Physica Medica 107 (2023) 102560

Contents lists available at ScienceDirect

Physica Medica

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

A novel deep learning model for breast lesion classification using ultrasound Images: A multicenter data evaluation

Nasim Sirjani [a,1], Mostafa Ghelich Oghli [a,*,1], Mohammad Kazem Tarzamni [b], Masoumeh Gity [c], Ali Shabanzadeh [a], Payam Ghaderi [d], Isaac Shiri [e], Ardavan Akhavan [a], Mehri Faraji [a], Mostafa Taghipour [a]

[a] Research and Development Department, Med Fanavaran Plus Co., Karaj, Iran
[b] Department of Radiology, Imam Reza Hospital, Tabriz University of Medical Sciences, Tabriz, Iran
[c] Department of Radiology, Advanced Diagnostic and Interventional Radiology Research Center (ADIR), Medical Imaging Center, Imam Khomeini Complex Hospital, Tehran, Iran
[d] Besat Hospital, Kurdistan University of Medical Sciences, Sanandaj, Iran
[e] Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, CH-1211 Geneva 4, Switzerland

ARTICLE INFO

Keywords:
Breast ultrasound
Deep learning
Convolutional neural network
Image classification

ABSTRACT

Purpose: Breast cancer is one of the major reasons of death due to cancer in women. Early diagnosis is the most critical key for disease screening, control, and reducing mortality. A robust diagnosis relies on the correct classification of breast lesions. While breast biopsy is referred to as the "gold standard" in assessing both the activity and degree of breast cancer, it is an invasive and time-consuming approach.
Method: The current study's primary objective was to develop a novel deep-learning architecture based on the InceptionV3 network to classify ultrasound breast lesions. The main promotions of the proposed architecture were converting the InceptionV3 modules to residual inception ones, increasing their number, and altering the hyperparameters. In addition, we used a combination of five datasets (three public datasets and two prepared from different imaging centers) for training and evaluating the model.
Results: The dataset was split into the train (80%) and test (20%) groups. The model achieved 0.83, 0.77, 0.8, 0.81, 0.81, 0.18, and 0.77 for the precision, recall, F1 score, accuracy, AUC, Root Mean Squared Error, and Cronbach's α in the test group, respectively.
Conclusions: This study illustrates that the improved InceptionV3 can robustly classify breast tumors, potentially reducing the need for biopsy in many cases.

Introduction

According to the statistics published by the American Cancer Society in 2021, female breast cancer is the second leading cause of death due to cancer (15%) and the most commonly diagnosed cancer in women. Specifically, it is the reason for 14.9% and 30% of new incidences in both sexes and females, respectively [1]. Early detection and more precise diagnosis methods, primarily relying on imaging techniques, can help reduce breast cancer's mortality rate [2]. Although biopsy serves as the gold standard for classifying breast lesions into benign and malignant, breast ultrasound has been emphasized as a valuable method as it is practically harmless, non-invasive, real-time, and cost-effective [2]. On the other hand, in patients in the early stages of breast cancer, accurate detection of the axillary lymph node metastasis status is essential to avoid unnecessary axillary surgery and complications [3]. Therefore, several studies have proposed ultrasound as a reliable alternative for unnecessary lymph node biopsy with a relatively remarkable correlation [4]. In addition, deep-learning approaches have been proposed for predicting lymph node status in recent years [5,6]. However, ultrasound imaging usually misses tumor classification and tends to detect small cancerous lumps without typical malignant features. Moreover, ultrasonography is an operator-dependent procedure affected by inter-user variability. Putting it together makes classifying breast masses as benign or malignant challenging [7]. Breast tumor classification is still

[*] Corresponding author at: 10th St. Shahid Babaee Blvd., Payam Special Economic Zone, Karaj, Iran.
E-mail address: m.g31_mesu@yahoo.com (M. Ghelich Oghli).
[1] Both authors contributed equally to this manuscript.

https://doi.org/10.1016/j.ejmp.2023.102560
Received 16 August 2022; Received in revised form 20 February 2023; Accepted 26 February 2023
Available online 4 March 2023

inquiring due to the low signal-to-noise ratio and variability of tumors in shape, size, appearance, texture, and location [8].

