10-23-2020

Characterization of indeterminate breast lesions on B-mode ultrasound using automated machine learning models

Shuo Wang

Sihua Niu

Enze Qu

Flemming Forsberg

Annina Wilkes

See next page for additional authors

Authors

Shuo Wang, Sihua Niu, Enze Qu, Flemming Forsberg, Annina Wilkes, Alexander Sevrukov, Kibo Nam, Robert F. Mattrey, Haydee Ojeda-Fournier, and John R. Eisenbrey

Characterization of indeterminate breast lesions on B-mode ultrasound using automated machine learning models

Shuo Wang, MS[1,2], Sihua Niu, MD, PhD[3], Enze Qu, MD[4], Flemming Forsberg, PhD[2], Annina Wilkes, MD[2], Alexander Sevrukov, MD[2], Kibo Nam, PhD[2], Robert F. Mattrey MD[5], Haydee Ojeda-Fournier, MD[6], John R. Eisenbrey, PhD[2]

1. Department of Biomedical Engineering, Drexel University, Philadelphia, PA

2. Department of Radiology, Thomas Jefferson University, Philadelphia, PA

3. Department of Ultrasound, Peking University People's Hospital, Beijing, China

4. Department of Ultrasound, The Third Affiliated Hospital of Sun Yat-Sen University, Guangzhou, China

5. Department of Radiology, UT Southwestern, Dallas, TX 75390, USA

6. Department of Radiology, University of California, San Diego, CA 92037, USA

Corresponding author:

John Eisenbrey, PhD

Department of Radiology

Thomas Jefferson University

132 S. 10th Street

Philadelphia, PA 19017, USA

e-mail: John.Eisenbrey@jefferson.edu

Funded in part by NIH R01 CA140338 and DoD W81XWH-11-1-0630

For the original clinical trial that data was obtained from, the ultrasound contrast agent was provided by Lantheus Medical Imaging and the ultrasound scanner provided by GE Healthcare.

Conflict of Interest: None

Key words: Artificial Intelligence; Machine learning; Deep learning; Ultrasound imaging; Breast lesions

Abstract

Purpose: While mammography has excellent sensitivity for the detection of breast lesions, its specificity is limited. Adjunct screening with ultrasound may partially alleviate this issue, but also increases false positives, resulting in unnecessary biopsies. This study investigated the use of Google AutoML Vision (Mountain View, CA), a commercially available machine learning service, to both identify and characterize indeterminate breast lesions on ultrasound.

Methods: B-mode images from 253 independent cases of indeterminate breast lesions scheduled for core biopsy were used for model creation and validation. The performances of two sub-models from AutoML Vision, the image classification model and object detection model were evaluated, while also investigating training strategies to enhance model performances. Pathology from the patient's biopsy were used as a reference standard.

Results: The image classification models trained under different conditions demonstrated areas under the precision recall curve (AUC) ranging from 0.85 to 0.96 during internal validation. Once deployed, the model with highest internal performance demonstrated a sensitivity of 100% (95% confidence interval (CI) of 73.5-100%), specificity of 83.3% (CI=51.6-97.9%), positive predictive value (PPV) of 85.7% (CI=62.9-95.5%), and negative predictive value (NPV) of 100% (CI non-evaluable) in an independent dataset. The object detection model demonstrated lower performance internally during development (AUC=0.67) and during prediction in the independent dataset (sensitivity=75.0% (CI=42.8-94.5), specificity=80.0% (CI=51.9-95.7), PPV=75.0% (CI=50.8-90.0), NPV=80.0% (CI=59.3-91.7%)), but was able to demonstrate the location of the lesion within the image.

Conclusions: Two models appear to be useful tools for identifying and classifying suspicious areas on B-mode images of indeterminate breast lesions.

Keywords: Artificial Intelligence; Machine learning; Deep learning; Ultrasound imaging; Breast lesions

Introduction

Breast cancer remains a primary health concern with 271,270 new cases diagnosed and more than 42,260 deaths in 2019 in the United States alone.[1] When the patient presents with metastases, the 5-year survival rate is only 26%.[2] However, early detection along with appropriate therapy can reduce mortality significantly.[3] Screening mammography remains the best modality for breast cancer detection with an overall sensitivity > 85%. However, in women with dense breasts, which make up more than 40% of women in the United States, the sensitivity lowers to as low as 48 %.[4] While adjunct screening with ultrasound imaging improves the sensitivity for cancer detection, the cost is reduced specificity: increased non-cancer recalls and more benign biopsies.[5]

