-RADS categories were automatically grouped under the malignant readings as biopsy is recommended for those categories. These were noted and tabulated as part of the final dataset for collection.

The BI-RAD readings were obtained from the imaging reports of the group of radiologists in the TMC Breast Center. The mammogram reports are from a consensus reading of at least two radiologists; in cases of discordance, a third and fourth reader come-in to read the mammogram independently. Readings were considered false negative if the malignant (pathology) group was categorized by the radiologist as BI-RADS 1, 2 or 3. Mammogram readings were considered false positive if the benign group was read by the radiologist as BI-RADS 4 or 5.

Data Anonymization

Data collection and anonymization were done in accordance with RA 10173, also known as the Data Privacy Act of 2012.[18] The four categories of data were stored in different databases and were unified into a master list and referenced properly. The overall process of data collection and unification, tagging, and anonymization is summarized in Figure 2.

AI-Powered Radiology (AIR) System Cancer Detection

After anonymization, transfer learning was used to train the data using the InceptionV4 [21] architecture with the model (latest weights) on the object detection as the starting weights. The InceptionV4 model was used as the base model for transfer learning, i.e. training the base model on mammograms dataset having two labels: benign and malignant. The labels were based on the biopsy diagnosis which served as the gold standard. Each training batch took approximately 3–72 hours (depending on the training hyperparameters) on a server with 62GB RAM and Tesla P100 graphics processing unit. The dataset was split into two sets: a training set (80%) and a testing set (20%). The training set was used on the algorithm to train the model, whereas the test set was used to measure the performance of the model after training. Sampling was done from both labels, with an equal number of labels for each training epoch. The summary of steps can be visualized in **Appendices 2A and 3A**.

Sample screenshots of the demo interface running the trained model on a mammogram image with malignant and benign diagnosis is shown in Figure 4. The demo interface is initially loaded, then the image for analysis is browsed by selecting the browse button. After loading, the image is processed by selecting the process button. After a couple of seconds, the results are shown as probabilities with the final diagnosis being the higher probability score. Figure 5 shows other images run through the demo interface.

Analysis and Performance Metrics

The performance of the model was quantified by calculating accuracy, sensitivity, and specificity, with the pathology results serving as the reference standard. The accuracy, sensitivity, and specificity of the radiologists' readings and of the AIR algorithm and their capability to correctly recognize malignant tissue from benign tissue using mammogram images were measured using standard statistical formulae.[20]

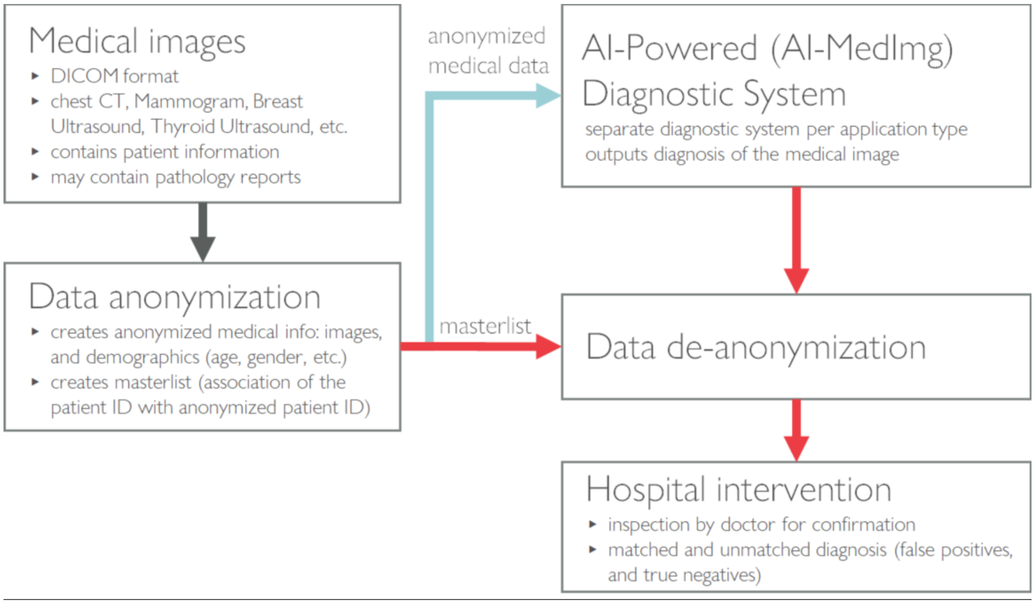
Medical images
DICOM format
chest CT, Mammogram, Breast Ultrasound, Thyroid Ultrasound, etc.
contains patient information
may contain pathology reports
anonymized medical data
AI-Powered (AI-MedImg) Diagnostic System
separate diagnostic system per application type
outputs diagnosis of the medical image
Data anonymization
creates anonymized medical info: images, and demographics (age, gender, etc.)
creates masterlist (association of the patient ID with anonymized patient ID)
masterlist
Data de-anonymization
Hospital intervention
inspection by doctor for confirmation
matched and unmatched diagnosis (false positives, and true negatives)

