

10-23-2020

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

Shuo Wang

*Thomas Jefferson University; Drexel University*

Sihua Niu

*Peking University People's Hospital*

Enze Qu

*The Third Affiliated Hospital of Sun Yat-Sen University*

Flemming Forsberg

*Thomas Jefferson University*

Annina Wilkes

*Thomas Jefferson University*

Follow this and additional works at: <https://jdc.jefferson.edu/radiologyfp>



Part of the [Radiology Commons](#)

*See next page for additional authors*

**[Let us know how access to this document benefits you](#)**

---

### Recommended Citation

Wang, Shuo; Niu, Sihua; Qu, Enze; Forsberg, Flemming; Wilkes, Annina; Sevrakov, Alexander; Nam, Kibo; Mattrey, Robert F.; Ojeda-Fournier, Haydee; and Eisenbrey, John R., "Characterization of indeterminate breast lesions on B-mode ultrasound using automated machine learning models" (2020). *Department of Radiology Faculty Papers*. Paper 94.

<https://jdc.jefferson.edu/radiologyfp/94>

This Article is brought to you for free and open access by the Jefferson Digital Commons. The Jefferson Digital Commons is a service of Thomas Jefferson University's [Center for Teaching and Learning \(CTL\)](#). The Commons is a showcase for Jefferson books and journals, peer-reviewed scholarly publications, unique historical collections from the University archives, and teaching tools. The Jefferson Digital Commons allows researchers and interested readers anywhere in the world to learn about and keep up to date with Jefferson scholarship. This article has been accepted for inclusion in Department of Radiology Faculty Papers by an authorized administrator of the Jefferson Digital Commons. For more information, please contact: [JeffersonDigitalCommons@jefferson.edu](mailto:JeffersonDigitalCommons@jefferson.edu).

---

**Authors**

Shuo Wang, Sihua Niu, Enze Qu, Flemming Forsberg, Annina Wilkes, Alexander Sevrakov, 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<sup>1,2</sup>, Sihua Niu, MD, PhD<sup>3</sup>, Enze Qu, MD<sup>4</sup>, Flemming Forsberg, PhD<sup>2</sup>,  
Annina Wilkes, MD<sup>2</sup>, Alexander Sevrakov, MD<sup>2</sup>, Kibo Nam, PhD<sup>2</sup>, Robert F. Mattrey  
MD<sup>5</sup>, Haydee Ojeda-Fournier, MD<sup>6</sup>, John R. Eisenbrey, PhD<sup>2</sup>

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. 10<sup>th</sup> Street

Philadelphia, PA 19017, USA

e-mail: [John.Eisenbrey@jefferson.edu](mailto: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

1 **Abstract**

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

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

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

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

26

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

29

30 **Introduction**

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

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

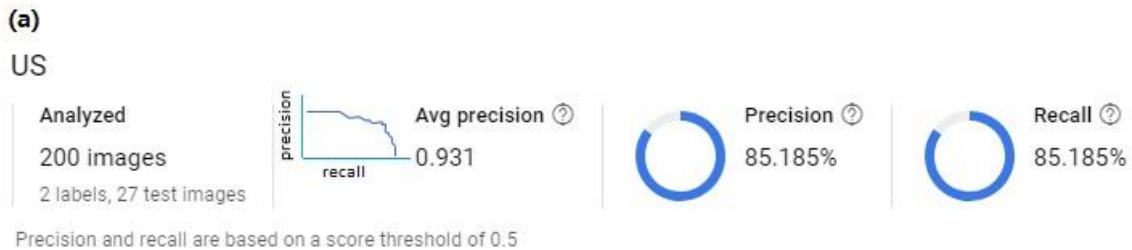
47 The use of artificial intelligence (AI) in radiology has the potential to reduce costs,  
48 save time, and improve diagnostic accuracy.<sup>8</sup> Multiple studies have shown that deep  
49 learning algorithms (one type of AI) outperform experienced radiologists in the diagnosis  
50 of breast lesions with 5-13% larger area under the receiver operating characteristic (ROC)  
51 curves.<sup>9,10,11</sup> However, using deep learning algorithms requires a large amount of data (e.g.,  
52 5,000-10,000 training images) and training a new deep learning algorithm is both time-  
53 consuming and expensive. Several commercially AI programs are available providing an  
54 opportunity to overcome these barriers. Google AutoML Vision (Google, Mountain View,  
55 CA) is a machine learning service from Google Cloud Platform that runs deep learning  
56 algorithms online and performs image-classification and image-recognition tasks on cloud  
57 services, reducing the need for expensive hardware. It enables a customized model to be  
58 created quickly by leveraging transfer learning and neural architecture search technologies,  
59 which can lead to more accurate results with less misclassifications than other generic  
60 machine learning services.<sup>12,13</sup> In addition, due to the transfer learning component, which  
61 takes the advantages of lower-level features from pre-trained convolutional neural  
62 networks (CNN), significantly fewer images are required for algorithm training.<sup>11</sup>

