Healthcare Analytics 3 (2023) 100186

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

A deep learning architecture with an object-detection algorithm and a convolutional neural network for breast mass detection and visualization

Steven J. Frank

*Med*A-Eye Technologies, Framingham, MA, United States of America*

ARTICLE INFO

Keywords:
Intelligence
Deep learning
Convolutional neural network
Decision support
Object detection
Radiology

ABSTRACT

This study presents an integrated deep learning architecture with an object-detection algorithm and a convolutional neural network (CNN) for breast mass detection and visualization. Mammograms are analyzed to identify and localize breast mass lesions to aid clinician review. Two complementary forms of deep learning are used to identify the regions of interest (ROIs). An object-detection algorithm, YOLO v5, analyzes the entire mammogram to identify discrete image regions likely to represent masses. Object detections exhibit high precision, but the object-detection stage alone has insufficient overall accuracy for a clinical application. A CNN independently analyzes the mammogram after it has been decomposed into subregion tiles and is trained to emphasize sensitivity (recall). The ROIs identified by each analysis are highlighted in different colors to facilitate an efficient staged review. The CNN stage nearly always detects tumor masses when present but typically occupies a larger area of the image. By inspecting the high-precision regions followed by the high-sensitivity regions, clinicians can quickly identify likely lesions before completing the review of the full mammogram. On average, the ROIs occupy less than 20% of the tissue in the mammograms, even without removing pectoral muscle from the analysis. As a result, the proposed system helps clinicians review mammograms with greater accuracy and efficiency.

1. Introduction

Breast cancer is one of the leading causes of death for women globally [1,2] and is the most frequently diagnosed non-skin cancer [3]. It occurs due to the uncontrolled growth of breast tissue, typically resulting in a lump or mass that may be detected at an early stage using different imaging modalities. Mammography, the most commonly employed modality, has contributed to sharp declines in mortality since the death rate peaked in 1989 [4,5].

Nonetheless, screening mammography misses about 20% of malignancies, often due to high breast density [6] and the low inherent sensitivity of the modality. Fibroglandular tissue and tumors have similar density, so the former may hide the latter in a mammogram image. These interpretive challenges are exacerbated by clinician fatigue, which can reduce accuracy further as radiologists are asked to shoulder ever-increasing workloads [7,8].

The potential for machine learning to combine the strengths of radiologists and computers has long been recognized. Indeed, the objective of much recent work has been to replace clinicians altogether [9], or to match or exceed human performance according to some benchmark [10,11]. Even efforts that do not seek explicitly to supplant radiologists may offer predictions that amount to clinical judgments [12,13]. Recognizing that such approaches have not actually reduced radiologists' workloads, developers of "decision-referral" systems seek to spare clinicians from reviewing exams with a high probability of being cancer-free [14–16]. Not surprisingly, the resulting efficiencies can come at the cost of an unacceptable reduction in sensitivity [17]. Indeed, the efficiencies may be illusory: Germany imposes a legal obligation on radiologists to review every mammogram [18], and elsewhere, institutional expectations may enforce similar policies in fact if not in law.

AI systems will achieve greater acceptance in radiology practice when they are perceived as having the clinician's back rather than doing their job. As a 2019 editorial in the journal *Radiology* observed, asking whether AI will replace radiologists is the wrong question. "The right answer is: Radiologists who use AI will replace radiologists who don't" [9].

To this end, the approach described here offers decision support by identifying tumor masses in mammograms in a manner consistent with common clinical practice, i.e., radiologist review of every image. The objective is to ensure that, if a mass is present, the clinician's attention is efficiently drawn to it.

2. Literature review

Assessing the utility of systems designed for decision support requires evaluation metrics that are both statistically valid and meaningful in measuring clinical usefulness. The latter critically depends on two

E-mail address: steve@medaeye.com.

https://doi.org/10.1016/j.health.2023.100186
Received 25 December 2022; Received in revised form 17 March 2023; Accepted 25 April 2023

factors: the reliability with which disease regions are highlighted and the utility of the tool in promoting review efficiency.

Prior studies rely on traditional analytics, which are based on similarity between detected regions and their "ground truth" counterparts. These metrics are widely used to score semantic segmentation systems that assign category labels to regions of an image, quantifying the pixel-level accuracy of the labels. They include "intersection over union" (IoU), precision, and sensitivity (or recall). IoU is defined as follows:

$$\text{IoU} = \frac{P \cap T}{P \cup T}$$

where P is the set of pixels representing the prediction and T corresponds to the ground truth (e.g., the disease regions in an image). Precision represents the proportion of pixels classified as positive (e.g., as lesion pixels) that are, in fact, positive while sensitivity corresponds to the proportion of all positive pixels correctly classified as such. In terms of true positives (*TP*), false positives (*FP*), and false negatives (*FN*),

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

The limitations of these conventional similarity metrics have been recognized in connection with tasks such as region-level identifications, where pixel-level identity to the ground truth may not fully capture the significance of a detection [19]. Analogously, these analytics may have less relevance to clinical practice than whether the diagnostic regions of a medical image are visually identified, even if they are not marked with perfect fidelity.

Object-detection systems are often scored in terms of detection accuracy, defined as the ratio of true detected cases to the total number of cases [20]. Numerous studies based on the YOLO architecture [21–27] discussed here as well as on other object-detection models [28–32], for example, have reported detection rates for masses in mammograms ranging from 0.74 to over 0.99. These detection rates may not account for spurious detections — i.e., false positives, which, like false alarms, can lead to reviewer frustration and diminish confidence in a tool's effectiveness. Such shortcomings have long plagued healthcare automation systems [33–36]. Even if a more rigorous accuracy metric is used, it is often unclear how false positives and false negatives are defined; for example, they may reflect pixel-level or object-level errors.