Figure 1. Overall pipeline of the methodology

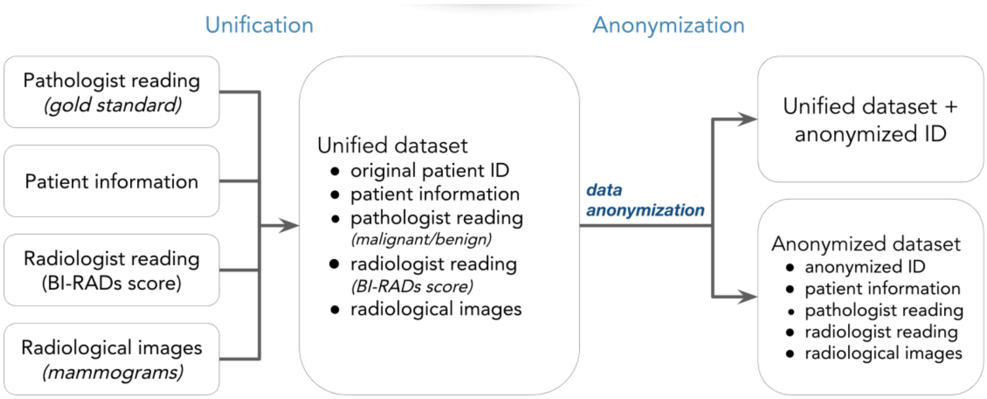

Figure 2. Data collection and preparation process

RESULTS

A total of 215 women with breast biopsies and mammograms conducted at TMC in 2017 and 2018 were included in this study. Of these, 159 had malignant pathology reports, and 56 had benign pathology reports. A summary of the dataset distribution is shown in Table 1. Sample mammograms for each pathology are shown in Figure 3.

The most frequent BI-RADS Classification (as shown in Table 1) identified in this set of mammograms was BI-RADS 0, with 63 of the 215 patients (29%) having this mammogram result. Fifty eight of the 215 (27%) had a BI-RADS classification of 5. Fifty two of the 215 (24%) had BI-RADS 4. Of these 52, 10 (19%) were classified as BI-RADS 4 (no subtype), 15 (29%) as BI-RADS 4A, 15 (29%) as BI-RADS 4B, and 12 (23%) as BI-RADS 4C. Twenty four of the 215 (11%) had BI-RADS 2. Fifteen of the 215 (7%) had BI-RADS 6, one of the 215 had BI-RADS 3, and 2 of the 215 (1%) had BI-RADS 1.

The performance of the radiologist's readings of benign (BI-RADS 1, 2 or 3) or malignant (BI-RADS 4 or 5) based on the biopsy results were as follows: 79% accuracy, 86% sensitivity, and 46% specificity (Tables 2 and 3).