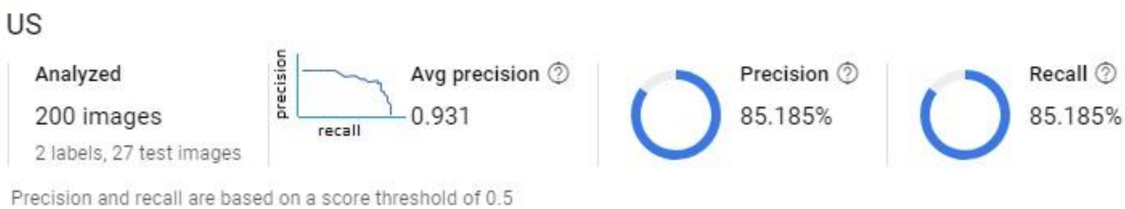
The Breast Imaging Reporting and Data System (BI-RADS®) is used by radiologists to classify breast lesions into several risk categories with different expected probabilities of malignancy. The course of clinical management is based on risk categories[6], with malignancy confirmed by biopsy. Nonetheless, even with using the BI-RADS data, inter and intra observer variability exists in classifying lesions and over 70% of all breast biopsy results are benign.[7] Thus, a better approach to differentiate between benign and malignant lesions from ultrasound images is needed.

The use of artificial intelligence (AI) in radiology has the potential to reduce costs, save time, and improve diagnostic accuracy.[8] Multiple studies have shown that deep learning algorithms (one type of AI) outperform experienced radiologists in the diagnosis of breast lesions with 5-13% larger area under the receiver operating characteristic (ROC) curves.[9,10,11] However, using deep learning algorithms requires a large amount of data (e.g., 5,000-10,000 training images) and training a new deep learning algorithm is both time-consuming and expensive. Several commercially AI programs are available providing an opportunity to overcome these barriers. Google AutoML Vision (Google, Mountain View, CA) is a machine learning service from Google Cloud Platform that runs deep learning algorithms online and performs image-classification and image-recognition tasks on cloud services, reducing the need for expensive hardware. It enables a customized model to be created quickly by leveraging transfer learning and neural architecture search technologies, which can lead to more accurate results with less misclassifications than other generic machine learning services.[12,13] In addition, due to the transfer learning component, which takes the advantages of lower-level features from pre-trained convolutional neural networks (CNN), significantly fewer images are required for algorithm training.[11]

Several sub-models are currently available for beta testing including an image classification mode and an object detection model. These models may provide distinct but useful roles within the field of radiology. The image classification model can train models to classify images (in this example cancer vs. not cancer), while the object detection model can be used to detect objects within an image and then assign a confidence score for a specific classification (in this example the likelihood of lesion being cancerous). Each of these sub-models perform self-validation and self-testing during the training process and generate model performance reports based on the training data (Figure 1).

While this technology has been used for a variety of product management applications, its use in radiological applications is relatively unexplored.[12,13] Thus, the purpose of this study was to evaluate the performance of both AutoML Vision's image classification and object detection models for the characterization of intermediate breast masses imaged with B-mode ultrasound. Specifically, we strove to identify the performance of AutoML's image classification and object detection mass for classifying breast masses as cancerous or non-cancerous in a population of suspicious masses scheduled for tissue biopsy. The influence of category balancing and image cropping on model performance was also investigated.

(a)



(b)

Important parameters	Description
Score thresholds	A minimal score for model to classify images to its correct labels. Score range: 0 to 1
Average Precision (AUC)	How well model performs across all score thresholds, area under precision-recall tradeoff curve. Range: 0 to 1
Precision (Positive Predict Value)	Higher precision, fewer false positives. Increase score threshold increase precision but lower recall. Range: 0 to 1
Recall (Sensitivity)	Higher recall, fewer false negatives. Lower score threshold increase recall but lower precision. Range: 0 to 1

Figure 1. (a) A model performance report is generated after each training process (b) Parameter descriptions and their equivalent ROC terminologies.