Different computer-aided diagnosis (CAD) systems have been developed to hasten breast tumor detection [9]. Conventional CAD systems encompass different steps toward a breast cancer diagnosis. In the first step, the suspicious region is segmented semi- or even fully automatically. Secondly, quantitative metrics, such as morphological or texture features, are extracted from the segmented area using different methods [2,10]. Finally, after feature extraction steps, eclectic classification algorithms learn conspicuous features of breast tumor regions to classify these tumors [8,11–16].

Nevertheless, these techniques' performance depends on the coincidence degree of classification algorithms and produced features [17]. For example, Moon et al. [18]developed a CAD system that uses speckle features of breast ultrasound images for breast tumor classification and attains good sensitivity and precision. Flores et al. [19] improved the accuracy of breast tumor classification by using a combination of morphological and texture features. Byra et al. [20] enhanced the classification precision in ultrasound breast tumors utilizing the segmented quantitative ultrasound maps of homodyned K distribution parameters. They demonstrated that examining the interior shifts in lesion parametric maps resulted in higher accuracy in classification. Uniyal et al. [21] utilized the ultrasound radiofrequency (RF) time series analysis to classify malignant tumors in ultrasound images. In all of these CAD systems, the availability of a distributed database from different resources provides several advantages. For instance, the performance of the CAD system can be improved in terms of sensitivity and specificity. In this regard, The MAGIC-5 Project aimed at developing an infrastructure connection for sharing resources and devising suitable CAD software [22]. According to the MAGIC-5 GRID philosophy, only images with a high probability of carrying a pathology are moved over the network to the diagnostic centers, where the physicians can analyze them by taking advantage of the CAD selection almost in real-time.

Deep learning (DL), specifically convolutional neural networks (CNNs), has received great attention for different tasks and achieved excellent performances in different imaging modalities, containing magnetic resonance imaging (MRI), microscopic imaging, computer tomography (CT) scan, ultrasound, X-ray imaging [23–28]. In one study, the authors proposed a deep learning approach to predict if a patient with advanced breast cancer gained a complete pathological response (pCR) [26]. They used a pre-trained CNN to extract low-level features from breast MRI images and a support vector machine (SVM) to classify an optimal set of the most stable ones. Liu et al. [27] used the same database to predict neoadjuvant chemotherapy (NAC) using a CNN with 3 × 3 kernels. Compared to conventional CAD, CNN architectures can automatically extract conspicuous features and perform feature selection and classification tasks in one module[29,30]. Wang et al. [31] proposed a modified Inception-v3 architecture for breast tumor classification as benign and malignant. Shi et al. [32] designed a stacked deep polynomial network (S-DPN) for breast tumor classification using ultrasound imaging. Cao et al. [33] examined object detection, such as Fast R-CNN, Faster R-CNN, SSD, YOLO, and classification methods based on several CNN architectures, like AlexNet, ZFNet, VGG16, GoogLeNet, Resnet, and Densenet, for lesion detection and classification in breast ultrasound images.

In this paper, we aimed to classify breast lesions between benign and malignant using a novel deep neural network architecture developed based on Inception-V3. Inception-V3 is a CNN architecture proposed by Szegedy et al. [34], improved with respect to other versions of Inception networks using an auxiliary classifier to propagate ground truth information more in the network's depth. The squeeze and excitation blocks are one of Inception V3′s vital features, exciting feature maps dynamically for classification improvement and suppressing the ones not aiding according to the feature map global averages pattern. In addition, the convolutional block attention module encompassing channel and spatial attention modules has been used to improve expression power using the attention mechanism, which focuses on prominent features and suppresses unneeded ones.

Three main modifications that are performed in our proposed architecture are (a)converting Inception modules to residual ones, (b) changing the number of modules, and (c) altering the hyperparameters. Other renowned networks were employed to evaluate this one, such as original InceptionV3, InceptionResnetV2, ResnetV1-V2, Mobilenets, Xception, VGG, Resnext, Densenet, and SqueezeNet, [34–41] with different variations containing 25 models in total.