Nine hundred images were generated from these 215 patients' mammograms: 620 images from the malignant mammograms and 280 images from the benign mammograms. Seven hundred twenty images were used to train the model, and 180 were used to test. The overall accuracy of the trained model was 90.7%, with a sensitivity of 95.2% and specificity of 87.0%. Performances of the models and radiologists' reading are summarized in Table 4.

Table 1. Summary of the dataset, showing the number of images with their corresponding BI-RADs assessment and biopsy diagnosis.

BI-RADs	Image Count	Mammograms		Patient Count	Patients	
		Biopsy			Biopsy	
		Malignant	Benign		Malignant	Benign
0	250	130	120	63	35	28
1	5	1	4	2	1	1
2	114	40	74	24	14	10
3	6	0	6	1	0	1
4	45	27	18	10	6	4
4A	51	16	35	15	6	9
4B	59	45	14	15	14	1
4C	53	53	0	12	12	0
5	256	256	0	58	58	0
6	61	52	9	15	13	2
Total	900	620	280	215	159	56

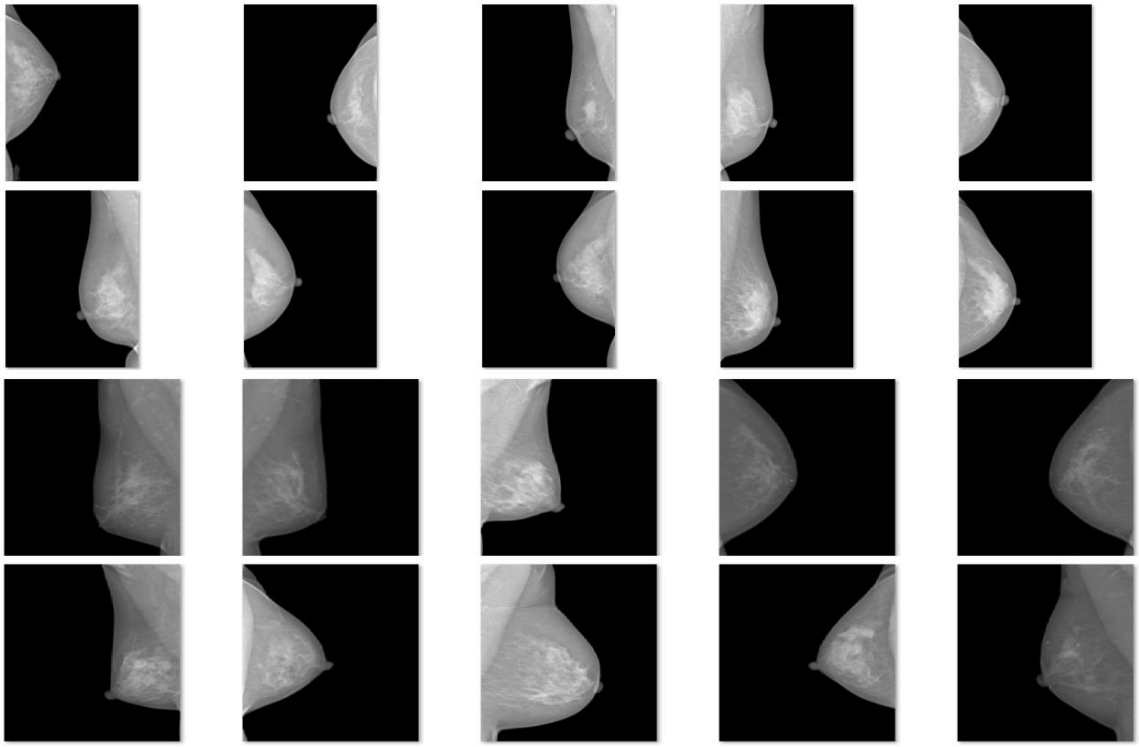

Figure 3. Sample mammogram images of patients with benign (top) and malignant (bottom) tumors

