Original Article

Evaluation of the performance of traditional machine learning algorithms, convolutional neural network and AutoML Vision in ultrasound breast lesions classification: a comparative study

Ka Wing Wan, Chun Hoi Wong, Ho Fung Ip, Dejian Fan, Pak Leung Yuen, Hoi Ying Fong, Michael Ying

Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China

Correspondence to: Dr. Michael Ying. Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China. Email: michael.ying@polyu.edu.hk.

Background: In recent years, there was an increasing popularity in applying artificial intelligence in the medical field from computer-aided diagnosis (CAD) to patient prognosis prediction. Given the fact that not all healthcare professionals have the required expertise to develop a CAD system, the aim of this study was to investigate the feasibility of using AutoML Vision, a highly automatic machine learning model, for future clinical applications by comparing AutoML Vision with some commonly used CAD algorithms in the differentiation of benign and malignant breast lesions on ultrasound.

Methods: A total of 895 breast ultrasound images were obtained from the two online open-access ultrasound breast images datasets. Traditional machine learning models (comprising of seven commonly used CAD algorithms) with three content-based radiomic features (Hu Moments, Color Histogram, Haralick Texture) extracted, and a convolutional neural network (CNN) model were built using python language. AutoML Vision was trained in Google Cloud Platform. Sensitivity, specificity, F1 score and average precision (AUCPR) were used to evaluate the diagnostic performance of the models. Cochran's Q test was used to evaluate the statistical significance between all studied models and McNemar test was used as the post-hoc test to perform pairwise comparisons. The proposed AutoML model was also compared with the current related studies that involve similar medical imaging modalities in characterizing benign or malignant breast lesions.

Results: There was significant difference in the diagnostic performance among all studied traditional machine learning classifiers ($P<0.05$). Random Forest achieved the best performance in the differentiation of benign and malignant breast lesions (accuracy: 90%; sensitivity: 71%; specificity: 100%; F1 score: 0.83; AUCPR: 0.90) which was statistically comparable to the performance of CNN (accuracy: 91%; sensitivity: 82%; specificity: 96%; F1 score: 0.87; AUCPR: 0.88) and AutoML Vision (accuracy: 86%; sensitivity: 84%; specificity: 88%; F1 score: 0.83; AUCPR: 0.95) based on Cochran's Q test ($P>0.05$).

Conclusions: In this study, the performance of AutoML Vision was not significantly different from that of Random Forest (the best classifier among traditional machine learning models) and CNN. AutoML Vision showed relatively high accuracy and comparable to current commonly used classifiers which may prompt for future application in clinical practice.

Keywords: Ultrasonography; computer-aided diagnosis (CAD); machine learning; AutoML Vision; breast cancer

Submitted Jun 29, 2020. Accepted for publication Oct 30, 2020.
doi: 10.21037/qims-20-922
View this article at: http://dx.doi.org/10.21037/qims-20-922

Introduction

Breast cancer is one of the leading causes of death in women around the globe. There were more than 1.6 million new cases and about 1.2 million people died from breast cancer per year in China (1,2). Breast cancer is the most common cancer in the Western countries. In 2018, there were 523,000 new cases and more than 130,000 deaths in breast cancer. Early diagnosis remains as an important aspect in breast cancer because it allows patients to have early treatment and thus a better prognosis and higher survival rate. On the contrary, according to the World Health Organization (3), the global shortage in healthcare professionals are expected to reach 12.9 million by 2035, meaning that there might not have sufficient radiologists to examine the large number of medical images of breast cancer patients. This increases the workload of existing radiologists and might lead to a delayed treatment and poor prognosis of patients.

Artificial intelligence, consists of machine learning and brain-inspired deep learning neural networks (4), may help to tackle the current issues in the healthcare systems around the world. Recent applications of artificial intelligence in medical ultrasound images have already included a variety of specific tasks ranging from image segmentation to biometric measurement. For example, previous studies successfully developed methods for automated breast lesion segmentation and computer-aided quantification of intranodal vascularity on ultrasound (5,6). Artificial intelligence works well with radiomics which extracts image information that cannot be obtained by human like textural data and wavelet features. With the availability of those radiomics features, artificial intelligence system can be trained to make its own diagnosis such as classifying a tumor as benign or malignant. However, building and training a top-tier machine learning model require thorough understanding on the mathematical and engineering aspects of artificial intelligence, including tuning hyperparameters of the model and selecting appropriate algorithms. These might already be a laborious task for many experienced engineers or computer scientists, let alone some healthcare professionals with limited experience in computer science. In view of this, Google Cloud AutoML Vision might be a possible solution to combat this barrier because it is a highly user-friendly interface. AutoML Vision provides a highly automated model development environment to users who have less experience in computer programming to develop and train their own machine learning models according to their classification needs. AutoML Vision credited with the advantage of transfer learning in machine learning and their neural architecture search technologies.

Currently, AutoML Vision has been widely used for business purpose but seldom applied in the medical applications. However, previous studies have shown that AutoML Vision has potential significance in medical diagnostic field (7-9). The performance of these AutoML Vision-based computer-aided diagnosis (CAD) models was comparable to professional specialists indicating that these models could help clinical decision making. However, until now, there is no prior study has been conducted to investigate the feasibility of AutoML Vision in the analysis of ultrasound images in particular of breast ultrasound in differentiating benign and malignant lesions. As ultrasonography is relatively different when compared to other imaging modalities like X-ray, CT, or MRI in terms of operator dependency and image quality, testing AutoML Vision with ultrasound images become essential. This study presents a novel work in tackling this research gap by examining the potential usage of AutoML Vision in future clinical setting.

There is an increasingly popularity in using transfer learning techniques with some well-established pre-trained convolutional neural network in the realm of machine learning. Several studies have been conducted on tumor characterization and different methods have been proposed. We reviewed some of the transfer learning approaches with fine tuning techniques of pre-trained convolutional neural networks (CNN), including AlexNet, ResNet and Inception.