Material and Methods

Clinical studies

To create training datasets for the AI image classification and object detection models, ultrasound images were extracted from two previous clinical studies. The first study was a multi-center clinical trial that was approved by the Institutional Review Boards of Thomas Jefferson University (TJU) and The University of California, San Diego (UCSD) and conducted between January 2011 and December 2015 in which contrast-enhanced ultrasound was used to characterize indeterminate breast masses scheduled for biopsy.[14,15] The second study was approved by the Institutional Review Boards of TJU and conducted between May 2014 and February 2016, in which a contrast-enhanced ultrasound technique was used to predict the response of breast cancer to neoadjuvant chemotherapy[16]. All patients from both studies provided written informed consent before participating. The imaging data for both studies were acquired using a commercially available Logiq 9 scanner (GE Healthcare, Waukesha, WI) equipped with a 4D10L probe and imaging parameters were optimized on an individual basis according to good clinical practice. There were 236 women enrolled in the first clinical study with an average age of 52 ± 13 years. The average lesion cross-sectional areas for malignant and benign lesions were 190.1 ± 35.7 mm^2 and 124.1 ± 15.5 mm^2, respectively. The second clinical study enrolled 17 participants who had invasive ductal carcinomas with an average age of 52.9 ± 10.4 years and an average lesion cross-sectional area of 604.6 ± 460.7 mm^2. In total, there were 253 cases. For this AI processing study, 242 patient cases with available biopsy results (reference standard) were selected. Within these 242 cases, 21 cases were then excluded by a blinded radiologist due to poor image quality resulting in 154 unique patients with benign breast lesions and 67 unique patients with malignant breast lesions (221 in total).

Data preprocessing

The B-mode ultrasound data were originally stored in DICOM format. A radiologist (S.N) with more than 10 years of experience in breast ultrasound who was blinded to pathology results selected representative views from each CINE loop for the 221 cases. The DICOM data were viewed with RadiAnt DICOM Viewer (4.6.9, Medixant, Poznan, Poland) software and selected images were stored into JPG format in order to meet the input format requirements for Google AutoML Vision. Images were further cropped using Matlab (2016a, The Mathworks Inc., Natick, MA) to generate three different groups of training data: Annotated (A; with black and white scale, depth scale, GE label and ultrasound image), de-Annotated (deA; scales and GE label were removed, ultrasound images only), and Lesion Only (LO; lesions were extracted from the ultrasound images). Example images for each three training groups are shown in Figure 2.

Based on model recommendations, 26 out of the 221 cases (19 malignant and 7 benign cases corresponding to 11% of the patients) were reserved to form an independent prediction dataset to evaluate the models' performance. In order to augment our prediction dataset, a second radiologist (E.Q) with over 10 years of experience in breast ultrasound selected 5-7 image from each of the 26 test cases. This resulted in a final prediction dataset of 154 images for prediction testing. The same prediction dataset was used to evaluate all models from both image classification and object detection. Additionally, findings were grouped on a lesion by lesion basis to evaluate model intra-reader agreement (i.e., the ability to predict malignancy in separate images from the same case).

Image Classification Model Training

The Google AutoML Vision Image Classification Model was first investigated for its ability to differentiate benign (non-cancerous) from malignant (cancerous) breast lesions within the population of suspicious masses referred for biopsy. This model requires input training data of at least 100 images from each outcome group for training. However, as there were only 48 unique patients with malignant lesions remaining in the overall dataset after excluding the 19 malignant cases that were used for independent testing, a radiologist (S.N) selected at least two images from the malignant lesion dataset. Consequently, the final training data for the image classification model consisted of 147 images of benign breast lesions and 117 images of malignant lesions (264 images in total).

The training data for the model was slightly unbalanced (with 147 in the benign group and 117 in the malignant group), which may impact the performance of the model.[17] Thus, 30 random benign images were removed from the data set in order to compare the impact of unbalanced training (147 benign lesion images vs. 117 images of malignant lesions) relative to balanced training (117 benign lesion images vs. 117 malignant lesion images) on the performance of the model. Therefore, in addition to three different training groups (Annotated, de-Annotated, and Lesion Only; Figure 2), 6 customized models were trained. These groups are summarized in Table 1.