Materials and methods

Dataset description

This study used 2D breast ultrasound images to train and evaluate the networks from public and private databases. Three public datasets and two in-house datasets are included. The description of all datasets is provided in Table 1. The public datasets include: (I) the Breast Ultrasound Images Dataset (BUSI) [42], (II) The BUS dataset, collected from the UDIAT Diagnostic Centre of the Parc Taulí Corporation, Sabadell, Spain, using a Siemens ACUSON Sequoia C512 ultrasound machine [43] and (III) A public dataset of 86 breast cancer ultrasound images obtained by a Philips iU22 ultrasound machine at the Department of Radiology, Thammasat University Hospital, Bangkok, Thailand [44]. The first in-house dataset was collected at the Tabesh imaging center in Tabriz, Iran. All images were acquired by the Voluson E8 ultrasound imaging system (GE, Boston, United States). Fig. 1 shows some examples of breast ultrasound images in this dataset.

In addition, another in-house dataset (Dr. Gity imaging center) was gathered and used for testing the network as the external dataset to evaluate the model's robustness, reliability, and generalizability. The images are acquired by the Vinno M80 ultrasound imaging system (Vinno, Suzhou, China).

For the in-house datasets, informed consent was obtained from all patients. The study was approved by our institutional review board. Fig. 2 summarizes the described datasets. This dataset contained BI-RADS scores from the radiologist reports, and its benign-malignant labels were assigned based on BI-RADS scores. This dataset's primary purpose is to compare our proposed model and a routine radiologist's diagnosis for classifying benign and malignant tumors.

Image preprocessing

As the preprocessing step, three-channel images were resized to 224 × 224 and converted to grayscale. Each image has a normalization degree in which its pixels are normalized to a range of 0 and 1. It is worth noting that different acquisitions and scanners pose different noise values to the image. However, we assume that deep learning could handle this variability across these centers. In addition, we split the data set using stratification with respect to centers, in which both training and test sets contain the same portion from each center. Our results show good performance across test sets, representing different data from different sources (acquired with different scanners and acquisition settings). Harmonization could be achieved in feature levels using conventional algorithms such as ComBat in radiomics and deep learning algorithms such as GAN at the image level. However, the harmonization of the data set across the different centers is out of the scope of the current study. In addition, the dynamic augmentation during the training step was also implemented. This idea increases the robustness of the model against image variations.

After data preprocessing, 80% of the dataset was appropriated to the training set and 20% to the test set. As various papers do, we could train our network on public datasets and test them on internal ones. However, we encountered two problems when implementing this idea. First, there was an intrinsic difference between public and in-house datasets. As described in the BUSI dataset description, all images were cropped to

Table 1
Characteristics of Patients in the Data set.

Dataset name	Public/private	Meta-data			benign	malignant	Total
		age	Images/patient	laterality			
BUSI	Public	25–75	1	NA	487	210	697
BUS	Public	NA	1	NA	110	53	163
Pubic dataset 3	Public	NA	1	NA	0	86	86
Tabesh	Private	23–78	1.06	1.1	382	178	560
Dr. Gity imaging center	Private	20–72	1.03	1.15	71	79	150

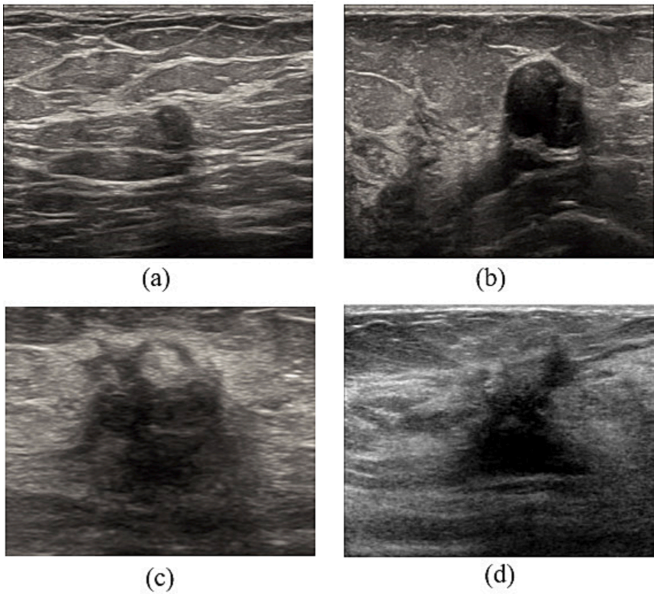

Fig. 1. The breast ultrasound tumor samples used in this article, (a) and (b), are benign tumors, and (c) and (d) are malignant ones.

remove unused and unimportant boundaries [42]. In contrast, our collected images were obtained from clinical routines with much more complicated ones and a lot of redundant data. Another problem was the percentage of in-house datasets relative to the total images, 42.87%. Obviously, training the network with 57% of the available images is not a reasonable approach.