3. Proposed model and analytic framework

The objective of the proposed system is to ensure that, if a mass is present, the clinician's attention is efficiently drawn to it. As illustrated in Fig. 1, two complementary neural network architectures are employed: an object detection stage and a probability mapping stage. Most of the time, a mass will appear within a tightly conforming zone highlighted as a high-precision region of interest (ROI) identified by the object-detection stage. The probability mapping stage defines ROIs with greater sensitivity but less precision; if a mass is not found within a high-precision ROI, it will almost certainly fall within a high-sensitivity ROI. To assess the proposed system's success in achieving the objectives of efficiency and reliability in decision support, we consider three analytic metrics in addition to the simple detection fraction: object-level sensitivity (recall), object-level precision, and an object-level accuracy metric that reflects false positives (FP, representing spurious detections) and false negatives (FN, representing masses that were missed) as well as true positives (TP). True negatives are null quantities that do not contribute to this overall accuracy metric, i.e., they constitute the normal background.

$$\text{Accuracy} = \frac{TP}{TP + FP + FN}$$

Here precision and sensitivity are defined as above but measure object detections rather than visual similarity, i.e., they are computed at an object level rather than the pixel level.

Object-level sensitivity, precision, and accuracy score success in achieving reliability. To measure efficiency, we utilize a fourth metric. Recognizing that lesions tend to be small and rarely occupy a significant fraction of the tissue in a mammogram, we measure the average fraction of tissue occupied by ROIs across a test set. If the high-precision and high-sensitivity ROIs collectively occupy too much of an image on average, the clinician's experience will approach conventional, *ab initio* review and the system provides little practical benefit.

At the same time, pixel-level visual similarity may be considered to assess how tightly a ROI encloses a mass. This enables comparison between high-precision and high-sensitivity ROIs in terms of promoting review efficiency.

4. Materials and data

The two system stages shown in Fig. 1 may be organized in parallel, as shown, or sequentially. Because the object-detection stage is so much faster than probability mapping, parallel execution by different processing entities generally offers little performance benefit. Indeed, a preliminary object-detection pass allows mammograms to be prioritized for the slower probability-mapping stage.

Most commonly used object-detection algorithms are based on deep learning, i.e., neural networks with multiple layers of processing that extract progressively higher-level features from image data. Two-stage object detectors first generate "region proposals" where objects may appear in an image, then analyze features within these regions to identify objects on which the system has been trained. Examples of two-stage architectures include the regional convolutional neural network (R-CNN), Faster R-CNN and Mask R-CNN. One-stage object detectors directly predict bounded regions around identified objects without the region proposal step. The best-known one-stage architecture is the aptly named YOLO ("You Only Look Once"), which exists in several versions — the latest being v8. Although single-stage detectors are considered less accurate in recognizing irregularly shaped objects or a group of small objects [37], experimentation across both types of detectors revealed YOLO to perform best for this application; and within the YOLO series, YOLO v5 outperformed its siblings based on the overall accuracy metric discussed earlier.

The probability mapping stage is trained to classify small, overlapping tiles derived from a mammogram as tumor or non-tumor breast tissue. The prediction probabilities are combined and averaged at the pixel level, and the ROI consists of pixels whose average probabilities exceed a decision boundary — typically, but not necessarily, the boundary used during training. The resulting ROI is a tissue segmentation map of likely tumor regions.

As described in [38], systematic testing across tile sizes generally reveals a single best-performing size; here, because the intrinsic capabilities of the CNN are being assessed, performance is scored using the traditional pixel-level similarity metrics IoU, precision, and sensitivity. The tile size and the degree of tile overlap determine the resolution of the ROI. Hence, if multiple tile sizes achieve similar scores, the smallest is selected. The degree of overlap for mapping purposes depends on several factors: the desired resolution, accuracy, processing time, and the size of the smallest expected mass relative to the tile size. Without overlap, adjacent tiles might split a mass, with neither tile including enough of its feature area to elicit a positive prediction. Overlap also ensures that pixel-level probabilities reflect the contribution of more than one tile, reducing random error. Finally, during training, overlap is important for data augmentation.

Numerous CNN architectures were tested, including ResNet50, ResNet101, VGG16, MobileNet v2, EfficientNetB0, and the five-layer CNN described in earlier work [38,39]. EfficientNetB0 slightly outperformed the five-layer CNN, and both achieved significantly better performance than the other architectures tested; here performance was measured in terms of tile-level classification accuracy and assessment of the ROI using the IoU similarity metric. A potential drawback of