Table 2. Radiologists reading of mammogram

BI-RADs	Malignant	Benign
1	1	1
2	14	10
3	0	1
4	38	14
5	58	0
	111	26

Table 3. Radiologist reading vis-à-vis histopathology of breast lesions seen on mammograms

	MALIGNANT PATHOLOGY	BENIGN PATHOLOGY	Total
BI-RADS 4 OR 5	96	14	110
BI-RADS 1, 2, OR 3	15	12	27
Total	111	26	137

Table 4. Summary of accuracy, sensitivity and specificity of radiologists and trained model

	Accuracy	Sensitivity	Specificity
Radiologists (BI-RADs 1-5)	79.0%	86.0%	46.0%
Trained Model (BI-RADs 0-6)	90.7%	95.2%	87.0%

DISCUSSION

At TMC, standard mammogram images are interpreted using the double reading strategy, and a report is considered final when a consensus is reached by at least two interpreting radiology physicians. The results of our study showed an overall 86% sensitivity given the assumptions and simplifications in the computation considering the different BI-RADs categories as either benign or malignant. Therefore, an 86% chance exists that a woman with a malignant breast lesion seen on mammogram will receive a BI-RADS 4 or 5 reading and will subsequently undergo a diagnostic biopsy. The results also showed a 14% chance that a malignant breast lesion seen on mammogram will receive a BI-RADS 1, 2, or 3 reading and will neither be biopsied nor removed. When there is a need for additional imaging evaluation or if previous images are not available at the time of reading, a BI-RADS 0 category is given and no information is provided on whether the imaging finding is benign or malignant; this often causes delays in clinical management.

The present study showed a much lower specificity of 46% based on the radiologists' reading of a mammogram with a breast lesion (Table 4). This result is expected of screening tests in general as there is a tendency to "overdiagnose" so as not to miss doing interventions on early cancer lesions. The price to pay for ruling in a potentially deadly disease in a screening test is the high rate of false positive readings. The overall accuracy of the radiologists' BI-RADS reading was 79%. Therefore, some women undergo biopsy due to BI-RADS 4 or 5 readings, but their final biopsy result is usually benign.

The model trained in this study showed a 95.2% sensitivity, which is higher than the sensitivity of radiologists' readings. The specificity of the study model was also higher at 87% compared to the 46% specificity of the radiologists BI-RADS reading. The model was trained to "read" a breast lesion on mammogram images as either benign or malignant only and has no "unclassifiable" reading akin to the BI-RADS 0 of human readers. With further training and exposure to more abnormal breast imaging findings, the model will most likely be able to better recognize true negative (benign) mammograms. The overall accuracy of the study CAD in predicting whether a breast lesion seen on mammogram is benign or malignant was 90.7% with an AUROC of 0.76. Higher values suggest that the model is better at predicting the correct category, in this case malignant and benign lesions. This value can still improve by exploring other data augmentation techniques, further adjusting the hyperparameters, and increasing the dataset.

In a study conducted by Rashad Kamal, et al, four main factors were identified as causes for the misdiagnosis of breast carcinomas: 1) patient, 2) tumor, 3) technical, and 4) provider factors.[10] Patient factors are those limitations when breasts are inherently dense. Tumor factors include subtle carcinomas that display features that are difficult to assess. Technical factors are those related to positioning, exposure, and processing of images. Finally, provider factors are related to wrong perceptions or misinterpretations by interpreting physicians or radiologists. In the same study, suggestions were made to avoid missing breast carcinomas. Aside from clinical correlation, good acquisition technique, and use of adjunct imaging studies, the study recommended double reading and the use of CAD, a form of AI, to minimize misdiagnosis.