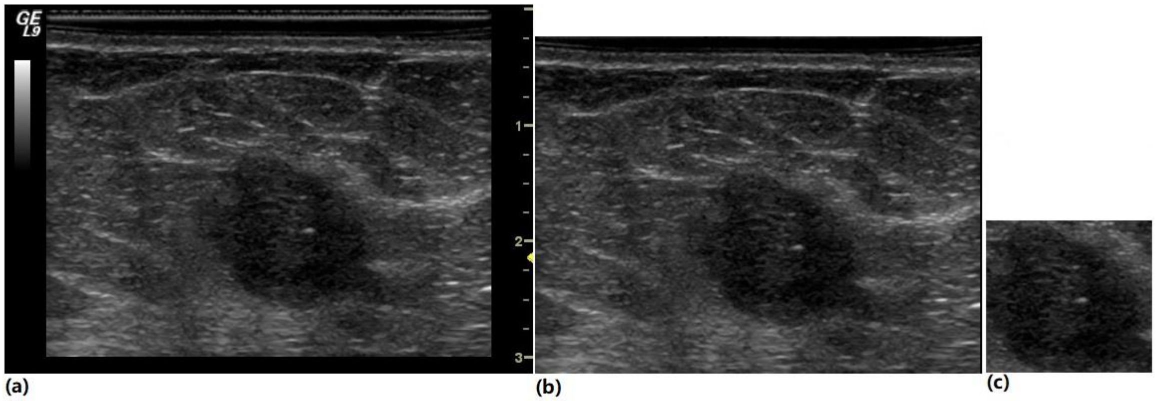


Figure 2. Example of the varying degrees of image cropping showing (a) the annotated image (A) containing the black and white scale bar, depth scale, GE label and ultrasound image, (b) the deAnnotated image (deA), in which the scales and GE label were removed leaving only the full ultrasound image, and (c) the lesion only (LO) image consisting of only the cropped breast mass.

Table 1. Summary of training data sets used for unbalanced (UB) and balanced (B) conditions. A stands for annotated images, deA stands for de-annotated images, and LO stands for lesion only images.

Unbalanced training	
Customized model	**Training Data Information (Number of benign lesion images, number of malignant lesion images, image group)**
A_UB	147 Benign, 117 Malignant, Annotated
deA_UB	147 Benign, 117 Malignant, deAnnotated
LO_UB	147 Benign, 117 Malignant, Lesion Only
Balanced training	
Customized model	**Training Data Information (Number of benign lesion images, number of malignant lesion images, image group)**
A_B	117 Benign, 117 Malignant, Annotated
deA_B	117 Benign, 117 Malignant, deAnnotated
LO_B	117 Benign, 117 Malignant, Lesion Only

Object Detection Model Training

The Google AutoML Vision Object Detection Model was investigated to determine the ability of this algorithm to first identify the suspicious breast mass, then subsequently assign a risk score on the likelihood of the image containing breast cancer. To train the object detection model, the same training data (147 benign and 117 malignant breast lesion images) as well as the same prediction images (154 breast images) described above were utilized. Data was first uploaded into Google Cloud Storage and then an Excel file that contained pathways for importing each image was generated from Python. The object detection model processes training image data within the model by using bounding boxes and labels to select objects that were important and intended to be detected inside an image. Therefore, only the full annotated images were imported into the model. Following upload, the model was trained by a blinded radiologist to identify the scale bars and manufacturer labels (as an algorithm validation check) and either malignant or benign masses within the three cropping approaches described above. An example of this training is provided in Figure 3.

Figure 3. Example figure showing image uploading and object identification training. Annotated images were imported into the object detection model during training and image labeling performed within the model. Labels were then manually added as shown on the left side by placing rectangle bounding boxes to on the desired objects as shown on the right side.

Evaluation of Model Performance

The performance of each model was evaluated using results from the participant's tissue biopsy as a reference standard. Performance reporting was separated by internal performance (self-reported by the model during training) and external prediction within the dataset reserved for testing. For internal validation, the area under the precision recall curve, sensitivity, specificity, negative predictive value, and positive predictive value were all reported with 95% confidence intervals. Model agreement was calculated for each of the six image classification models and the object detection model by quantifying the rate of agreement amongst images taken from the same lesion for each of the 26 external prediction cases. All statistical analysis was performed in GraphPad Prism Version 8.0 (San Diego, CA) with comparisons across multiple groups performed using a one-way ANOVA and direct comparisons between individual groups determined using a Student's t-test. Statistical significance was determined using $p < 0.05$.

Results

Image Classification Model Performance

Following training of the image classification model, internal performance reports were generated for each of the training conditions summarized in Table 1. Model performance reports from these six conditions are shown in Table 2. For external validation the model was deployed, and the 154 independent images analyzed. Figure 4 shows one prediction example from a model providing confidence scores for different labels. In order to draw decisions from the prediction results, a confidence score of 0.72 was utilized. This cutoff criteria was initially optimized by the model software based on optimization of the ROC curve during training and adjusted to minimize the number of cases in which a decision could not be made, while also mimicking the prevalence of malignancy in the prediction dataset. The decision for the prediction (either malignant or benign) relied on the label that had a confidence score greater than 0.72. If a prediction generated a confidence scores lower than 0.72 or if it generated both malignant and benign labels higher than 0.72, the prediction was considered as a not-applicable (N/A) case. The sensitivity, specificity, positive predictive value, negative predictive value, 95% confidence interval values and number of N/A cases for the 154 prediction images at a confidence score threshold of 0.72 are shown in Table 3.