Ragab *et al.* (10) used some of the well-known CNN like Alex-Net with transfer learning fine tuning techniques and support vector machine to classify benign and malignant breast masses using the digital database for screening mammography (DDSM) and the Curated Breast Imaging Subset of DDSM (CBIS-DDSM) which yielded an accuracy of 80% and 87% respectively.

Xiao *et al.* (11) constructed other renowned CNN architectures ranging from ResNet to Inception to classify a total of 2,058 benign and malignant ultrasound breast lesions and achieved an accuracy of around 85%.

Byra *et al.* (12) utilized VGG19 pre-trained deep CNN and employed fine tuning techniques to classify 882 ultrasound breast images into benign or malignant. They obtained around 88% of accuracy in the classification.

The aim of the present study was to investigate the feasibility of AutoML Vision in future clinical applications

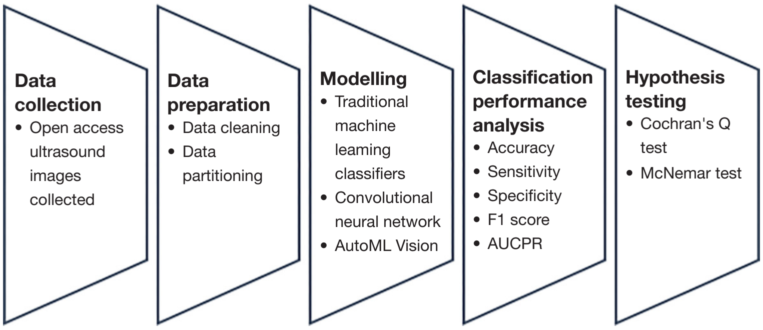


Figure 1 Flowchart of the design of the present study.

by testing the model with breast ultrasound images and compare AutoML Vision with some commonly used CAD algorithms in distinguishing benign and malignant breast lesions on ultrasound. Moreover, with reference to appropriate literature, this study would also compare the proposed AutoML model with some proven state-of-the-art transfer learning models such as ResNet and Inception to confirm the difference in performance between AutoML Vision and deep convolutional neural networks.

Methods

In this retrospective study, we proposed and compared different CAD models for ultrasound breast lesion classification by using common traditional machine learning classifiers, CNN, and Google AutoML Vision algorithms. *Figure 1* shows the overview of the methodology of the present study. This study was approved by the Human Subject Ethics Subcommittee of the authors' institution (Reference number: HSEARS20200311005).

Data source

Ultrasound images of benign and malignant breast tumors were collected from two online public datasets (13,14). These datasets have been widely used in other peer-reviewed literatures and proven to be effective in training machine learning models for detecting, classifying and segmenting the breast tumor (15,16). *Figure 2* show the samples images of the dataset. The details of the datasets were shown in *Table 1*.

Data preparation

As the ultrasound images were collected from two datasets, we firstly integrated the two sets of images into a single folder with according label as either "benign" or "malignant". Only ultrasound images of benign or malignant breast lesions were included in this study because we focused on the binary classification performance of different CAD models. Duplicated images were found from Cairo University Breast Ultrasound Images (BUSI) dataset probably due to inherent dataset error. The duplicated images (one image of a benign lesion and one image of a malignant lesion) were excluded from the study to prevent confusion during modelling of classifiers. Finally, a total of 895 ultrasonic breast ultrasound images were included in the study with 536 images of benign lesions and 359 images of malignant lesions.

The dataset of the 895 images was randomly segregated into two groups by using a self-developed python program. The first group contained 800 images (89.4%) and the second group contained 95 images (10.6%). The first group of the image dataset was used for the modelling and evaluation of the classification performance of different models. The second group of the image dataset was used for the final hypothesis testing among all studied models to evaluate if there is any significant difference in the performance of classification. In other words, the second group of the dataset was used to explore the statistical properties among various models. A summary of the distribution of the number of cases is shown in *Table 2*.

In the subsequent modelling steps, images in the first group (800 images) were randomly classified into three subgroups which are training (80%, 640 images), validation (10%, 80 images), and testing (10%, 80 images) respectively for all studied models, i.e., traditional machine learning classifiers and CNN. According to Google's documentation, AutoML Vision can automatically classify the data into the above-mentioned subgroups (17).