Proposed architecture

The InceptionV3 deep convolutional architecture, introduced by Szegedy et al. [34], and inspired by GoogleNet, was used as our proposed algorithm's base network. The architecture's main superiority over VGGNet is deploying less computational power by modifying the previous Inception architectures. Thus, any further modification of the architecture needs to preserve this advantage. The original InceptionV3 architecture's major features contain:

1- Factorized convolution kernels: converting N × N convolution kernels to a product of two in Nx1 and 1xN Shapes. This trick reduces the computational complexity by reducing the number of parameters involved in the network. A few parameters of Inception-v3 make it more practical to implement it efficiently in a standard server to provide a rapid response service. In addition, kernels with diverse sizes enable them to possess distinct areas' receptive fields.
2- Smaller convolution kernels, reducing the computational complexity
3- Asymmetric convolutions: This trick replaces a 3 × 3 convolution with a 1 × 3 followed by a 3 × 1 one.
4- Auxiliary classifiers: an auxiliary classifier is a small CNN component that helps very deep architectures to converge by pushing proper gradients to the initial layers and preventing the gradient from vanishing.
5- Grid size reduction: This means pooling after a convolution operation and activation to avoid the representational bottleneck.

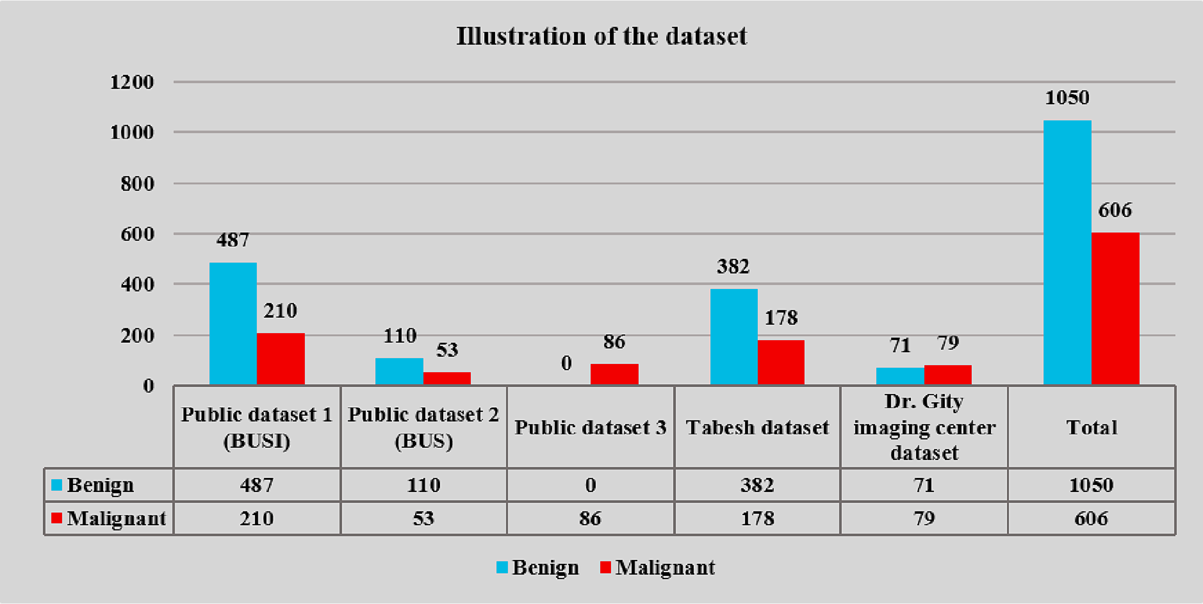

Fig. 2. The datasets' details. The first four datasets were used for training and evaluating the network, while the Dr. Gity imaging center dataset was used as an unseen data.

Between the auxiliary classifier and the fully connected (FC) layer in Inception-v3, a batch normalization (BN) layer as a regularizer is implemented to accelerate the training pace and model convergence. In addition, the batch normalization layer optimizes the learning influence by normalizing the input into each layer.