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

71 While this technology has been used for a variety of product management  
72 applications, its use in radiological applications is relatively unexplored.<sup>12,13</sup> Thus, the  
73 purpose of this study was to evaluate the performance of both AutoML Vision's image  
74 classification and object detection models for the characterization of intermediate breast  
75 masses imaged with B-mode ultrasound. Specifically, we strove to identify the

76 performance of AutoML’s image classification and object detection mass for classifying  
 77 breast masses as cancerous or non-cancerous in a population of suspicious masses  
 78 scheduled for tissue biopsy. The influence of category balancing and image cropping on  
 79 model performance was also investigated.

80  
 81



(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

82  
 83 **Figure 1.** (a) A model performance report is generated after each training process (b)  
 84 Parameter descriptions and their equivalent ROC terminologies.

85

## 86 **Material and Methods**

### 87 *Clinical studies*

88 To create training datasets for the AI image classification and object detection  
 89 models, ultrasound images were extracted from two previous clinical studies. The first  
 90 study was a multi-center clinical trial that was approved by the Institutional Review Boards  
 91 of Thomas Jefferson University (TJU) and The University of California, San Diego (UCSD)  
 92 and conducted between January 2011 and December 2015 in which contrast-enhanced  
 93 ultrasound was used to characterize indeterminate breast masses scheduled for biopsy.<sup>14,15</sup>  
 94 The second study was approved by the Institutional Review Boards of TJU and conducted  
 95 between May 2014 and February 2016, in which a contrast-enhanced ultrasound technique  
 96 was used to predict the response of breast cancer to neoadjuvant chemotherapy<sup>16</sup>. All  
 97 patients from both studies provided written informed consent before participating. The  
 98 imaging data for both studies were acquired using a commercially available Logiq 9  
 99 scanner (GE Healthcare, Waukesha, WI) equipped with a 4D10L probe and imaging  
 100 parameters were optimized on an individual basis according to good clinical practice. There  
 101 were 236 women enrolled in the first clinical study with an average age of  $52 \pm 13$  years.  
 102 The average lesion cross-sectional areas for malignant and benign lesions were  $190.1 \pm$   
 103  $35.7 \text{ mm}^2$  and  $124.1 \pm 15.5 \text{ mm}^2$ , respectively. The second clinical study enrolled 17  
 104 participants who had invasive ductal carcinomas with an average age of  $52.9 \pm 10.4$  years  
 105 and an average lesion cross-sectional area of  $604.6 \pm 460.7 \text{ mm}^2$ . In total, there were 253  
 106 cases. For this AI processing study, 242 patient cases with available biopsy results  
 107 (reference standard) were selected. Within these 242 cases, 21 cases were then excluded

108 by a blinded radiologist due to poor image quality resulting in 154 unique patients with  
109 benign breast lesions and 67 unique patients with malignant breast lesions (221 in total).

110

### 111 *Data preprocessing*

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

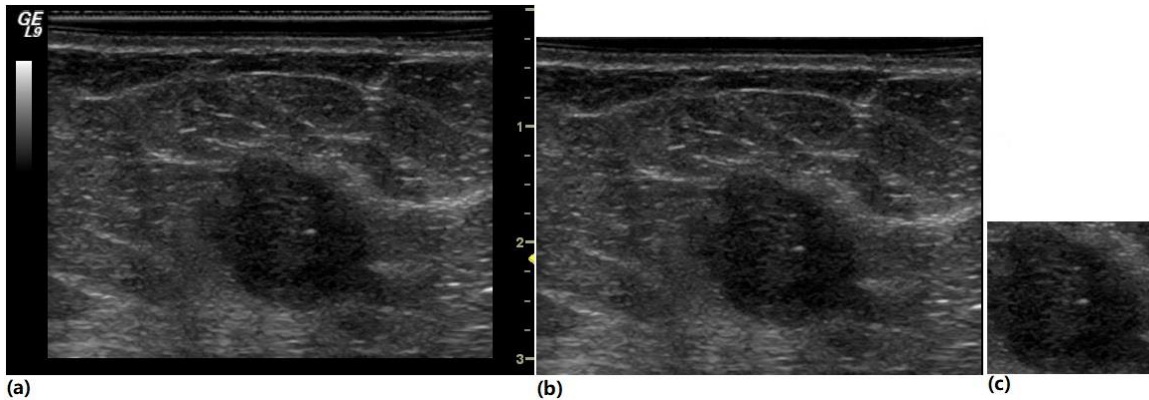
123 Based on model recommendations, 26 out of the 221 cases (19 malignant and 7  
124 benign cases corresponding to 11% of the patients) were reserved to form an independent  
125 prediction dataset to evaluate the models' performance. In order to augment our prediction  
126 dataset, a second radiologist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