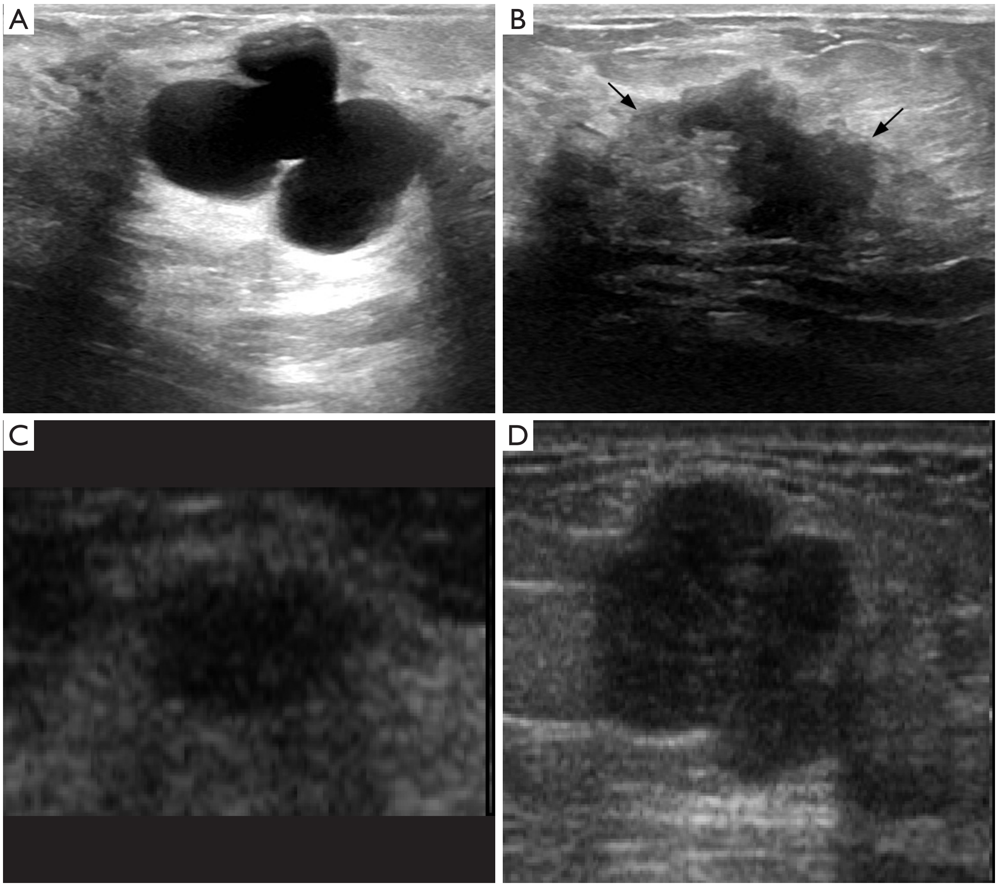


Figure 2 Sample ultrasound images of the dataset. (A) Grayscale ultrasound image shows a benign breast lesion which appears anechoic and has well-defined borders. The acoustic enhancement indicates that the lesion is cystic. [This image is obtained from a public domain (13)]. (B) Grayscale ultrasound image shows a malignant breast lesion which has ill-defined borders and heterogeneous echotexture (arrows). [This image is obtained from a public domain (13)]. (C) Grayscale ultrasound image shows a benign breast lesion which is hypoechoic and has well-defined borders. [This image is obtained from a public domain (14)]. (D) Grayscale ultrasound image shows a malignant breast lesion which is ill-defined and hypoechoic. The lesion appears heterogeneous with hypoechoic and hyperechoic areas. [This image is obtained from a public domain (14)].

Table 1 Details of the open-access data sources of ultrasound images

Source	Cairo University Breast Ultrasound Images (BUSI) dataset	Mendeley Data BUS dataset
Image distribution	780 images (133 normal, 437 benign and 210 malignant)	350 images (100 benign and 150 malignant)
Accessibility	https://scholar.cu.edu.eg/?q=afahmy/pages/dataset	http://dx.doi.org/10.17632/wmy84gzngw.1
Breast lesion classification	Normal, benign, and malignant	Benign and malignant
Related articles published	Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Deep learning approaches for data augmentation and classification of breast masses using ultrasound images (15).	Singh VK, Rashwan HA, Abdel-Nasser M, Sarker M, Kamal M, Akram F, Pandey N, Romani S, Puig D. An efficient solution for breast tumor segmentation and classification in ultrasound images using deep adversarial learning (16).

Table 2 Distribution of cases in the dataset used in this study

Type	1st group dataset (modelling), n (%)	2nd group dataset (hypothesis testing), n (%)	Entire database, n (%)
Benign lesion	480 (54)	56 (6)	536 (60)
Malignant lesion	320 (36)	39 (4)	359 (40)
Total	800 (89)	95 (11)	895 (100)

Modelling

In this study, three CAD systems were built: (I) traditional CAD system comprising the use of machine learning classifiers; (II) deep convolutional CAD system makes use of CNN as the building blocks; (III) AutoML Vision CAD system using Google Cloud.

The modelling for traditional and deep convolutional CAD models is performed on a computer workstation with Intel(R) HD Graphics 520 GPU, Intel(R) Core (™) i5-6,200 U CPUs and 8,192 MB RAM. Modelling environment for traditional and deep convolutional CAD models were launched on Jupyter Notebook (Version 6.0.3, New York, USA) using python codes. Several essential python libraries were imported for modelling which included sci-kit learn (18,19), Tensorflow (20), OpenCV (21), Mahotas, NumPy, h5Py, and Matplotlib. The modelling environment for AutoML Vision was carried out on the Google Cloud Platform.

Traditional machine learning classifiers

Several commonly used traditional machine learning algorithms for ultrasound CAD systems were identified from the literature (22) which includes Random Forest (RF), K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Logistic Regression (LR), Support Vector Machine (SVM), Naïve Bayes (NB), and AdaBoost. Feature extraction from the image is required for traditional models. However, feature descriptor in breast imaging reporting and data system (BI-RADS) handcrafted by radiologists such as depth-to-width ratio of lesion, echotexture and microcalcification characteristic in categorizing breast lesions incurs high cost and expertise. Therefore, this study extracted content-based radiomics image features using computer vision libraries ranging from OpenCV to Mahotas. Three common global image features descriptors (Haralick texture, Hu Moments, color histogram) that have been identified in a previous study (23), were extracted from all images and were used to model the traditional classifiers. Regarding Haralick texture, all images were converted to grayscale 8-bit and the fundamental theorem underlying was used to compute the Gray Level Co-occurrence Matrix (GLCM) and to calculate the 13 Haralick texture feature descriptors based on the GLCM, with the use of a set of formulae given by Haralick (24). Mahotas library would automatically compute these values. A total of 13 texture descriptors would then quantify the texture of the image, e.g., fine-coarse, rough-smooth, hard-soft. Regarding Hu Moments, they were calculated from central moments that could quantify the shape of the grayscale image independent of translation, scale, and rotation. Using OpenCV library, the analysis yielded a total of 7 moment descriptors (25). For color histogram, images were converted to hue, saturation, and value color space. OpenCV library automatically calculated the histogram for each image in terms of the image's hue, saturation, and quantitative value of the color features which yielded a total of 512 feature vectors. In general, a total of 532 feature vectors are generated from the above-mentioned feature descriptors. Once the above features for training images were extracted, built in machine learning classifiers from sci-kit learn libraries were used to train the traditional models (*Figure 3*). K-fold cross-validation (number of splits =10) was employed for traditional models. Built-in traditional machine learning classifiers from sci-kit learn packages were used in this study. The number of trees used in RF and AdaBoost were set to 100, the number of neighbors used in KNN was set to 5.