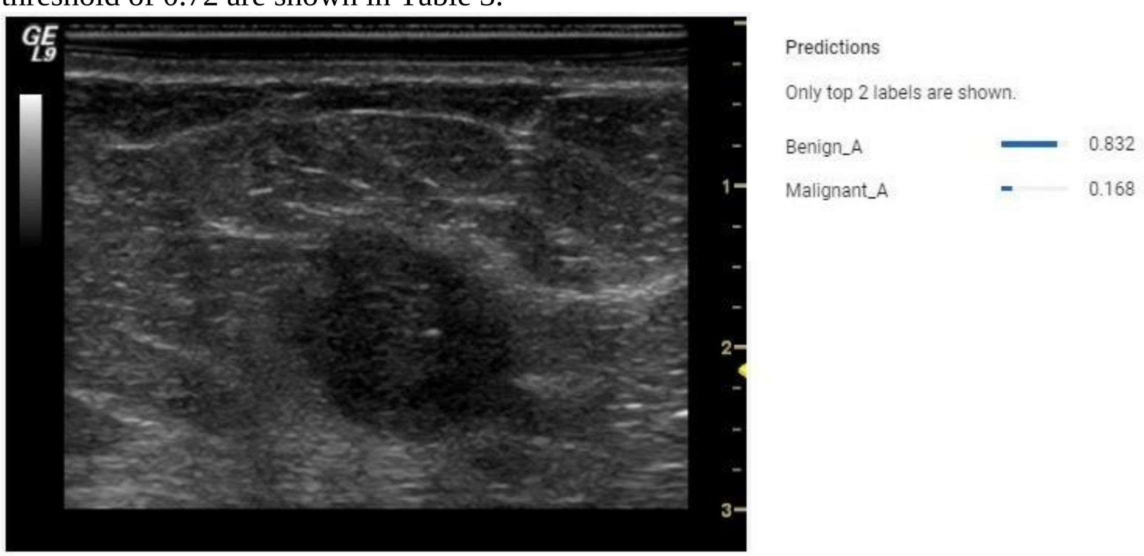


Figure 4. Example result from the image classification model during the post-training prediction phase of a benign mass. From the model's perspective, it had 83.2% certainty that the lesion was benign and 16.8% certainty that the lesion was malignant.

Table 2. Internal model performance reports obtained during model training from the 6 customized image classification models. AUC: Area under the precision recall curve. PPV:

Positive predictive value. NPV: Negative predictive value. 95% CI: 95% Confidence Interval.

Customized Models	AUC	Sensitivity(%) 95% CI	Specificity(%) 95% CI	PPV (%) 95% CI	NPV(%) 95% CI
A_UB	0.871	63.6 (30.8 - 89.1)	83.3 (51.6 - 97.9)	77.8 (47.8 – 93)	71.5 (52.4 - 85.1)
A_B	0.882	72.7 (39.0 – 94.0)	80.0 (51.9 - 95.7)	72.7 (47.6 - 88.7)	80 (59.6 - 91.6)
deA_UB	0.955	100.0 (73.5- 100.0)	86.7 (59.5 - 98.3)	85.7 (62.2 - 95.6)	100.0 non-evaluable*
deA_B	0.966	100.0 (73.5 – 100.0)	83.3 (51.6 - 97.9)	85.7 (62.9 - 95.5)	100.0 non-evaluable*
LO_UB	0.911	80 (44.4 - 97.5)	76.5 (50.1 - 93.2)	66.6 (44.5 - 83.2)	86.7 (64.7 - 98.9)
LO_B	0.853	81.8 (48.2 - 97.7)	76.9 (46.2 - 94.7)	75.0 (51.7 - 89.4)	83.4 (58.0 - 94.8)

* NPV non-evaluable due to lack of false negative cases.

Table 3. The calculated sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), for all customized image classification models as well as number of N/A cases in the prediction (post-training) dataset. 95% CI: 95% Confidence Interval.