132

### 133 *Image Classification Model Training*

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

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



151

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

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

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

160

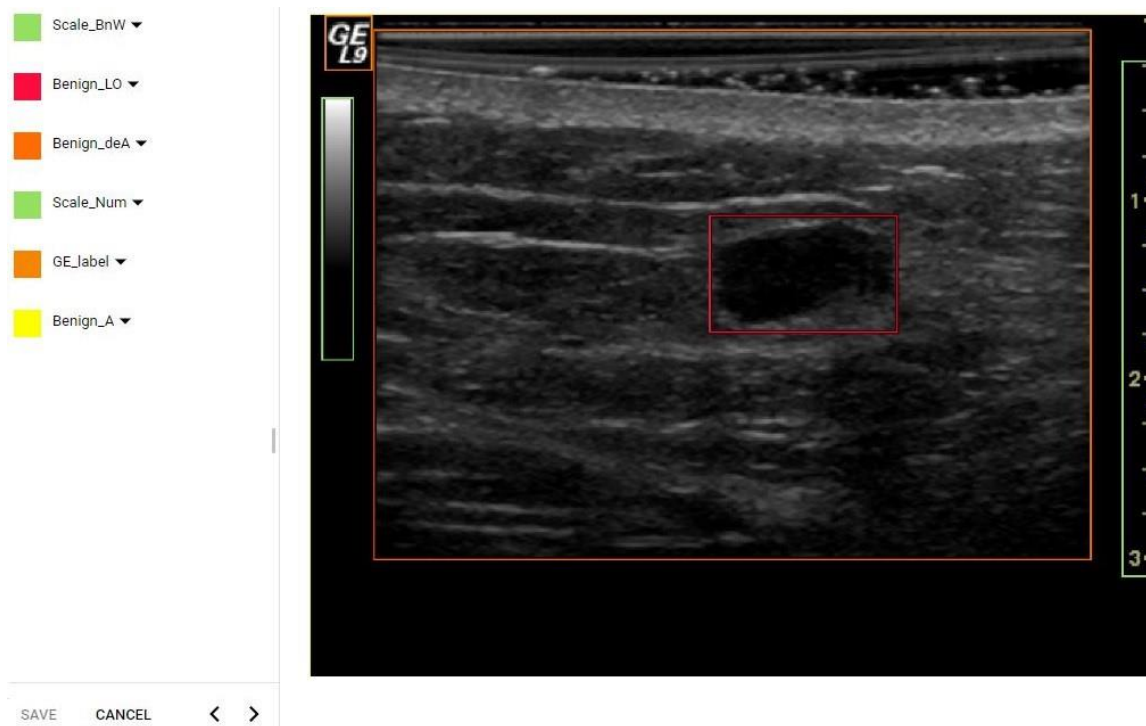
161 *Object Detection Model Training*

162 The Google AutoML Vision Object Detection Model was investigated to determine  
 163 the ability of this algorithm to first identify the suspicious breast mass, then subsequently  
 164 assign a risk score on the likelihood of the image containing breast cancer. To train the



165 object detection model, the same training data (147 benign and 117 malignant breast lesion  
166 images) as well as the same prediction images (154 breast images) described above were  
167 utilized. Data was first uploaded into Google Cloud Storage and then an Excel file that  
168 contained pathways for importing each image was generated from Python. The object  
169 detection model processes training image data within the model by using bounding boxes  
170 and labels to select objects that were important and intended to be detected inside an image.  
171 Therefore, only the full annotated images were imported into the model. Following upload,  
172 the model was trained by a blinded radiologist to identify the scale bars and manufacturer  
173 labels (as an algorithm validation check) and either malignant or benign masses within the  
174 three cropping approaches described above. An example of this training is provided in  
175 Figure 3.