Convolutional neural network

Tensorflow was used to build the CNN architecture. As a relatively small data size was used in the present study, a shallow CNN architecture was built. The overall structure of the CNN includes two convolutional layers and two fully connected layers with sigmoid function for the final binary classification result (*Figure 4*). All convolutional layers make use of 3×3 kernels stacked together with Rectified Linear Units (ReLUs) as activation between each other

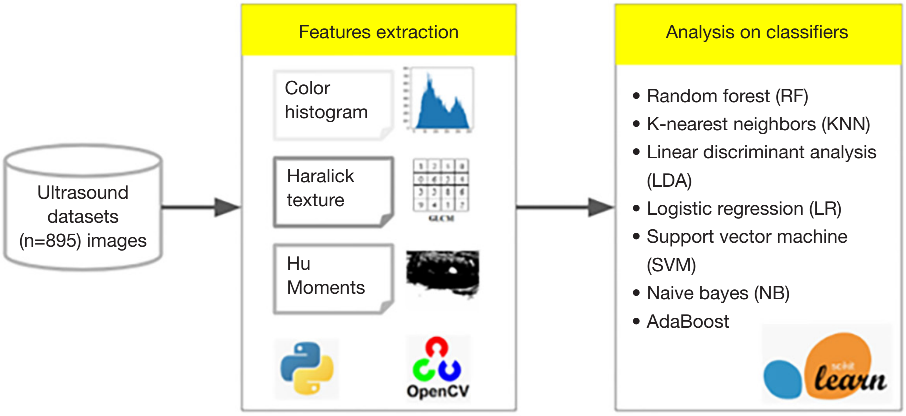

Ultrasound datasets (n=895) images
Features extraction
Color histogram
Haralick texture
GLCM
Hu Moments
OpenCV
Analysis on classifiers
• Random forest (RF)
• K-nearest neighbors (KNN)
• Linear discriminant analysis (LDA)
• Logistic regression (LR)
• Support vector machine (SVM)
• Naive bayes (NB)
• AdaBoost
scikit learn

Figure 3 Features extraction for traditional machine learning classifiers.

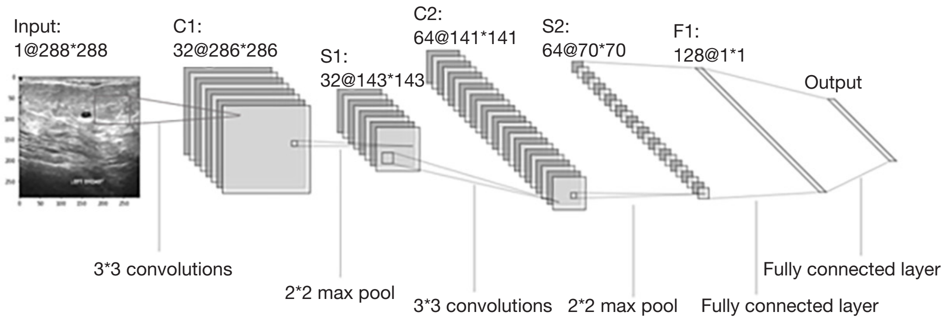


Figure 4 Convolutional Neural Network Architecture constructed in this study.

and followed by maximum pooling layers with 2×2 kernels. Hold-out validation (validation split =10%) was used to tune and refine the hyper-parameters of the model and check for any overfitting of the model. An early stop function was employed to determine the stopping epochs for CNN that is pivotal to prevent overfitting of the model. Images input to CNN were normalized with echogenicity (i.e., all pixel intensity values were normalized from 0 to 1).

AutoML Vision

For the modelling in AutoML Vision, the steps were simpler as Google provides a user-friendly user interface whereby the images were simply uploaded to a Google Cloud Platform bucket by a zip file in which AutoML Vision can recognize the label for each individual image. Images were distributed to the training, validation, and test datasets (80%, 10%, and 10%, respectively) automatically by Google Cloud (17). The training cost for this model is set to 16 node hours. The model type was set to "Cloud". AutoML Vision provides with its highly automated training process makes the modelling step much faster and easier.

Result analytic approach

To evaluate the classification performance of all classifiers, the accuracy, sensitivity, specificity and F1 score were calculated and compared whereby malignant lesions were considered to be "positive cases", while benign lesions were "negative cases". Additionally, average precision, also known as the area under the precision-recall curve, was used to evaluate the compare the performance of different CAD models.

For the hypothesis testing, the 95 images in the second group of the dataset were used to test any statistical significance among different classifiers. Cochran's Q was used to evaluate the significance of difference of diagnostic performance among different models, and McNemar

Table 3 Summary of classification performance of all studied classifiers

Model	Accuracy	Sensitivity	Specificity	F1 score	AUCPR
LR	0.74	0.64	0.79	0.63	0.75
LDA	0.76	0.64	0.83	0.65	0.66
KNN	0.84	0.75	0.88	0.76	0.81
RF	0.90	0.71	1.00	0.83	0.90
NB	0.35	1.00	0.00	0.52	0.35
SVM	0.73	0.57	0.81	0.59	0.67
AdaBoost	0.84	0.68	0.92	0.75	0.82
CNN	0.91	0.82	0.96	0.87	0.88
AutoML	0.86	0.84	0.88	0.83	0.95

LR, logistic regression; LDA, linear discriminant analysis; KNN, K-Nearest neighbors; RF, random forest; NB, naïve bayes; SVM, support vector machine; CNN, Convolutional Neural Network; AutoML, AutoML Vision; AUCPR, area under the precision-recall curve (Average Precision).

test was the post hoc test for pairwise comparison. All statistical analyses were performed using the Statistical Package for the Social Sciences (SPSS, version 23.0 for Windows, Chicago, IL, USA). A P value lesser than 0.05 was considered significant.