Models	Sensitivity(%) 95% CI	Specificity(%) 95% CI	PPV(%) 95% CI	NPV(%) 95% CI	# of N/A
A_UB	75.2 (66.4 - 82.7)	51.5 (33.5 - 69.2)	80.8 (74.4 - 85.8)	43.6 (32.7 - 54.8)	4
A_B	70.4 (61.2 - 78.6)	63.9 (46.2 - 79.2)	84.1 (77.1 - 89.2)	44.2 (35.5 - 53.7)	3
deA_UB	83.1 (75 - 89.3)	36.1 (20.8 - 53.8)	77.9 (73.1 –82.0)	44.1 (30.4 - 58.7)	0
deA_B	81.9 (73.7 - 88.4)	36.1 (20.8 - 53.8)	77.6 (72.8 - 81.8)	42.5 (29.2 - 56.9)	2
LO_UB	78.9 (70.3 – 86.0)	76.5 (58.8 - 89.3)	90.1 (83.1 - 94.4)	57.3 (47.4 - 66.7)	6
LO_B	87.8 (80.4 - 93.2)	12.9 (3.63 - 29.8)	73.2 (70.1 –76.0)	28.2 (12.2 - 52.5)	8

Object Detection Model Performance

Annotated images from the training dataset were uploaded into the Google Cloud platform and the object detection model trained as described above. The internal performance report during training is provided in Table 4.

Table 4. Internal performance report from the object detection model during training. AUC: Area under the precision recall curve. PPV: Positive predictive Value. NPV: Negative predictive value. 95% CI: 95% Confidence Interval.

Score Threshold	AUC	Sensitivity(%) 95% CI	Specificity(%) 95% CI	PPV(%) 95% CI	NPV(%) 95% CI
0.47	0.667	75.0 (42.8 - 94.5)	80.0 (51.9 - 95.7)	75.0 (50.8-90.0)	80.0 (59.3-91.7)

Following training, the 154 prediction images were uploaded into the model and the predictions showed three distinct behaviors. In the first behavior, the model detected the lesions as well as the area where the lesion was located using the bounding boxes and provided confidence scores (Figure 5a, 5b). In the second behavior, the model detected no distinct lesion but predicted either benign or malignant areas within the image (Figure 5c). In the third behavior, the model detected lesions but assigned both malignant and benign labels to the lesions with different confidence scores (Figure 5d). The performance metrics of the object detection model within the independent prediction dataset is provided in Table 5.

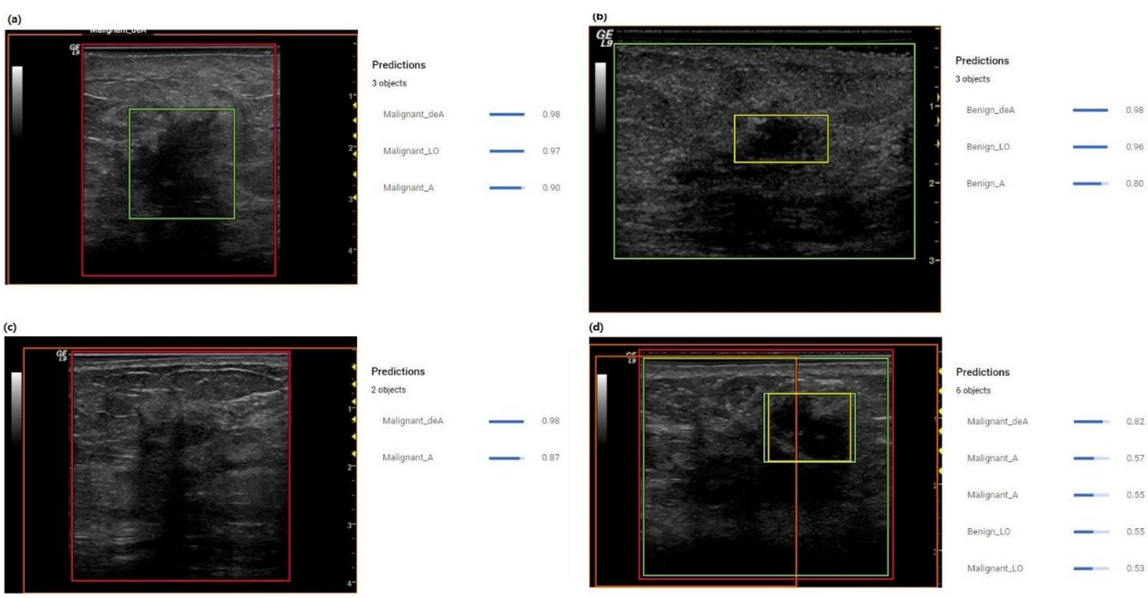


Figure 5. (a) Example case where the model detected both lesion and suspicious areas in the image with confidence scores of 0.97, 0.98 and 0.9. The position of the malignant lesion was marked by the green color bounding box drawn by the model. (b) Example case where the model detected both lesion and suspicious areas in the image with confidence scores of 0.96, 0.98 and 0.8 for the lesion and areas to be benign. The position of the benign lesion was marked by the yellow bounding box drawn by the model. (c) Example case where the model detected no lesions but malignant areas with confidence scores of 0.98 and 0.87. (d) The model detected the lesion but assigned both malignant

and benign labels. The model provided a confidence score of 0.55 for the lesion to be benign and a confidence score of 0.53 for the lesion to be malignant. The model also indicated malignant areas with confidence score of 0.82 and 0.57.