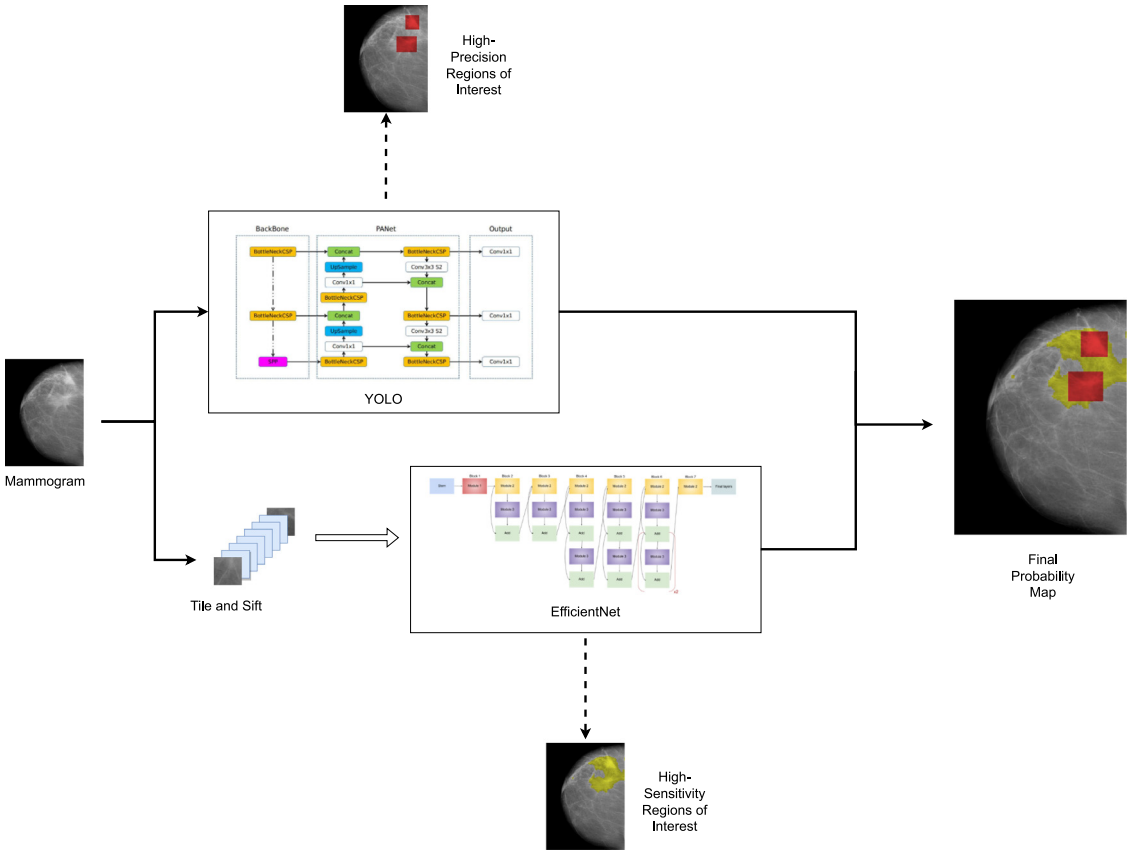

Fig. 1. Organization and operation of stacked YOLO and EfficientNetB0 models. The YOLO stage circumscribes features identified as masses with high precision (red regions). The EfficientNetB0 stage identifies ROIs with less precision but high sensitivity (yellow regions). The output analysis represents the union of these ROIs, with the high-precision red ROIs overwriting the yellow. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

EfficientNet is its fixed image input size. The designers attempt to accommodate a range of image sizes by making eight variants available with different input requirements (ranging from 224 × 224 pixels to 600 × 600 pixels). The apparent floor of 224 × 224 pixels can be reduced further, however, by upsampling smaller tiles, which will not degrade prediction accuracy so long as the rescaling does not disturb critical feature-level spatial relationships – a condition that can only be verified through testing.

Once trained, YOLO v5 and EfficientNetB0 operate independently on new candidate mammograms: YOLO v5 on the entire image, which is reduced in size during processing, and EfficientNetB0 on small tiles derived from the entire mammogram. Two publicly available datasets were used in this study. The CBIS-DDSM dataset [40] contains 2907 mammograms from 1555 patients, including pixelwise annotations for ROIs corresponding to masses and calcifications. The dataset conveniently includes training and testing subsets of 1216 and 311 images, respectively. CBIS-DDSM images have an average size of about 3000 × 4800 pixels. The INBreast dataset [41] contains 410 mammograms, 108 of which include mass abnormalities. Images have an average size of about 3300 × 4100 pixels. The INBreast ground-truth annotations generally surround rather than conform to lesion shapes, and so have limited value for training.

For all experiments involving the YOLO stage, we used stochastic gradient descent optimization with a learning rate of 0.01 and a batch size of 8. The YOLO model was pretrained on the COCO dataset, which has a native resolution of 640 × 640 pixels; although larger and a few smaller dimensions were tested, best performance was obtained by shrinking the mammograms to 640 × 640 pixels despite the loss of resolution.

For the EfficientNetB0 stage, we tested tile sizes ranging from 50 × 50 pixels to 200 × 200 pixels in steps of 50 pixels, as well as the native 224 × 224 input size. The 150 × 150 tiles, upsampled to 224 × 224 pixels for training and testing, performed best as measured by IoU against the ground-truth annotations of the CBIS-DDSM validation set. This modest upsampling, in other words, did not undermine CNN performance. Data augmentation consisted of random horizontal and vertical flips as well as brightness variation. More significant data augmentation resulted from the degree of tile overlap. Other than removal of background regions, source images were not preprocessed; we have found that equalization, contrast adjustment and similar operations actually degrade results, presumably due to loss of information. We used an Adam optimizer, a learning rate of 0.0001, a batch size of 16, and no pre-training.

Labeled tiles were drawn separately from lesion and normal tissue regions of the training mammograms. Masses typically occupy a small portion of the overall tissue image area, which contributes to the challenge of identifying them by traditional visual inspection. This also produces a class imbalance between lesion and normal tissue. To address this discrepancy, we obtained similar numbers of lesion and normal-tissue tiles by overlapping the lesion tiles by 80%. In this way we obtained 13,500 lesion and 13,500 normal-tissue tiles for training. While the underlying data remains imbalanced, this did not impair the ability to train models that produce useful predictions; the trained models do not exhibit overfitting to lesion regions, for example.

For both stages, we trained first using the 1216 images in the CBIS-DDSM mass-lesion training set, with 50 of the 361 mammograms in the test set used for validation. We performed four-fold cross validation using three additional training and test sets derived from the entire CBIS-DDSM dataset. As noted, the INBreast images are generally unsuitable for training, but we tested against the 108 mammograms from that dataset which include lesions.

Table 1
Performance of YOLO object detection on CBIS-DDSM and INBreast datasets.

YOLO v5 Alone – Object Level							
Dataset	# Test images	Detection fraction	False positives	False negatives	Object-Level Recall	Object-Level Precision	Overall accuracy
CBIS-DDSM	311	.76	62	76	.69	.74	.56
INBreast	108	.76	7	26	.74	.91	.69

Table 2
Performance of YOLO object detection on CBIS-DDSM and INBreast datasets.

YOLO v5 Alone – Pixel-Level Performance on CBIS-DDSM dataset		
Pixel-Level recall	Pixel-Level precision (mAP@0.5)	F_1 Score
.83	.74	.78

Table 3
Performance of two-stage system on CBIS-DDSM and INBreast datasets.