The original InceptionV3 consists of three inception modules: A, B, and C. These modules are composed of several convolutional and pooling layers parallel, as depicted in Fig. 3. Inception Module A employs the factorization technique to decline the number of parameters, which instead of using one layer of 5 × 5 filter, utilizes two layers of 3 × 3 filters. On the other hand, Inception Module B decreases the number of parameters by operating the factorization into asymmetric convolutions approach in which one 3 × 1 convolution is followed by one 1 × 3 convolution instead of one 3 × 3 convolution. In addition, Inception Module C is also proposed for facilitating high-dimensional representations. Thus, by reducing the number of parameters, these three kinds of Inception Modules decrease the likelihood of overfitting and by which the network can go deeper.

We revised the structure by converting these modules to residual inception ones to obtain more precise results, changing their number in the construction, and altering the hyperparameters. As shown in Fig. 3, there is a shortcut connection at the bottom of each module; this shortcut connection helps to go deeper into ResNet architecture and simultaneously to hand gradient vanishing and exploding problems.

Breast tumor classification suffers from different problems, such as the tumors' shape diversity and the limited number of samples. Consequently, the suggested model was pre-trained over the large dataset of ImageNet to improve the results using the transfer learning technique. Various studies have achieved acceptable performance in medical image classification regardless of the data size, how the data is augmented, how CNN models are used, or whether transfer learning is used [45]. In other words, transfer learning using ImageNet, a non-medical dataset, might be an effective strategy for approaching medical problems. To fix the problem of using the grayscale image as the input layer, we concatenated three copies of the input grayscale image to create a 3D (color) image, as it was in ImageNet. Fig. 3 depicts the proposed network's complete architecture, modified InceptionV3 [46]. The network takes the batch of the one-channel breast ultrasound images as input and yields the probabilities that show each class's liking for inputs.

Training the network

Using the backpropagation algorithm, the network illustrated in Fig. 3 is trained end-to-end on the breast ultrasound datasets. The initial weights for the network are random. The images were randomly chosen from each epoch from the comprehensive training dataset as a batch. The loss function used in the training procedure is binary cross-entropy loss.

Implementation details

This study utilizes a system with 16 GB of RAM, a GPU-based graphic card with 2176 CUDA cores (GeForce RTX 2060-A8G), and an Intel Xeon CPU. We implemented the network in the Python environment with Tensorflowr 2.8.0 [47] and Keras 2.8.0 [48]. We used random search on a few iterations for the hyperparameter optimization, such as regularization parameters and learning rate. We employed RMSprop as an optimizer to train the network with a learning rate of 10e-4.

Evaluation metrics

The model is evaluated by comparing the network's predictions to the true labels using some of the key metrics in the classification task, including precision, recall, F1 score, accuracy, area under the curve (AUC), root mean squared error (RMSE), and Cronbach's α. The first five metrics are binary classification metrics using a confusion matrix to calculate true positive (TP), true negative (TN), false positive (FP), and false negative (FN) [49]. On the other hand, RMSE and Cronbach's α evaluate the predicted values utilizing probabilities of the network outcomes (e.g., from a least-squares fit). These metrics are concisely