Results

Classification performance

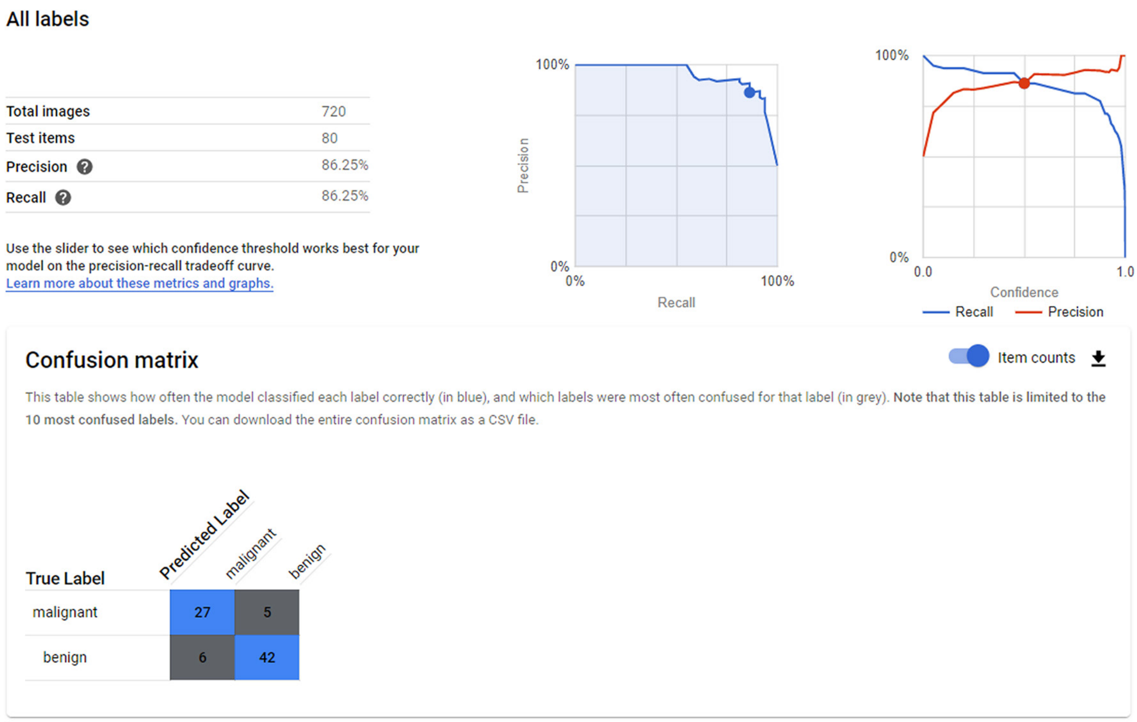
The diagnostic performance of different classifiers in the differentiation of benign and malignant breast lesions is summarized in *Table 3*. Among all studied traditional machine learning classifiers, Random Forest (RF) demonstrated the best performance, followed by K-Nearest Neighbors (KNN) and AdaBoost. These three traditional classifiers demonstrated relatively high classification performance (i.e., accuracy >80%) while others showed lower accuracy. Naïve Bayes (NB) showed lowest accuracy (35%) among all studied traditional machine learning classifiers with a sensitivity of 100% and a specificity of 0%. The CNN model built in the present study showed relatively high accuracy (91%). AutoML Vision also revealed high accuracy (86%) in differentiating benign and malignant breast lesions on ultrasound. AutoML Vision demonstrated highest average precision when compared with other classifiers (AUCPR: 0.95). *Figure 5* is a screenshot of the model evaluation setting of AutoML Vision where confusion matrix and precision-recall curve are shown.

Hypothesis testing

Results showed that there were significant differences among the classifiers in the performance of distinguishing benign and malignant breast lesions (Cochran's Q statistics: 291.28, P<0.05). Results of the pairwise post-hoc McNemar test demonstrated that four pairs of comparison exhibit did not show significant difference: AdaBoost *vs.* RF, AdaBoost *vs.* KNN, SVM *vs.* KNN, RF *vs.* KNN (P=0.82, 0.38, 0.14, 0.61 respectively), whereas other comparisons showed significant difference (P<0.05). When compared the best-performed traditional machine learning classifier (i.e., RF) with CNN and AutoML Vision, there was no significant differences in the diagnostic performance among these three models (Cochran's Q statistics: 3.06, P>0.05). *Table 4* shows the post hoc test result verifying there is no the significance difference for the classification accuracy of the RF, CNN, and AutoML Vision.

Discussion

In the present study, random forest (RF) demonstrated the best diagnostic performance among all studied traditional machine learning classifiers. RF is an ensemble learning method which aims to build multiple independent decision trees in calculation whereby improving the generalizability and robustness over a single decision tree. Each tree in the random forest can learn from each other and rectify the mistake. Another ensemble learning classifier Adaboost, which aims to eventually build a strong classification tree



All labels

Total images	720
Test items	80
Precision ?	86.25%
Recall ?	86.25%

Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve.
Learn more about these metrics and graphs.

Confusion matrix

This table shows how often the model classified each label correctly (in blue), and which labels were most often confused for that label (in grey). Note that this table is limited to the 10 most confused labels. You can download the entire confusion matrix as a CSV file.

True Label \ Predicted Label	malignant	benign
malignant	27	5
benign	6	42

Figure 5 Image shows the AutoML Vision model evaluation.