YOLO + EfficientNetB0			
Dataset	Mass detection fraction	Case detection fraction	ROI tissue fraction
CBIS-DDSM	.91	.97	.132
INBreast	.96	1.0	.183

For the YOLO stage, validation loss reached a minimum after fewer than 27 epochs for all cross-validation folds; for the EfficientNet stage, the traditional IoU similarity metric was used as the loss function, and the minimum validation loss occurred below 20 epochs. Trained using a decision boundary of 0.5, the EfficientNet stage exhibited sufficient sensitivity without additional measures, possibly due to the imbalance favoring the lesion class [42]. Were this not the case, we would have attempted to increase sensitivity by lowering the decision boundary used for testing or combining prediction probabilities from multiple models trained on different training/test splits [43].

5. Results

We first review the performance of the YOLO-based object-detection stage.

The figures provided in Table 1 represent cross-validation averages; the spread across the four folds was less than 5%. The detection fraction (i.e., the ratio of masses detected to the total number of masses in the test images) was identical for both datasets. As noted, however, this quantity does not reflect erroneous detections and is easily adjusted by varying the confidence level at which a detection is triggered. Ultimately, we identified a confidence threshold of 0.3 as optimal based on criteria derived from the other metrics. In particular, this confidence threshold nearly balanced false positives and false negatives while reducing overall accuracy only slightly from the maximum value obtained across all confidence thresholds tested. False negatives (misses) reduce object-level sensitivity, i.e., the fraction of masses actually detected, meaning that more clinician review time is spent on the larger regions of lower probability. False positives, as observed earlier, undermine confidence in a tool's effectiveness.

Were the YOLO stage used alone, safety concerns would favor a focus on sensitivity. In this two-stage system, however, with lesions missed by YOLO quite likely to fall within high-sensitivity ROIs, attempting to balance error types for object detection seemed reasonable. Using the confidence threshold of 0.3 on the smaller INBreast dataset skewed the error distribution to false negatives, but recall still exceeded the value obtained for CBIS-DDSM.

For the YOLO stage, precision is important at object and pixel levels; that is, there should be few false-positive object detections and the detections themselves should conform tightly to the detected masses. Standard mammograms include bilateral craniocaudal (CC) and mediolateral oblique (MLO) views, and for each patient-level case, the CBIS-DDSM and INBreast datasets generally include both views. We can assess the detection fraction at the mass level (i.e., the proportion of masses partly or entirely contained within a ROI and visibly identified to a reviewing clinician) or the case level (the fraction of patients whose masses are identified in at least one of the mammogram views). Table 3 shows both quantities as well as the ROI tissue fraction.

More than 90% of masses fall at least partly within the high-sensitivity ROI. Of the 419 test images analyzed in both datasets, only six were missed at a case level; all of these were in the CBIS-DDSM dataset. The INBreast dataset comprises more recently acquired full-field digital mammograms, which better represent current practice than the CBIS-DDSM digitized film mammograms. Despite the high detection fraction, the ROIs occupy a small proportion of breast tissue in the mammograms. (The tissue fraction, it should be noted, excludes background (black) regions of the images; the reported values represent fractions of the tissue depicted in a mammogram.)

Table 2 reports pixel-level metrics that indicate both how tightly the YOLO detection boundaries surround a mass (precision) as well as how effectively they capture the entire mass (recall). The mean pixel-level precision of 0.74 exhibited by successful YOLO detections reflects the inevitable mismatch between a regular bounding box and an uneven mass feature; even perfect detections will encompass some precision-degrading excess area (see Fig. 2).

Frequently, the high-sensitivity ROIs encompass non-diagnostic pectoral muscle, which exhibits mammogram densities similar to fibrous or lesioned breast tissue. Efforts have been made to computationally identify and exclude pectoral muscle in diagnostic procedures such as quantitative analysis of breast parenchyma [44,45], and similar approaches have been proposed for analysis of mammograms for cancerous lesions [46]. Unfortunately, as shown in Fig. 3, particularly in MLO views, a mass can overlap with the pectoral muscle in the two-dimensional image. Excluding the pectoral muscle will remove the mass as well. As a result, despite the detrimental effect on ROI tissue fraction, the proposed system retains pectoral tissue in the analysis.

6. Discussion

Mammograms can be challenging even for experienced radiologists to read. Modern digital mammography equipment provides substantially greater dynamic range ("latitude") and contrast resolution than film mammography. Still, subtle abnormalities, particularly in a background of dense breast tissue, remain difficult to identify visually or computationally. In this study, we demonstrate the advantage of stacking complementary CNN architectures so as to maximize their strengths in analyzing mammograms. The practical objective is to help clinicians use their time efficiently and to recognize masses that might otherwise elude visual detection.

Object-detection systems identify suspicious lesions with high precision and closely conforming boundaries, but are prone to false positives and false negatives. As noted, some previous studies report very high detection rates, and we easily obtained a detection fraction of 0.94 using our YOLO stage on the CBIS-DDSM test image set by reducing the confidence threshold from 0.3 to 0.075; but false positives soared from 62 to 238, reducing object-level precision from 0.74 to 0.18 — essentially an invitation to hunt for the true needle in a haystack of erroneous detections. The overall accuracy metric reported in Table 1