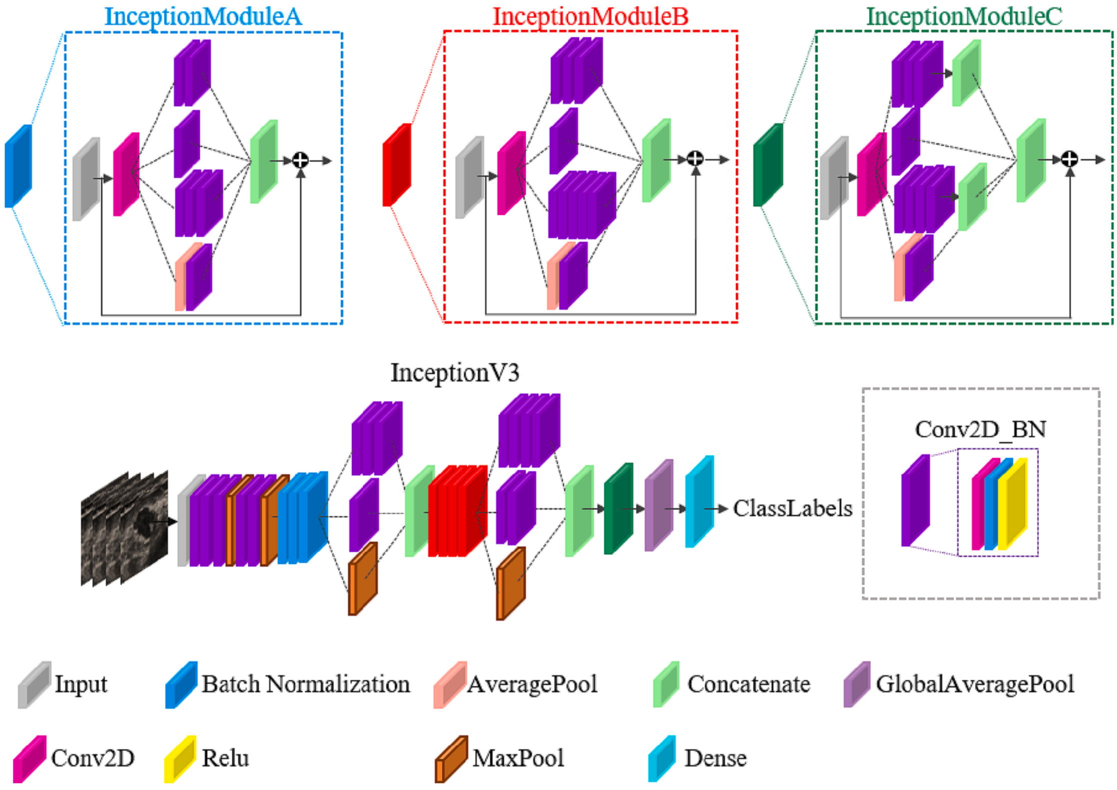

Fig. 3. Improved InceptionV3 structure with residual inception blocks.

explained below.

Cronbach's α is a reliability metric determining how closely related a set of items (also known as scale) are as a group [50]. The scores from each scale item are compared to the total score for each observation, and the variance is calculated for each item to compute Cronbach's α. Equation (1) indicates this metric's formula, where σ_i is the covariance between the prediction and ground truth and σ_x is the prediction's sum and ground truth covariance separately.

$$Cronbach^{'}s\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n}\sigma_i^2}{\sigma_x^2}\right) \quad (1)$$

According to the equation, Cronbach's α reliability coefficient ranges typically between zero and one. It has a higher value when closer to 1.0, meaning the items in a scale have a higher consistency level (or homogeneity between them).

The RMSE is defined as the root of the mean of the squared errors over the data instances [51]. Equations (2) describe the RMSE formula.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i-1}^{n}(y_i - \widehat{y}_i)^2} \quad (2)$$

Results

Table 2 summarizes the proposed network's results and state-of-the-art convolutional neural networks. We used precision, recall, F1 score, accuracy, AUC, RMSE, and Cronbach's α to assess the approach. The train-test split rate was 80% and 20%, respectively. The proposed architecture attained 83% precision, 77% recall, 80% of F1 score, 81% accuracy, 18% of RMSE, and 77% of Cronbach's α for breast tumors classification. Table 2 presents the results of the proposed network and conventional, well-known ones; as shown in this table, the proposed network outperforms all conventional CNN architectures.

We evaluated the proposed network with an unseen external dataset (Dr. Gity imaging center) to compare our proposed model's performance and radiologists'. Since our method is designed to classify the images into "benign" and "malignant", we need to assign a label to each image based on the sonography report. On the other hand, the radiologist does not determine whether the lesion is benign or malignant in the report. Therefore, we had to use either definitely benign or malignant images (i.e., they did not need a biopsy). The benign-malignant labels of this dataset were assigned based on BI-RADS scores of the lesions that the radiologist reports. In this regard, lesions with BI-RADS 1 to 3 were considered benign, and tumors with B-RADS 5 and 6 were considered malignant. Since lesions with BI-RADS 4 (a, b, and c) scores are suspicious, we did not enroll these images in our test sets. Since we aimed to compare the network performance with radiologists' efficiency, we just tested the best model in terms of accuracy, the Improved InceptionV3. The results of all metrics are illustrated in Table 3. Table 3 shows our proposed network to achieve high performance when faced with an unseen dataset labeled based on the radiologist's diagnosis. The receiver operating characteristic curve (ROC curve) and area the under the curve (AUC) for this test procedure is shown in Fig. 4.