Table 4 Summary of the hypothesis testing between RF, CNN, and AutoML Vision

Data analysis information	CNN & RF	CNN & AutoML	RF & AutoML
Number of samples	95	95	95
P values	0.839	0.77	0.345

by training and refining the originally weak tree into strong tree sequentially, showed similar but relatively lower classification performance with RF. The unique property of multiple trees in RF makes it less susceptible to overfitting of the model such that it can improve the overall classification performance. Moreover, RF is good at handling large data sets with high dimensionality (i.e., feature vector size is 532 in the present study) and identifying important features while neglecting unimportant features in the computation (26). *Figure 6* shows the relative features importance of the top ten most important features. Interestingly, most of them (7 out of 10) are Haralick textural features, followed by Hu Moment features and lastly color histogram features. This might be related to the data used in this study whereby most ultrasound images are in grayscale and only some of them include color Doppler signal in aiding the diagnosis for malignancy, leading to relatively low importance of color histogram features. Moreover, the Hu Moment features that quantify the shape of the image have lower importance due to the operator dependency and image formation characteristics of ultrasound in which a slight change in transducer position can greatly alter the shape of the image or tumor. The reason for Haralick textural features have a greatest influence on classification might be attributed to the fact that it is pertained to the distribution pattern of nuclei inside the breast tissues and some studies also demonstrated that Haralick descriptor, which is correlated with the risk of breast cancer, has been deemed as a more important feature than histogram values in characterizing breast densities (27,28).

Both LR and LDA have been categorized as linear classifiers, which are widely used in medical imaging. However, the performance of these classifiers is limited by

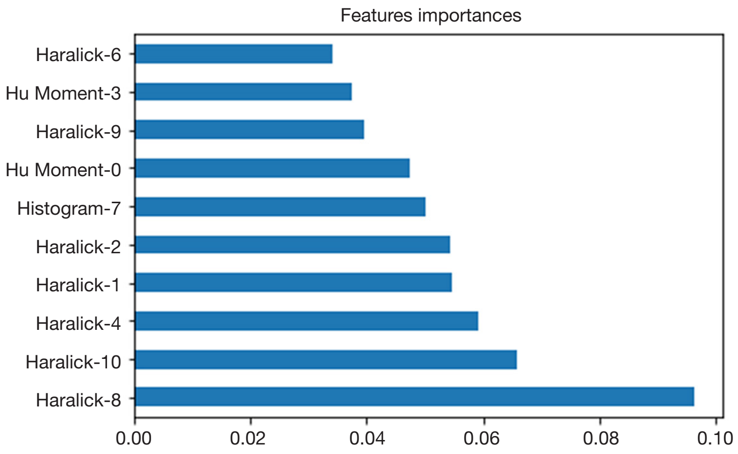


Figure 6 Bar chart shows the relative features' importance.

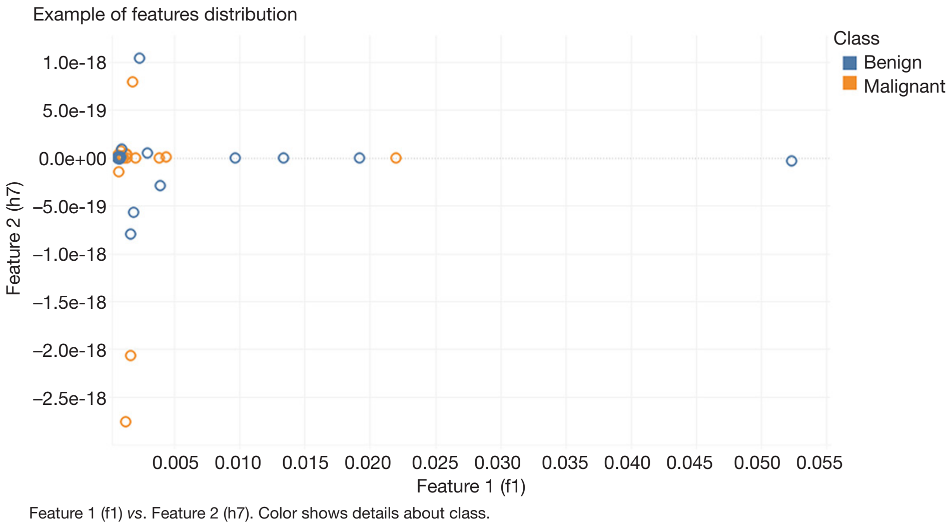


Figure 7 An example displays of the feature distribution of benign and malignant classes.

data features linearity, and the classification performance may be lowered if the data cannot be linearly separated (22). We speculate this may probably be the reason that average accuracy is attained by both linear classifiers. For getting a more in-depth understanding about features space distribution, we have randomly selected 40 images (20 benign and 20 malignant cases) for the visualization of two randomly selected features distribution in 2D space. *Figure 7* is the graph showing the distribution of breast tumor classes across two features (Haralick feature f1 and Hu Moment feature h7). From the graph, it is clear that the distribution is perhaps not linearly separable and thus it would be difficult for linear classifiers to draw a separation line to classify these two classes without huge error.

KNN, being one of the simplest classifiers and non-parametric machine learning algorithms aims to find a predefined number of neighbors closest in distance to the new data point and predict the class of the new data point. It could be one option to deal with dataset where the decision boundary is very irregular (18,19). The decision boundary in this study seems to be irregular and thus yielding a relatively accurate performance of KNN classifier (*Figure 5*).

SVM theorem aims to construct a hyperplane to divide the data into different categories in higher dimension which is basically suitable for study that involves a high dimension of feature vector space (22). Moreover, SVM is particularly useful when the dataset is non-linear separable (*Figure 5*). However, the performance of SVM in this study is neither the best nor the worst. We speculate that its relatively average performance could be attributed to the relation between the number of training samples per class and the number of features. As pointed out by previous study, the best practice to lower the error rate of the classifier is having a training sample that is at least three times of the number of features (29).

Among all traditional machine learning classifiers, Naïve Bayes Classifier (NB) demonstrated the lowest accuracy, specificity, F1 score and average precision. This probably can be explained by the underlying Bayes theorem which assumes all features are independent of each other or dependence can cancel out each other (30), which may not be applicable for the sonographic image data in the present study. Documentation from the sci-kit learn pointed out that NB is used when dealing with text data (19) and this classifier is more commonly used in email spam filtering rather than in image classification. The poor classification performance of NB in the present study highlighted the importance of feeding appropriate features into machine learning algorithms because the underlying mathematical principles of the classifiers could affect the outcomes.