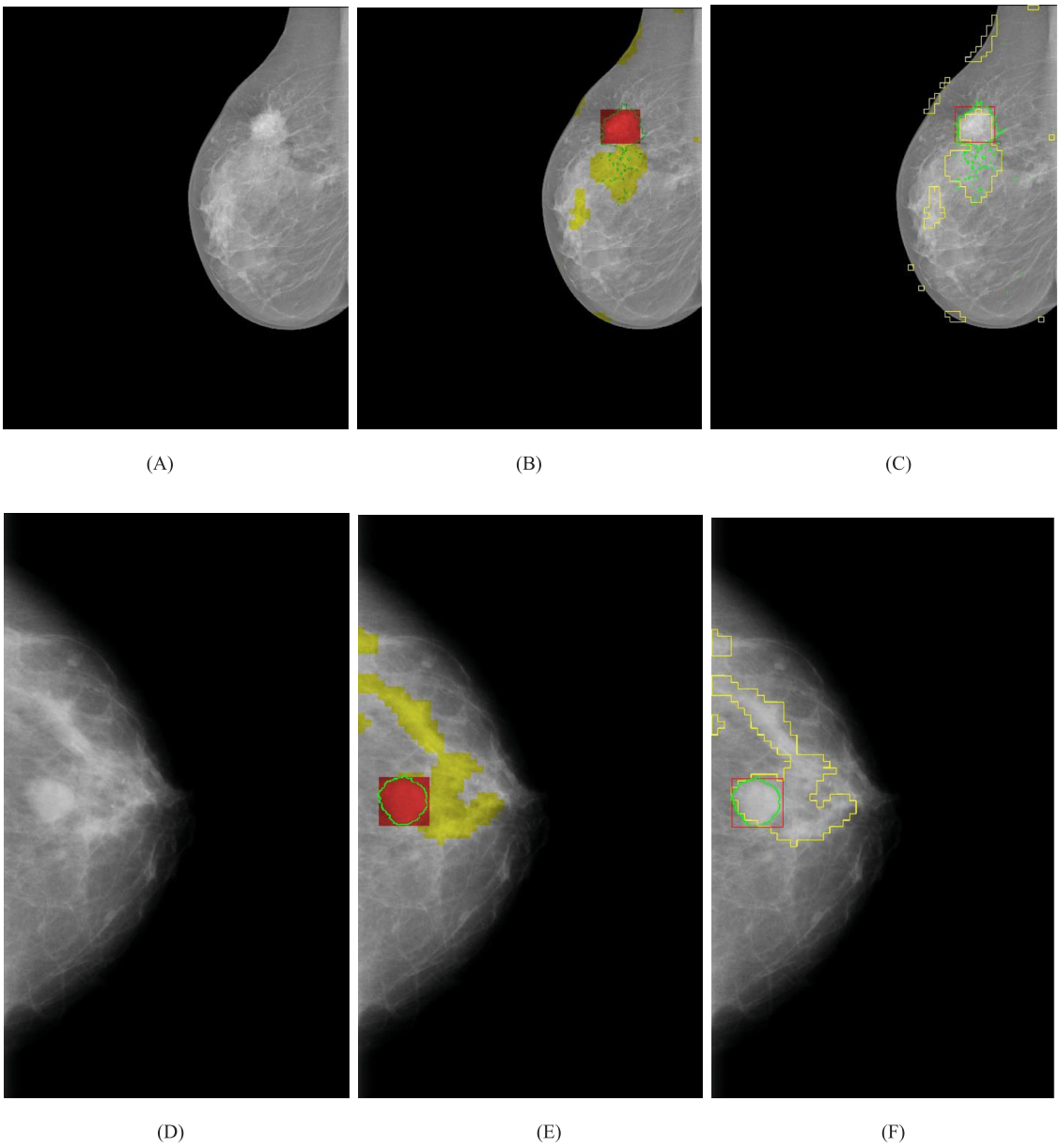

Fig. 2. *Top*: (A) Mediolateral oblique mammogram from INBreast dataset. (B) High-precision region colored red, high-sensitivity regions colored yellow, mass outlined in green. Minuscule mass features elude object detections but are fully within a high-sensitivity region. (C) Alternative representation with colored outlines rather than filled regions. *Bottom*: (D) Bilateral craniocaudal mammogram from CBIS-DDSM dataset. (E) High-precision region colored red, high-sensitivity regions colored yellow, mass outlined in green. (F) Alternative representation with colored outlines rather than filled regions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

is a harsh critic, penalizing false positives and false negatives equally. Suboptimal as the reported values may seem, in our experiments, YOLO outperformed other commonly used object-detection architectures in successfully locating mass lesions in mammograms. Hence the need for a second stage.

Using a traditional CNN to derive pixel-level probabilities from image subregions captures what object detection misses but with much less precision. Combining these architectures captures the benefits of high precision without sacrificing sensitivity. Specifically, the above performance statistics suggest that about 76% of the time, the reviewer's primary task will be completed almost immediately; if a mass is detected by the YOLO stage, it is always the primary mass and will be well defined within the ROI. A scan for additional abnormalities will begin with the high-sensitivity ROIs, which will very likely reveal any additional lesions. A reviewer's attention is thus drawn to the regions where abnormalities most likely lurk. At the same time, suspicious features outside a ROI are unlikely to represent masses; combined with other patient risk factors, this finding can help tip the balance away from unnecessary biopsies.

The proposed system is straightforwardly applied to other mammogram features such as calcifications. These are small calcium deposits that appear as white specks in the mammogram, and may signify ductal carcinoma in situ or early breast cancer if they appear in specific patterns. The YOLO architecture has been successfully employed to detect calcifications and distinguish them from masses [27,47,48], as have more traditional CNN architectures [49].

Screening mammography has long demonstrated its value in preventing deaths from breast cancer. Nonetheless, the National Cancer Institute cautions that the benefits of mammography "need to be balanced against its harms," which include false positive and false negative results [6]. The approach proposed here is intended to address both types of harm.