Other studies have compared the performance of a CAD system to the interpretation of a radiologist.[8,5,21,22] In a study conducted by Thijs Kooi, et al., AI using a deep CNN could classify regions of interest of malignant soft tissue lesions in a mammography as effectively as experienced radiologists.[12] Another study which compared two AI systems, a state-of-the-art system in CAD and a CNN, reported that experienced human readers and the CNN had similar performance.[7] A review of CAD in a mammography shows that this AI technique may decrease oversights made by interpreting radiologists.[22] Another study showed that radiologists read mammogram examinations more effectively when they are assisted by an AI system than when they are unaided, without costing additional time.[23] A recent review by Mendelson on AI on Breast Imaging concluded that AI can support breast images in diagnosis and patient management. The authors stated that the limitation of AI is that AI is still unreliable for decisions that may affect survival.[22]

CONCLUSION AND RECOMMENDATION

The results of this pilot study showed that the deep learning model trained on reading mammogram images with corresponding histopathologic results of biopsied lesions is potentially more accurate, sensitive and specific than the current double-reading of radiologists in interpretation of abnormal breast lesions seen on mammogram images.

Future studies should be conducted to validate these findings in a prospective manner to test whether aiding radiologists' readings with the trained model (via a CAD tool) will be more efficient and accurate than the current standard in diagnosing breast lesions seen on mammographic images. Another avenue which can be explored is the performance of other base models (such as YoloV4) for the transfer learning method and comparing its performance with the trained model.

LIMITATIONS OF THE STUDY

Being a pilot study in TMC on the use of AI in imaging, simplification and presuppositions were made in computation of the sensitivity, specificity and accuracy of the radiologists' reading. To remove the complexities, the BI-RADS classification was simply categorized as either "benign" or "malignant". BI-RADS 6 category was not included in the computation in the radiologists' reading. Future studies need to take into consideration the inherent intricacies of the BI-RADS classification.

Exploration of the architecture of InceptionV4, and other architectures such as ResNet, AlexNet was not performed in this study

ACKNOWLEDGEMENTS

The authors would like to acknowledge and thank the following people and institutions: Aimee Rarugal, Chino Rodriguez, Charles Chan, Al Asaad, Nico Gonzales, Rachel Domingo, Christina Jamerlan, Ghizelle Pulusan, Jen Azuela, Jo Pabona, Noel Pinton, Nelson Tiongson, and Jayvenson Javier.

REFERENCES

[1] Laia Domingo, Solveig Hofvind, Rebecca A Hubbard, Marta Román, David Benkeser, Maria Sala, and Xavier Castells. Cross-national comparison of screening mammography accuracy measures in US, Norway, and Spain. European radiology, 26(8):2520–2528, 2016.

[2] Bonnie C Yankaskas, Michael J Schell, Richard E Bird, and David A Desrochers. Reassess- ment of breast cancers missed during routine screening mammography: a community-based study. American Journal of Roentgenology, 177(3):535–541, 2001.

[3] Alejandro Rodríguez-Ruiz, Elizabeth Krupinski, Jan-Jurre Mordang, Kathy Schilling, Sylvia H Heywang-Köbrunner, Ioannis Sechopoulos, and Ritse M Mann. Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology, 290(2):305–314, 2018.

[4] Alyssa T Watanabe, Vivian Lim, Hoanh X Vu, Richard Chim, Eric Weise, Jenna Liu, Willam G Bradley, and Christopher E Comstock. Improved Cancer Detection Using Artificial Intelligence: a Retrospective Evaluation of Missed Cancers on Mammography. Journal of Digital Imaging. 2019.

[5] Janine Katzen and Katerina Dodelzon. A review of computer aided detection in mammography. Clinical imaging, 52:305–309, 2018.

[6] Jan-Jurre Mordang, Tim Janssen, Alessandro Bria, Thijs Kooi, Albert Gubern-Mérida, and Nico Karssemeijer. Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks. In International Workshop on Breast Imaging, pages 35–42. Springer, 2016.