176



177

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

### 183 *Evaluation of Model Performance*

184 The performance of each model was evaluated using results from the participant's  
185 tissue biopsy as a reference standard. Performance reporting was separated by internal  
186 performance (self-reported by the model during training) and external prediction within the  
187 dataset reserved for testing. For internal validation, the area under the precision recall curve,  
188 sensitivity, specificity, negative predictive value, and positive predictive value were all  
189 reported with 95% confidence intervals. Model agreement was calculated for each of the  
190 six image classification models and the object detection model by quantifying the rate of

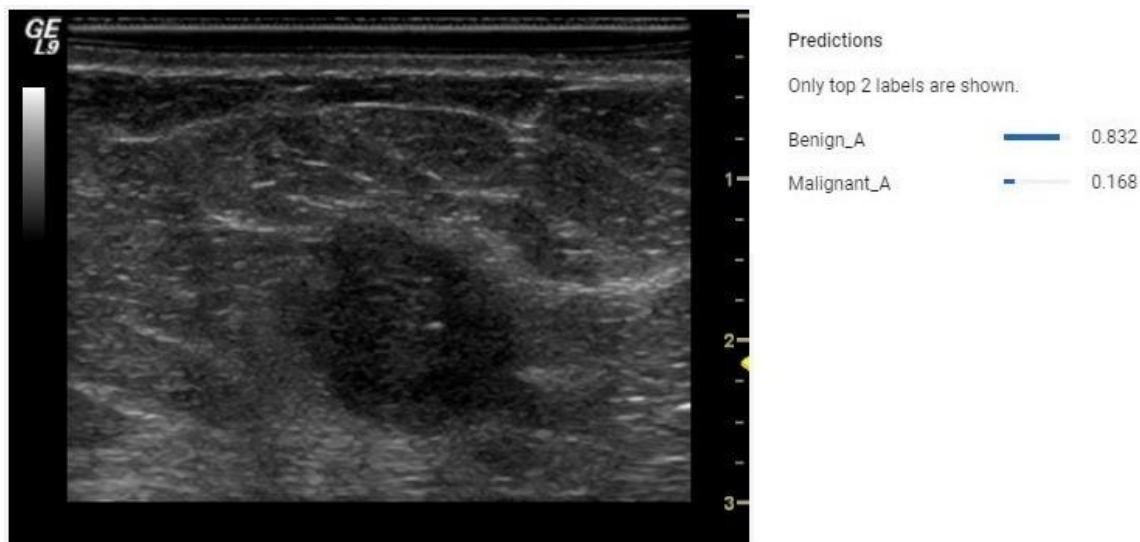
191 agreement amongst images taken from the same lesion for each of the 26 external  
192 prediction cases. All statistical analysis was performed in GraphPad Prism Version 8.0  
193 (San Diego, CA) with comparisons across multiple groups performed using a one-way  
194 ANOVA and direct comparisons between individual groups determined using a Student's  
195 t-test. Statistical significance was determined using  $p < 0.05$ .

196

## 197 **Results**

### 198 *Image Classification Model Performance*

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



215

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

219

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

222 Positive predictive value. NPV: Negative predictive value. 95% CI: 95% Confidence  
 223 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)

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

225

226 **Table 3.** The calculated sensitivity, specificity, positive predictive value (PPV), and  
 227 negative predictive value (NPV), for all customized image classification models as well  
 228 as number of N/A cases in the prediction (post-training) dataset. 95% CI: 95%  
 229 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