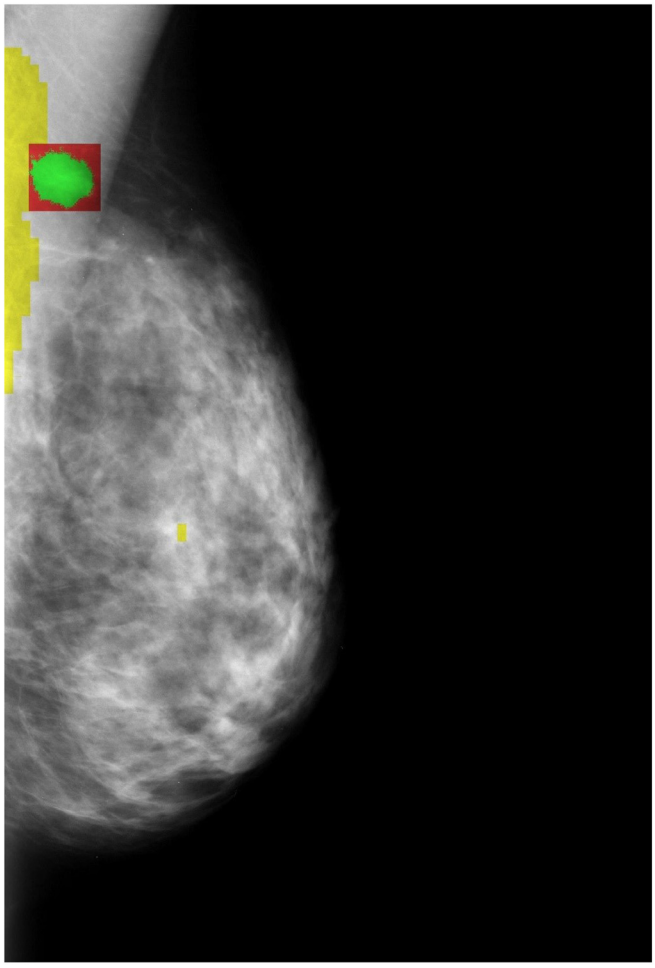

Fig. 3. Mediolateral oblique mammogram illustrating a lesion (green) in a tissue region aligned with the pectoral muscle, exclusion of which would also preclude detection of the mass. The red high-precision ROI has a pixel-level precision relative to the ground-truth mass of 0.63 due in part to the rectangular boundary, but the extra intercepted area is not visually significant. The yellow high-sensitivity ROIs do not occupy a significant portion of the breast tissue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Declaration of competing interest

There are no conflicts of interest, competing interests or overlap with prior publications in connection with this paper.

Data availability

Other than the publicly available CBIS-DDSM and INBreast datasets cited below, no data was used for the research described in the article.

References

[1] Breast Cancer Facts & Figures 2022-2024, Am. Cancer Soc., 2022, https://www.cancer.org/research/cancer-facts-statistics/breast-cancer-facts-figures.html.

[2] T. Jia, Y. Liu, Y. Fan, L. Wang, E. Jiang, Association of healthy diet and physical activity with breast cancer: Lifestyle interventions and oncology education, Front. Public Heal. 10 (2022) http://dx.doi.org/10.3389/fpubh.2022.797794.

[3] R.L. Siegel, K.D. Miller, H.E. Fuchs, A. Jemal, Cancer statistics, 2022, CA. Cancer J. Clin. 72 (2022) http://dx.doi.org/10.3322/caac.21708.

[4] R.E. Hendrick, J.A. Baker, M.A. Helvie, Breast cancer deaths averted over 3 decades, Cancer. 125 (2019) http://dx.doi.org/10.1002/cncr.31954.

[5] J.L., M.H. Zi Lin Lim, Peh Joo Ho, Alexis Jiaying Khng, Yen Shing Yeoh, Amanda Tse Woon Ong, Benita Kiat Tee Tan, Ern Yu Tan, Su-Ming Tan, Geok Hoon Lim, Jung Ah Lee, Veronique Kiak-Mien Tan, Jesse Hu, Mammography screening is associated with more favourable breast cancer tumour characteristics and better overall survival: case-only analysis of 3739 Asian breast cancer patients, BMC Med. (2022) 239, http://dx.doi.org/10.1186/s12916-022-02440-y.

[6] Mammograms - NCI, Natl. Cancer Inst. (2021) https://www.cancer.gov/types/breast/mammograms-fact-sheet, (Accessed October 10, 2022).

[7] E.U. Ekpo, M. Alakhras, P. Brennan, Errors in mammography cannot be solved through technology alone, Asian Pacific J. Cancer Prev. 19 (2018) http://dx.doi.org/10.22034/APJCP.2018.19.2.291.

[8] B.M. Geller, E.J.A. Bowles, H.Y. Sohng, R.J. Brenner, D.L. Miglioretti, P.A. Carney, J.G. Elmore, Radiologists' performance and their enjoyment of interpreting screening mammograms, Am. J. Roentgenol. 192 (2009) 361–369, http://dx.doi.org/10.2214/AJR.08.1647.

[9] C.P. Langlotz, Will artificial intelligence replace radiologists? Radiol. Artif. Intell. 1 (2019) http://dx.doi.org/10.1148/ryai.2019190058.

[10] S. Gadgil, M. Endo, E. Wen, A.Y. Ng, P. Rajpurkar, CheXseg: Combining expert annotations with DNN-generated saliency maps for X-ray segmentation, 2021, https://arxiv.org/abs/2102.10484v2, (Accessed June 4, 2022).

[11] P. Lakhani, B. Sundaram, Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks, Radiology 284 (2017) http://dx.doi.org/10.1148/radiol.2017162326.