[7] Thijs Kooi, Albert Gubern-Merida, Jan-Jurre Mordang, Ritse Mann, Ruud Pijnappel, Klaas Schuur, Ard den Heeten, and Nico Karssemeijer. A comparison between a deep convolutional neural network and radiologists for classifying regions of interest in mammography. In International Workshop on Breast Imaging, pages 51–56. Springer, 2016.

[8] Manisha Bahl, Niveditha Pinnamaneni, Sarah Mercaldo, Anne Marie McCarthy, and Con- stance D Lehman. Digital 2d versus tomosynthesis screening mammography among women aged 65 and older in the united states. Radiology, 291(3):582–590, 2019.

[9] Farahnaz Sadoughi, Zahra Kazemy, Farahnaz Hamedan, Leila Owji, Meysam Rahmanikati- gari, and Tahere Talebi Azadboni. Artificial intelligence methods for the diagnosis of breast cancer by image processing: a review. Breast Cancer: Targets and Therapy, 10:219, 2018.

[10] RASHA M Kamal, NM Abdel Razek, MOHAMED A Hassan, and MOHAMED A Shaalan. Missed breast carcinoma; why and how to avoid. J Egypt Natl Canc Inst, 19(3):178–94, 2007.

[11] EA Sickles, CJ d'Orsi, LW Bassett, CM Appleton, WA Berg, ES Burnside, et al. ACR BI-RADS mammography. ACR BI-RADS atlas, breast imaging reporting and data system, 5:2013, 2013.

[12] Thijs Kooi, Geert Litjens, Bram Van Ginneken, Albert Gubern-Mérida, Clara I Sánchez, Ritse Mann, Ard den Heeten, and Nico Karssemeijer. Large scale deep learning for computer aided detection of mammographic lesions. Medical image analysis, 35:303–312, 2017.

[13] Lukas H Kus, Manish Shah, Spiro Eski, Paul G Walfish, and Jeremy L Freeman. Thyroid cancer outcomes in Filipino patients. Archives of Otolaryngology–Head & Neck Surgery, 136(2):138–142, 2010.

[14] J Dheeba, N Albert Singh, and S Tamil Selvi. Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach. Journal of biomedical informatics, 49:45–52, 2014.

[15] Fan Jiang, Hui Liu, Shaode Yu, and Yaoqin Xie. Breast mass lesion classification in mammograms by transfer learning. In Proceedings of the 5th international conference on bioinformatics and computational biology, pages 59–62, 2017.

[16] Ravi K Samala, Heang-Ping Chan, Lubomir Hadjiiski, Mark A Helvie, Caleb D Richter, and Kenny H Cha. Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets. IEEE transactions on medical imaging, 38(3):686–696, 2018.

[17] Lenin G Falconí, María Pérez, and Wilbert G Aguilar. Transfer learning in breast mammogram abnormalities classification with mobilenet and nasnet. In 2019 International Conference on Systems, Signals, and Image Processing (IWSSIP), pages 109–114. IEEE, 2019.

[18] Official Gazette. Republic act 10173.

[19] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of

residual connections on learning. In Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[20] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Under- standing transfer learning for medical imaging. In Advances in neural information processing systems, pages 3347–3357, 2019.

https://www.privacy.gov.ph/data-privacy-act/, 2012. Accessed: 2019-08-02.

[21] Jay A Baker, Eric L Rosen, Joseph Y Lo, Edgardo I Gimenez, Ruth Walsh, and Mary Scott Soo. Computer-aided detection (cad) in screening mammography: sensitivity of commercial cad systems for detecting architectural distortion. American Journal of Roentgenology, 181(4):1083–1088, 2003.

[22] Ellen B Mendelson. Artificial intelligence in breast imaging: potentials and limitations. American Journal of Roentgenology, 212(2):293–299, 2019.

[23] Laurence N Kolonel. Cancer incidence among Filipinos in Hawaii and the Philippines. National Cancer Institute monograph, 69:93–98, 1985.