From the viewpoint of CAD systems using traditional machine learning classifiers, the advantages include simpler architecture and flexibility in choosing multiple classifiers. On the contrary, the drawbacks of using traditional machine learning classifiers include the requirements of feature engineering and underperformance of the system in extremely large amount of data for analysis. Study has found that owing to the simple structure, the classification performance of traditional machine learning algorithms become steady when dealing with large amount of data whereas deep learning like CNN increases with increasing amount of large data (31).

CNN is considered as the gold standard in image classification as it consists of different layers like convolution and pooling, which can preserve the spatial relationship of the image. In addition, features like edges or blobs are automatically extracted within these layers and the neural network optimizes the classification performance by minimizing the loss of binary cross entropy (32). With these characteristics, CNN achieved high accuracy in differentiating benign and malignant breast lesions on ultrasound images in the present study. However, to build a CNN that can provide accurate results, many trial-and-error and experiences in determining the number of convolutional layers, choosing activation function and their combinations are needed. Therefore, considering the pros and cons of CNN, it is expected that more CAD systems will adopt the deep learning approach.

The present study found that the diagnostic accuracy of AutoML Vision was similar to that of CNN and RF in distinguishing benign and malignant breast lesions on ultrasound. Similar finding was also found in another paper which constitutes the use of AutoML Vision as an automated deep learning classifier and convolution neural networks in classifying invasive ductal carcinoma using histopathology images (33). Result from that study indicated comparable classification accuracy and F1 score for both AutoML Vision and CNN. The similar performance observed between AutoML Vision and CNN is probably due to the machine learning principle behind. Theoretically speaking, both classifiers utilize the deep learning technique where multilayer perceptron are stacked with convolution filters and pooling function. AutoML Vision can automatically find the best performing neural network architecture and hyperparameters with Google neural architecture search technology. Thus, with the same machine learning principle with CNN, AutoML Vision by Google demonstrated their strengths and uniqueness by developing a highly automated deep learning classifier with the added value of Google's neural architecture search technology, targeting different users to create their machine learning model in a more convenient manner. According to the available documentation about AutoML Vision, the main machine learning techniques is probably attributed to transfer learning and neural architecture search techniques by Google. Future research of utilizing both transfer learning and neural architecture search techniques to develop open sources automated diagnosis system is needed.

Although the classification performances and statistical testing in this study demonstrated no significant difference between AutoML Vision with the current CNN constructed in this study, further comparison with some state-of-the-art CNNs in similar classification tasks across different literatures is needed. As transfer learning is now becoming a more popular and robust option than merely convolutional neural network in classification task, this study has identified several related works in comparing their results with our proposed AutoML Vision model to further justify the statistical result obtained. In *Table 5*, some of the related works that involved the use of transfer learning techniques

Table 5 Comparison of the results of the present study with other related works involving the use of most proven, state-of-the-art models, and with mammography

Reference	Pre-trained network employed	Dataset	Data classes	Accuracy	AUC	AUCPR
Ragab *et al.* (10)	Fine-tuned AlexNet	Mammography DDSM (n=1,840)	Benign and malignant breast masses	0.81	0.88	–
		Mammography CBIS-DDSM (n=5,272)		0.87	0.94	–
Xiao *et al.* (11)	Fine-tuned ResNet50	Breast ultrasound images (n=2,058)	Benign and malignant breast masses	0.85	0.91	–
	Fine-tuned InceptionV3			0.85	0.91	–
Byra *et al*. (12)	Fine-tuned VGG19 + match layer techniques	Breast ultrasound images (n=882)	Benign and malignant breast masses	0.89	0.94	–
The proposed AutoML model	–	Breast ultrasound images (n=895)	Benign and malignant breast masses	0.86	–	0.95

and pre-trained well-established deep convolutional neural network were shown. The proposed AutoML model shows the highest AUCPR value (an alternative value of AUC) of 0.95 and the accuracy also shows a decent score of 0.85 which is comparable to other pre-trained CNNs. Therefore, we believe that there might not be much difference in the performance between CNN and AutoML Vision model.

One must also note that even though AutoML Vision can automatically produce a CAD model for users, detailed information about the model performance is limited in the version of AutoML Vision that we used in the present study.

There are limitations in this study. First, the images collected from the open-access databases were not in DICOM format and thus information such as patient demographic data was not available. Secondly, the sample size of the study was small, and we could not investigate the capability of different classifiers in handling of large amount of data. Thirdly, due to the limited resources, this study constructed relatively simple CAD models which do not include image preprocessing and image segmentation components which are considered as common practice in establishing complex CAD models (34). Lastly, this study compared the diagnostic performance of different classifiers, however, the relationship between the image data used in the study and the performance of different classification methods remains to be investigated.

Conclusions

In this study, we had built three CAD models (traditional machine learning models comprising of seven commonly used classifiers, CNN, and AutoML Vision) and investigated their performance in distinguishing benign and malignant breast lesions on ultrasound. Among all traditional machine learning models, RF demonstrated the best classification performance. AutoML Vision had a comparable classification performance with RF and CNN. AutoML Vision with its user-friendly interface has high potential for clinical practice aiding physicians in decision-making.

Acknowledgments

Funding: None.

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/qims-20-922). The authors have no conflicts of interest to declare.

Ethical Statement: This study was approved by the Human Subject Ethics Subcommittee of the authors' institution (Reference number: HSEARS20200311005).