[12] T. Schaffter, D.S.M. Buist, C.I. Lee, Y. Nikulin, D. Ribli, Y. Guan, W. Lotter, Z. Jie, H. Du, S. Wang, J. Feng, M. Feng, H.E. Kim, F. Albiol, A. Albiol, S. Morrell, Z. Wojna, M.E. Ahsen, U. Asif, A. Jimeno Yepes, S. Yohanandan, S. Rabinovici-Cohen, D. Yi, B. Hoff, T. Yu, E. Chaibub Neto, D.L. Rubin, P. Lindholm, L.R. Margolies, R.B. McBride, J.H. Rothstein, W. Sieh, R. Ben-Ari, S. Harrer, A. Trister, S. Friend, T. Norman, B. Sahiner, F. Strand, J. Guinney, G. Stolovitzky, L. Mackey, J. Cahoon, L. Shen, J.H. Sohn, H. Trivedi, Y. Shen, L. Buturovic, J.C. Pereira, J.S. Cardoso, E. Castro, K.T. Kalleberg, O. Pelka, I. Nedjar, K.J. Geras, F. Nensa, E. Goan, S. Koitka, L. Caballero, D.D. Cox, P. Krishnaswamy, G. Pandey, C.M. Friedrich, D. Perrin, C. Fookes, B. Shi, G. Cardoso Negrie, M. Kawczynski, K. Cho, C.S. Khoo, J.Y. Lo, A.G. Sorensen, H. Jung, Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms, JAMA Netw. Open. 3 (2020) http://dx.doi.org/10.1001/jamanetworkopen.2020.0265.

[13] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzebski, T. Fevry, J. Katsnelson, E. Kim, S. Wolfson, U. Parikh, S. Gaddam, L.L.Y. Lin, K. Ho, J.D. Weinstein, B. Reig, Y. Gao, H. Toth, K. Pysarenko, A. Lewin, J. Lee, K. Airola, E. Mema, S. Chung, E. Hwang, N. Samreen, S.G. Kim, L. Heacock, L. Moy, K. Cho, K.J. Geras, Deep neural networks improve radiologists' performance in breast cancer screening, IEEE Trans. Med. Imaging. 39 (2020) http://dx.doi.org/10.1109/TMI.2019.2945514.

[14] K. Dembrower, E. Wåhlin, Y. Liu, M. Salim, K. Smith, P. Lindholm, M. Eklund, F. Strand, Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study, Lancet Digit. Heal. 2 (2020) http://dx.doi.org/10.1016/S2589-7500(20)30185-0.

[15] J.L. Raya-Povedano, S. Romero-Martín, E. Elías-Cabot, A. Gubern-Mérida, A. Rodríguez-Ruiz, M. Álvarez-Benito, AI-based strategies to reduce workload in breast cancer screening with mammography and tomosynthesis: A retrospective evaluation, Radiology. 300 (2021) http://dx.doi.org/10.1148/radiol.2021203555.

[16] T. Kyono, F.J. Gilbert, M. van der Schaar, Improving workflow efficiency for mammography using machine learning, J. Am. Coll. Radiol. 17 (2020) http://dx.doi.org/10.1016/j.jacr.2019.05.012.

[17] C. Leibig, M. Brehmer, S. Bunk, D. Byng, K. Pinker, L. Umutlu, Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis, Lancet Digit. Heal. 4 (2022) e507–e519, http://dx.doi.org/10.1016/S2589-7500(22)00070-X.

[18] H. Kiros, Doctors using AI catch breast cancer more often than either does alone | MIT technology review, MIT Technol. Rev. (2022) https://www.technologyreview.com/2022/07/11/1055677/ai-diagnose-breast-cancer-mammograms/, (Accessed October 10, 2022).

[19] Y. Zhang, S. Mehta, A. Caspi, Rethinking semantic segmentation evaluation for explainability and model selection, 2023, arXiv https://arxiv.org/abs/2101.08418.

[20] F. Samuelson, C. Abbey, Using relative statistics and approximate disease prevalence to compare screening tests, Int. J. Biostat. 12 (2016) http://dx.doi.org/10.1515/ijb-2016-0017.

[21] M.A. Al-Masni, M.A. Al-Antari, J.M. Park, G. Gi, T.Y. Kim, P. Rivera, E. Valarezo, S.M. Han, T.S. Kim, Detection and classification of the breast abnormalities in digital mammograms via regional convolutional neural network, in: Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS, 2017, http://dx.doi.org/10.1109/EMBC.2017.8037053.

[22] M.A. Al-masni, M.A. Al-antari, J.M. Park, G. Gi, T.Y. Kim, P. Rivera, E. Valarezo, M.T. Choi, S.M. Han, T.S. Kim, Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system, Comput. Methods Programs Biomed. 157 (2018) http://dx.doi.org/10.1016/j.cmpb.2018.01.017.

[23] G. Hamed, M. Marey, S. Amin, M.F. Tolba, Automated breast cancer detection and classification in full field digital mammograms using two full and

cropped detection paths approach, IEEE Access. (2021) http://dx.doi.org/10.1109/ACCESS.2021.3105924.

[24] A. Baccouche, B. Garcia-Zapirain, C.C. Olea, A.S. Elmaghraby, Breast lesions detection and classification via YOLO-based fusion models, Comput. Mater. Contin. 69 (2021) http://dx.doi.org/10.32604/cmc.2021.018461.

[25] G.H. Aly, M. Marey, S.A. El-Sayed, M.F. Tolba, YOLO based breast masses detection and classification in full-field digital mammograms, Comput. Methods Programs Biomed. 200 (2021) http://dx.doi.org/10.1016/j.cmpb.2020.105823.

[26] Y. Su, Q. Liu, W. Xie, P. Hu, YOLO-LOGO: A transformer-based YOLO segmentation model for breast mass detection and segmentation in digital mammograms, Comput. Methods Programs Biomed. 221 (2022) 106903, http://dx.doi.org/10.1016/j.cmpb.2022.106903.

[27] A. Baccouche, B. Garcia-Zapirain, A.S. Elmaghraby, An integrated framework for breast mass classification and diagnosis using stacked ensemble of residual neural networks, Sci. Rep. 12 (2022) 12259, http://dx.doi.org/10.1038/s41598-022-15632-6.

[28] D. Ribli, A. Horváth, Z. Unger, P. Pollner, I. Csabai, Detecting and classifying lesions in mammograms with deep learning, Sci. Rep. 8 (2018) http://dx.doi.org/10.1038/s41598-018-22437-z.

[29] J. Peng, C. Bao, C. Hu, X. Wang, W. Jian, W. Liu, Automated mammographic mass detection using deformable convolution and multiscale features, Med. Biol. Eng. Comput. 58 (2020) http://dx.doi.org/10.1007/s11517-020-02170-4.