230

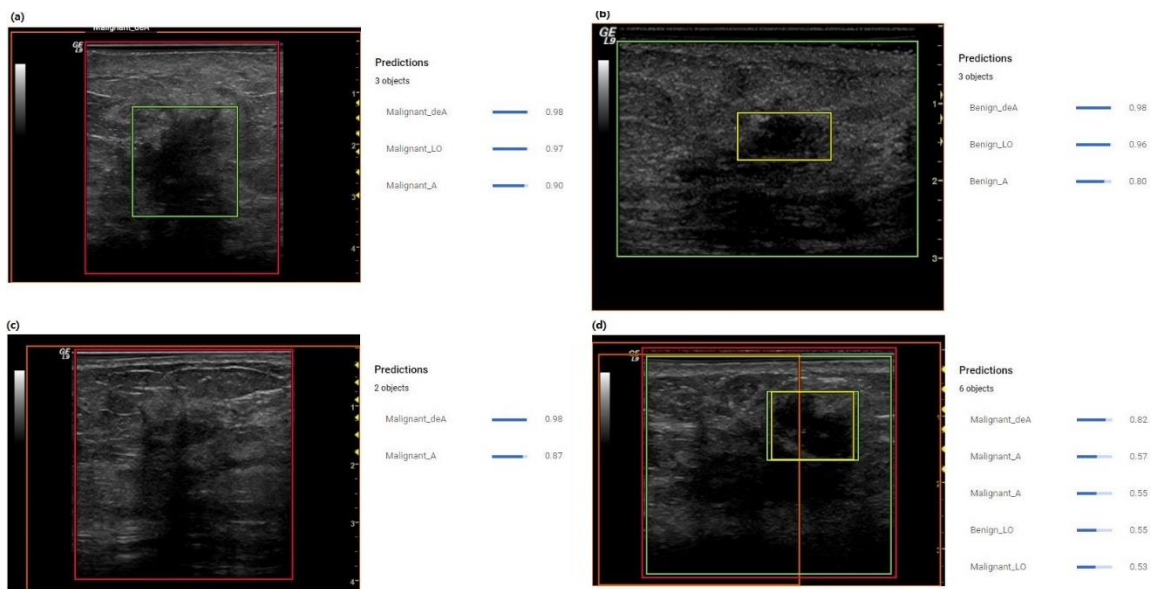
231 *Object Detection Model Performance*

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

235 **Table 4.** Internal performance report from the object detection model during training.  
 236 AUC: Area under the precision recall curve. PPV: Positive predictive Value. NPV:  
 237 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)

238  
 239 Following training, the 154 prediction images were uploaded into the model and  
 240 the predictions showed three distinct behaviors. In the first behavior, the model detected  
 241 the lesions as well as the area where the lesion was located using the bounding boxes and  
 242 provided confidence scores (Figure 5a, 5b). In the second behavior, the model detected no  
 243 distinct lesion but predicted either benign or malignant areas within the image (Figure 5c).  
 244 In the third behavior, the model detected lesions but assigned both malignant and benign  
 245 labels to the lesions with different confidence scores (Figure 5d). The performance metrics  
 246 of the object detection model within the independent prediction dataset is provided in Table  
 247 5.  
 248



249  
 250 **Figure 5.** (a) Example case where the model detected both lesion and suspicious areas in  
 251 the image with confidence scores of 0.97, 0.98 and 0.9. The position of the malignant  
 252 lesion was marked by the green color bounding box drawn by the model. (b) Example  
 253 case where the model detected both lesion and suspicious areas in the image with  
 254 confidence scores of 0.96, 0.98 and 0.8 for the lesion and areas to be benign. The position  
 255 of the benign lesion was marked by the yellow bounding box drawn by the model. (c)  
 256 Example case where the model detected no lesions but malignant areas with confidence  
 257 scores of 0.98 and 0.87. (d) The model detected the lesion but assigned both malignant

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

261

262 **Table 5.** The calculated sensitivity, specificity, positive predictive value (PPV), and  
 263 negative predictive value (NPV) for the object detection model in the prediction (post-  
 264 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

265

266 *Rate of Prediction Agreement*

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

271

272 **Table 6.** Average percentage of model prediction agreement with standard deviation across  
 273 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%

274

275 **Discussion**

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

285 Within the image classification model, 6 different training data setups were  
 286 investigated. Performance during internal testing from these methods was similar with  
 287 areas under the precision recall curve ranging from 0.85 to 0.96, indicating the influence

288 of label balancing and image cropping were negligible in this dataset. The object detection  
289 model had an area under the precision recall curve of 0.67 during internal validation. While  
290 this performance is less encouraging than the classification model, the object detection  
291 could locate the position of lesion in the image. It is anticipated that this will enable  
292 radiologist adoption by providing a clear rationale for diagnosis while also streamlining  
293 workflow.