References

1. Atlanta G. American Cancer Society. Cancer facts and figures 2013. American Cancer Society, 2013.
2. Fan L, Strasser-Weippl K, Li JJ, St Louis J, Finkelstein DM, Yu KD, Chen WQ, Shao ZM, Goss PE. Breast cancer in China. Lancet Oncol 2014;15:e279-89.
3. Gulland A. Shortage of health workers is set to double, says WHO. British Medical Journal Publishing Group, 2013.
4. Pakdemirli E. Artificial intelligence in radiology: friend or foe? Where are we now and where are we heading? Acta Radiologica Open 2019;8:2058460119830222.
5. Cheng SC, Ahuja AT, Ying M. Quantification of intranodal vascularity by computer pixel-counting method enhances the accuracy of ultrasound in distinguishing metastatic and tuberculous cervical lymph nodes. Quant Imaging Med Surg 2019;9:1773.
6. Lee CY, Chang TF, Chou YH, Yang KC. Fully automated lesion segmentation and visualization in automated whole breast ultrasound (ABUS) images. Quant Imaging Med Surg 2020;10:568.
7. Faes L, Wagner SK, Fu DJ, Liu X, Korot E, Ledsam JR, Back T, Chopra R, Pontikos N, Kern C. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. Lancet Digital Health 2019;1:e232-42.
8. Livingstone D, Chau J. Otoscopic diagnosis using computer vision: An automated machine learning approach. Laryngoscope 2020;130:1408-13.
9. Yang J, Zhang C, Wang E, Chen Y, Yu W. Utility of a public-available artificial intelligence in diagnosis of polypoidal choroidal vasculopathy. Graefes Arch Clin Exp Ophthalmol 2020;258:17-21.
10. Ragab DA, Sharkas M, Marshall S, Ren J. Breast cancer detection using deep convolutional neural networks and support vector machines. PeerJ 2019;7:e6201.
11. Xiao T, Liu L, Li K, Qin W, Yu S, Li Z. Comparison of transferred deep neural networks in ultrasonic breast masses discrimination. Biomed Res Int 2018;2018:4605191.
12. Byra M, Galperin M, Ojeda-Fournier H, Olson L, O'Boyle M, Comstock C, Andre M. Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion. Med Phys 2019;46:746-55.
13. Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. Data Brief 2019;28:104863.
14. Rodrigues PS. Breast Ultrasound Image. Mendeley Data 2017. doi: 10.17632/wmy84gzngw.1
15. Al-Dhabyani W, Gomaa M, Khaled H, Aly F. Deep learning approaches for data augmentation and classification of breast masses using ultrasound images. Int J Adv Comput Sci Appl 2019;10:e14464.
16. Singh VK, Rashwan HA, Abdel-Nasser M, Sarker M, Kamal M, Akram F, Pandey N, Romani S, Puig D. An Efficient Solution for Breast Tumor Segmentation and Classification in Ultrasound Images Using Deep Adversarial Learning. arXiv preprint arXiv:190700887 2019.
17. Bisong E. Google AutoML: Cloud Vision. Building Machine Learning and Deep Learning Models on Google Cloud Platform. Verlag: Springer, 2019:581-98.
18. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, Niculae V, Prettenhofer P, Gramfort A, Grobler J. API design for machine learning software: experiences from the scikit-learn project. arXiv preprint arXiv:13090238 2013.
19. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825-30.
20. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M (eds). Tensorflow: A system for large-scale machine learning. Savannah, USA: 12th USENIX Symposium on Operating Systems Design and Implementation, 2016.
21. Bradski G, Kaehler A. The opencv library. Dr Dobb's J. Software Tools 2000;25:120-5.
22. Huang Q, Zhang F, Li X. Machine learning in ultrasound computer-aided diagnostic systems: a survey. Biomed Res Int 2018;2018:5137904.
23. Kumar RM, Sreekumar K. A survey on image feature descriptors. Int J Comput Sci Inf Technol 2014;5:7668-73.
24. Miyamoto E, Merryman T. Fast calculation of Haralick texture features. Human computer interaction institute. Carnegie Mellon University, Pittsburgh, USA: Japanese Restaurant Office, 2005.
25. Huang Z, Leng J (eds). Analysis of Hu's moment invariants on image scaling and rotation. Chengdu, China: 2010 2nd International Conference on Computer Engineering and Technology, 2010.
26. Qi Y. Random forest for bioinformatics. Ensemble machine learning. Springer, 2012:307-23.
27. Aswathy M, Jagannath M. Detection of breast cancer on digital histopathology images: Present status and future

possibilities. Inform Med Unlocked 2017;8:74-9.

28. Carneiro PC, Franco MLN, Thomaz RL, Patrocinio AC. Breast density pattern characterization by histogram features and texture descriptors. Biomed Eng Res 2017;33:69-77.
29. Foley D. Considerations of sample and feature size. IEEE Trans Inf Theory 1972;18:618-26.
30. Zhang H. The optimality of naive Bayes. Menlo Park: Proceedings of 17th International Florida Artificial Intelligence Research Society Conference, 2004:562-7.
31. Alom MZ, Taha TM, Yakopcic C, Westberg S, Sidike P, Nasrin MS, Hasan M, Van Essen BC, Awwal AA, Asari VK. A state-of-the-art survey on deep learning theory and architectures. Electronics 2019;8:292.
32. Liu S, Wang Y, Yang X, Lei B, Liu L, Li SX, Ni D, Wang T. Deep learning in medical ultrasound analysis: a review. Engineering 2019;5:261-75.
33. Zeng Y, Zhang J. A machine learning model for detecting invasive ductal carcinoma with Google Cloud AutoML Vision. Comput Biol Med 2020;122:103861.
34. Cheng HD, Shan J, Ju W, Guo Y, Zhang L. Automated breast cancer detection and classification using ultrasound images: A survey. Pattern Recognition 2010;43:299-317.

Cite this article as: Wan KW, Wong CH, Ip HF, Fan D, Yuen PL, Fong HY, Ying M. Evaluation of the performance of traditional machine learning algorithms, convolutional neural network and AutoML Vision in ultrasound breast lesions classification: a comparative study. Quant Imaging Med Surg 2021;11(4):1381-1393. doi: 10.21037/qims-20-922