[30] Y. Li, L. Zhang, H. Chen, L. Cheng, Mass detection in mammograms by bilateral analysis using convolution neural network, Comput. Methods Programs Biomed. 195 (2020) http://dx.doi.org/10.1016/j.cmpb.2020.105518.

[31] V.K. Singh, H.A. Rashwan, S. Romani, F. Akram, N. Pandey, M.M.K. Sarker, A. Saleh, M. Arenas, M. Arquez, D. Puig, J. Torrents-Barrena, Breast tumor segmentation and shape classification in mammograms using generative adversarial and convolutional neural network, Expert Syst. Appl. 139 (2020) http://dx.doi.org/10.1016/j.eswa.2019.112855.

[32] W. Ansar, A.R. Shahid, B. Raza, A.H. Dar, Breast cancer detection and localization using mobilenet based transfer learning for mammograms, in: Commun. Comput. Inf. Sci., 2020, http://dx.doi.org/10.1007/978-3-030-43364-2_2.

[33] M. Cvach, Monitor alarm fatigue: An integrative review, Biomed. Instrum. Technol. 46 (2012) http://dx.doi.org/10.2345/0899-8205-46.4.268.

[34] J.P. Keller, Clinical alarm hazards: A top ten health technology safety concern, J. Electrocardiol. 45 (2012) http://dx.doi.org/10.1016/j.jelectrocard.2012.08.050.

[35] W.T.M. Au-Yeung, A.K. Sahani, E.M. Isselbacher, A.A. Armoundas, Reduction of false alarms in the intensive care unit using an optimized machine learning based approach, Npj Digit. Med. 2 (2019) http://dx.doi.org/10.1038/s41746-019-0160-7.

[36] Neda Asadi, Fatemeh Salmani, Narges Asgari, Mahin Salmani, Alarm fatigue and moral distress in ICU nurses in COVID-19 pandemic, BMC Nurs. 21 (2022) 125, http://dx.doi.org/10.1186/s12912-022-00909-y.

[37] P. Soviany, R.T. Ionescu, Optimizing the trade-off between single-stage and two-stage deep object detectors using image difficulty prediction, in: Proc. - 2018 20th Int. Symp. Symb. Numer. Algorithms Sci. Comput. SYNASC 2018, 2018, http://dx.doi.org/10.1109/SYNASC.2018.00041.

[38] S.J. Frank, Resource-frugal classification and analysis of pathology slides using image entropy, Biomed. Signal Process. Control. 66 (2021) 102388, http://dx.doi.org/10.1016/j.bspc.2020.102388.

[39] S. Frank, Accurate diagnostic tissue segmentation and concurrent disease subtyping with small datasets, Res. Sq. (2022) http://dx.doi.org/10.21203/RS.3.RS-1777977/V1.

[40] R.S. Lee, F. Gimenez, A. Hoogi, K.K. Miyake, M. Gorovoy, D.L. Rubin, Data descriptor: A curated mammography data set for use in computer-aided detection and diagnosis research, Sci. Data. 4 (2017) http://dx.doi.org/10.1038/sdata.2017.177.

[41] I.C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M.J. Cardoso, J.S. Cardoso, Inbreast: Toward a full-field digital mammographic database, Acad. Radiol. 19 (2012) http://dx.doi.org/10.1016/j.acra.2011.09.014.

[42] B. Juba, H.S. Le, Precision-recall versus accuracy and the role of large data sets, in: 33rd AAAI Conf. Artif. Intell. AAAI 2019, 31st Innov. Appl. Artif. Intell. Conf. IAAI 2019 9th AAAI Symp. Educ. Adv. Artif. Intell. EAAI 2019, 2019, http://dx.doi.org/10.1609/aaai.v33i01.33014039.

[43] S.J. Frank, Accurate diagnostic tissue segmentation and concurrent disease subtyping with small datasets, J. Pathol. Inform. 14 (2023) 100174, http://dx.doi.org/10.1016/j.jpi.2022.100174.

[44] X. Ma, J. Wei, C. Zhou, M.A. Helvie, H.-P. Chan, L.M. Hadjiiski, Y. Lu, Automated pectoral muscle identification on MLO-view mammograms: Comparison of deep neural network to conventional computer vision HHS public access, Med. Phys. 46 (2019) 2103–2114, http://dx.doi.org/10.1002/mp.13451.

[45] J. Wei, C. Zhou, H.-P. Chan, H.M. Lubomir, Y. Lu, X. Ma, Fully automated pectoral muscle identification on MLO-view mammograms with deep convolutional neural network, 2018, http://dx.doi.org/10.1117/12.2318124.

[46] K.S. Camilus, V.K. Govindan, P.S. Sathidevi, Computer-aided identification of the pectoral muscle in digitized mammograms, J. Digit. Imaging. 23 (2010) http://dx.doi.org/10.1007/s10278-009-9240-6.

[47] A.A. Yurdusev, K. Adem, M. Hekim, Detection and classification of microcalcifications in mammograms images using difference filter and Yolov4 deep learning model, Biomed. Signal Process. Control. (2023) 104360, http://dx.doi.org/10.1016/j.bspc.2022.104360.

[48] K. Wang, N. Khan, R. Highnam, Automated segmentation of breast arterial calcifications from digital mammography, in: Int. Conf. Image Vis. Comput. New Zeal, 2019, http://dx.doi.org/10.1109/IVCNZ48456.2019.8960956.

[49] Q. Lin, W.-M. Tan, S.-Y. Zhu, Y. Huang, Q. Xiao, Y.-Y. Xu, Y.-T. Jin, Z.-M. Shao, Y.-J. Gu, B. Yan, K.-D. Yu, Deep learning-based microcalcification detection and classification of mammography for diagnosis of breast cancer, SSRN Electron. J. (2022) http://dx.doi.org/10.2139/ssrn.4001825.