294 Comparing the performance of LO\_UB with prior studies on classifying B-mode  
295 ultrasound breast mass using deep learning algorithms, the 91.1% AUC was similar to the  
296 89.6% AUC from Cheng et al.<sup>18</sup> and 93.6% from Byra et al.<sup>10</sup> but lower than the 96% from  
297 Han et al.<sup>18</sup> or the 99% reported by Yap et al.<sup>19</sup> Importantly however, studies that have  
298 reported exceptional overall AUCs have employed datasets consisting of large numbers of  
299 lesions that were clearly benign (BI-RADS < 3) or highly likely to be malignant (BI-RADS  
300 5)<sup>19,20</sup>. Data from our study primarily consisted of indeterminate breast masses scheduled  
301 for biopsy in which lower performance is expected, but this scenario more closely  
302 resembles the clinical need for improved diagnosis. Therefore, we believe the image  
303 classification model provides acceptable diagnostic performance under the appropriate  
304 training setups.

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

322

### 323 **Conclusion**

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

332

### 333 **Conflicts of Interest**

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

337

### 338 **References**

- 339 1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin.*  
340 2019;69(1):7-34.
- 341 2. Koual M, Cano-Sancho G, Bats AS, Tomkiewicz C, Kaddouch-Amar Y, Douay-  
342 Hauser N, et al. Associations between persistent organic pollutants and risk of breast  
343 cancer metastasis. *Environ Int.* 2019; 132:105028
- 344 3. Etzioni R, Urban N, Ramsey S, et al. The case for early detection. *Nat Rev Cancer*  
345 2003; 3:243–252
- 346 4. Brem RF, Lenihan MJ, Lieberman J, Torrente J. Screening breast ultrasound: past,  
347 present, and future. *AJR Am J Roentgenol.* 2015;204(2):234-40.
- 348 5. Guo R, Lu G, Qin B, Fei B. Ultrasound Imaging Technologies for Breast Cancer  
349 Detection and Management: A Review. *Ultrasound Med Biol.* 2018;44(1):37-70.
- 350 6. Stavros T. *Breast Ultrasound.* Philadelphia, USA: LIPPINCOTT WILLIAMS &  
351 WILKINS; 2004.
- 352 7. Kolb TM, Lichy J, Newhouse JH. Comparison of the performance of screening  
353 mammography, physical examination, and breast US and evaluation of factors that  
354 influence them: an analysis of 27,825 patient evaluations. *Radiology* 2002;225(1):165–  
355 175
- 356 8. O'Connor M. Radiology spending on AI expected to surpass \$2B by 2023. *Health*  
357 *Imaging.* 2018.
- 358 9. Chang RF, Kuo WJ, Chen DR, Huang YL, Lee JH, Chou YH. Computer-Aided  
359 Diagnosis for Surgical Office-Based Breast Ultrasound. *Arch Surg.* 2000; 135:696-699
- 360 10. Byra M, Galperin M, Ojeda-Fournier H, Olson L, O'Boyle M, Comstock C, et al.  
361 Breast mass classification in sonography with transfer learning using a deep  
362 convolutional neural network and color conversion. *Med Phys.* 2019;46(2):746-55.
- 363 11. Wu GG, Zhou LQ, Xu JW, Wang JY, Wei Q, Deng YB, et al. Artificial intelligence  
364 in breast ultrasound. *World J Radiol.* 2019;11(2):19-26.
- 365 12. Li FF, Li J. *Cloud AutoML: Making AI accessible to every business.* Google Cloud.  
366 2018
- 367 13. Metz CE. ROC methodology in radiologic imaging. *Invest Radiol* 1986; 21(9):720–  
368 733
- 369 14. A. Sridharan, J. R. Eisenbrey, P. Machado, H. Ojeda-Fournier, A. Wilkes, A.  
370 Sevrakov, R. F. Mattrey, K. Wallace, C. L. Chalek, K. E. Thomenius, F. Forsberg.

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

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

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

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

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

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

387 20. Yap M.H., Pons G., Marti J., et al: Automated breast ultrasound lesions detection  
388 using convolutional neural networks. *IEEE J Biomed Health Inform* 2018; 22: pp. 1218-  
389 1226.

390 21. Shuo W, Liu JB, Zhu Z, Eisenbrey J. Artificial Intelligence in Ultrasound Imaging:  
391 Current Research and Applications. *Advanced Ultrasound in Diagnosis and Therapy*  
392 2019;03:053–061.

393 22. Handelman GS, Kok HK, Chandra RV, et al. Peering into the black box of artificial  
394 intelligence: evaluation metrics of machine learning methods. *AJR Am J Roentgenol*.  
395 2019; 212(1):38-43.

396

397

398