Table 5. The calculated sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for the object detection model in the prediction (post-training) dataset. 95% CI: 95% Confidence Interval.

Score Threshold	Sensitivity(%) 95% CI	Specificity(%) 95% CI	PPV(%) 95% CI	NPV(%) 95% CI	# of N/A
0.72	78.8 (70.3 - 85.8)	69.4 (51.9 - 83.7)	87.5 (80.9-92.0)	54.8 (44.6 - 64.6)	0

Rate of Prediction Agreement

The presence of multiple images and predictions (5-7) from each independent case (n=26) allowed for quantification of intra-reader agreement of each model. This data is summarized in Table 6. All models demonstrated a reasonably high rate of agreement, with no statistical difference observed across models (p=0.8).

Table 6. Average percentage of model prediction agreement with standard deviation across the 26 cases for all models.

Models	Prediction Agreement
OBJ	88 ± 18.2%
A_B	82 ± 18.1%
A_UB	87 ± 16.7%
deA_B	88 ± 13%
deA_UB	90 ± 13%
LO_B	86 ± 22%
LO_UB	89 ± 16.5%

Discussion

Ultrasound is a nonionizing, readily available, low-cost, and real-time imaging modality that has shown good diagnostic performance in breast cancer detection and diagnosis. In recent years, radiologists have explored the potential of AI technology to improve clinical practice, including the accuracy of ultrasound for breast cancer diagnosis.[9,10,11] Google AutoML Vision, released in 2018, may aid in the characterization of indeterminate breast masses by building of customized image-classification and image-recognition models on cloud services. Thus, this study explored the potential of AutoML Vision to classify and evaluate breast ultrasound images, using its image classification and object detection model.

Within the image classification model, 6 different training data setups were investigated. Performance during internal testing from these methods was similar with areas under the precision recall curve ranging from 0.85 to 0.96, indicating the influence of label balancing and image cropping were negligible in this dataset. The object detection model had an area under the precision recall curve of 0.67 during internal validation. While this performance is less encouraging than the classification model, the object detection could locate the position of lesion in the image. It is anticipated that this will enable radiologist adoption by providing a clear rationale for diagnosis while also streamlining workflow.

Comparing the performance of LO_UB with prior studies on classifying B-mode ultrasound breast mass using deep learning algorithms, the 91.1% AUC was similar to the 89.6% AUC from Cheng et al.[18] and 93.6% from Byra et al.[10] but lower than the 96% from Han et al.[18]or the 99% reported by Yap et al.[19] Importantly however, studies that have reported exceptional overall AUCs have employed datasets consisting of large numbers of lesions that were clearly benign (BI-RADS < 3) or highly likely to be malignant (BI-RADS 5)[19,20]. Data from our study primarily consisted of indeterminate breast masses scheduled for biopsy in which lower performance is expected, but this scenario more closely resembles the clinical need for improved diagnosis. Therefore, we believe the image classification model provides acceptable diagnostic performance under the appropriate training setups.

While encouraging, several limitations exist and should be addressed in the future. Within the object detection model, the input regions of interest are required to be in rectangular shape. The result of this is that all LO images will contain surrounding tissue. Based on the size and shape of the lesion, the amount of surrounding tissues could vary, which may introduce unwanted variability. Thus, potential improvement maybe achieved by allowing customize-shaped input images for the model or automatic segmentation prior to image upload. Meanwhile, more training images could be added to increase the model performance as only 264 training images were used in study. Finally, while the AutoML program stresses ease of use and off-the shelf capabilities, its limited flexibility also results in limitations compared to traditional AI platforms [21, 22]. For example, traditional methods of sample size augmentation and testing such as leave-one-out cross-validation methods cannot be used in applications where multiple images/lesion are generated without compromising independence. Additionally, once the model is deployed it provides a binary decision on images used for prediction, which prohibits traditional performance evaluations such as areas under the ROC and precision-recall curves. Despite these limitations, results to date are encouraging and the platform should be further explored moving forward.

Conclusion

The Google AutoML Vision platform showed an acceptable performance to classify breast ultrasound images under appropriate training setups and the use of both the Image Classification and Object Detection Models should be further explored. The platform also showed cost-effective advantage as all customized models were run on cloud services minimizing local hardware requirements. Our results indicated the platform could potentially be a useful tool in assisting radiologists in the characterization of indeterminate breast masses identified during screening. Ultimately, this approach could reduce the number of unnecessary biopsies.

Conflicts of Interest

For the original clinical trial that data was obtained from, the ultrasound contrast agent was provided by Lantheus Medical Imaging and the ultrasound scanner provided by GE Healthcare. No other conflicts of interest are declared.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin. 2019;69(1):7-34.

2. Koual M, Cano-Sancho G, Bats AS, Tomkiewicz C, Kaddouch-Amar Y, Douay-Hauser N, et al. Associations between persistent organic pollutants and risk of breast cancer metastasis. Environ Int. 2019; 132:105028

3. Etzioni R, Urban N, Ramsey S, et al. The case for early detection. Nat Rev Cancer 2003; 3:243–252

4. Brem RF, Lenihan MJ, Lieberman J, Torrente J. Screening breast ultrasound: past, present, and future. AJR Am J Roentgenol. 2015;204(2):234-40.

5. Guo R, Lu G, Qin B, Fei B. Ultrasound Imaging Technologies for Breast Cancer Detection and Management: A Review. Ultrasound Med Biol. 2018;44(1):37-70.

6. Stavros T. Breast Ultrasound. Philadelphia, USA: LIPPINCOTT WILLIAMS & WILKINS; 2004.

7. Kolb TM, Lichy J, Newhouse JH. Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations. Radiology 2002;225(1):165–175

8. O’Connor M. Radiology spending on AI expected to surpass $2B by 2023. Health Imaging. 2018.

9. Chang RF, Kuo WJ, Chen DR, Huang YL, Lee JH, Chou YH. Computer-Aided Diagnosis for Surgical Office-Based Breast Ultrasound. Arch Surg. 2000; 135:696-699

10. Byra M, Galperin M, Ojeda-Fournier H, Olson L, O'Boyle M, Comstock C, et al. Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. Med Phys. 2019;46(2):746-55.

11. Wu GG, Zhou LQ, Xu JW, Wang JY, Wei Q, Deng YB, et al. Artificial intelligence in breast ultrasound. World J Radiol. 2019;11(2):19-26.

12. Li FF, Li J. Cloud AutoML: Making AI accessible to every business. Google Cloud. 2018

13. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986; 21(9):720–733

14. A. Sridharan, J. R. Eisenbrey, P. Machado, H. Ojeda-Fournier, A. Wilkes, A. Sevrukov, R. F. Mattrey, K. Wallace, C. L. Chalek, K. E. Thomenius, F. Forsberg.

Quantitative analysis of vascular heterogeneity in breast lesions using contrast-enhanced three-dimensional harmonic and subharmonic ultrasound imaging. IEEE Trans Ultrason Ferroelectr Freq Control, vol. 62, no. 3, pp. 502 – 510, 2015.

15. A. Sridharan, J. R. Eisenbrey, M. Stanczak, P. Machado, D. A. Merton, A. Wilkes, A. Sevrukov, H. Ojeda-Fournier, R. F. Mattrey, K. Wallace, F. Forsberg. Characterizing breast lesions using quantitative parametric 3D subharmonic imaging: A multi-center study. Academic Radiology, In Press.

16. Nam K, Eisenbrey JR, Stanczak M, et al. Monitoring neoadjuvant chemotherapy for breast cancer by using three-dimensional subharmonic aided pressure estimation and imaging with US contrast agents: preliminary experience. Radiology 2017; 285: 53–62.

17. Seif G. Handling Imbalanced Datasets in Deep Learning. Medium. 2018

18. Cheng J-Z, Ni D, Chou Y-H, et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Sci Rep 2016;6:24454.

19. Han S, Kang H-K, Jeong J-Y, et al. A deep learning framework for supporting the classification of breast lesions in ultrasound images. Phys Med Biol 2017;62:7714-28.

20. Yap M.H., Pons G., Marti J., et al: Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J Biomed Health Inform 2018; 22: pp. 1218-1226.

21. Shuo W, Liu JB, Zhu Z, Eisenbrey J. Artificial Intelligence in Ultrasound Imaging: Current Research and Applications. Advanced Ultrasound in Diagnosis and Therapy 2019;03:053–061.

22. Handelman GS, Kok HK, Chandra RV, et al. Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. AJR Am J Roentgenol. 2019; 212(1):38-43.