Fig. 5 shows the ROC curve for all the tested models. As expected, the proposed architecture's AUC surpasses all other methods. The comparison between AUC models is also shown in Fig. 6.

Discussion

In this study, we present a novel CNN architecture, namely Improved InceptionV3, which helps radiologists classify benign and malignant lesions in breast ultrasound images. The network was trained and evaluated on comprehensive multi-national and multi-institutional datasets. This diversity maintained a high level of diagnostic accuracy for our system. In addition, testing the performance by an unseen dataset demonstrates the model's generalization and robustness across different images with different ultrasound devices and pre-sets. A fast and accurate distinction of tumors is essential and could result in lower expenditure and death rates [52]. Furthermore, the proposed CNN-based models can automatically predict the type of breast tumor using ultrasound images.

We compared the results of our method with 24 CNN models and a radiologist's diagnosis. To our knowledge, this is the first attempt to compare these many models using the same datasets, providing a good insight into the result of utilizing various CNN blocks and modules. The Improved InceptionV3 results lead to a precise appraisal of the breast tumor kinds that can be malignant or benign. The best precision (83%), accuracy (81%), F1 score (80%), and AUC (91%) were achieved by our model. The proposed model is also a relatively sensitive approach that (considering the higher accuracy and precision) can be a candidate for a screening model. The SqueezeNet's sensitivity is slightly higher than our

Table 2
The result of the state-of-the-art CNN models on the common dataset. Best values are bold.

Network	Precision	Recall	F1	Accuracy	AUC	RMSE	Cronbach's α
InceptionV3	0.72	0.75	0.74	0.74	0.74	0.50	0.65
InceptionV3-se	0.73	0.82	0.77	0.76	0.74	0.48	0.70
InceptionV3-cbam	0.80	0.63	0.70	0.74	0.77	0.50	0.66
InceptionResnetV2	0.82	0.72	0.76	0.78	0.8	0.46	0.73
InceptionResnetV2-se	0.82	0.75	0.78	0.80	0.78	0.44	0.75
InceptionResnetV2-cbam	0.78	0.78	0.78	0.79	0.78	0.45	0.73
Densenet	0.78	0.64	0.70	0.73	0.71	0.51	0.64
Densenet-se	0.73	0.55	0.63	0.68	0.67	0.31	0.55
Densenet-cbam	0.8	0.56	0.66	0.71	0.7	0.28	0.62
Mobilenets	0	0	0	0.51	0.5	0.69	0
Mobilenets-se	0	0	0	0.51	0.5	0.69	0
Mobilenets-cbam	0	0	0	0.51	0.5	0.69	0
ResnetV1	0.70	0.58	0.64	0.67	0.65	0.56	0.53
ResnetV1-se	0.79	0.50	0.61	0.69	0.66	0.55	0.57
ResnetV1-cbam	0.62	0.46	0.52	0.59	0.59	0.63	0.33
ResnetV2	0.80	0.41	0.54	0.66	0.66	0.57	0.53
ResnetV2-se	0.65	0.60	0.62	0.65	0.63	0.58	0.46
ResnetV2-cbam	0.81	0.43	0.56	0.67	0.68	0.56	0.55
Resnext	0.62	0.53	0.57	0.61	0.6	0.61	0.38
Resnext-se	0.70	0.40	0.51	0.62	0.63	0.61	0.42
Resnext-cbam	0.71	0.42	0.52	0.63	0.63	0.60	0.44
VGG	0.76	0.71	0.73	0.75	0.76	0.24	0.67
Xception	0.64	0.60	0.62	0.64	0.64	0.35	0.43
SqueezeNet	0.48	1	0.65	0.48	0.5	0.51	0
Improved-InceptionV3	0.83	0.77	0.80	0.81	0.81	0.18	0.77

Table 3
The result of the proposed model on an un-seen data (Dr. Gity imaging center dataset).

Network	Precision	Recall	F1	Accuracy	RMSE	Cronbach's α
Improved-InceptionV3	0.85	1	0.92	0.91	0.09	0.91

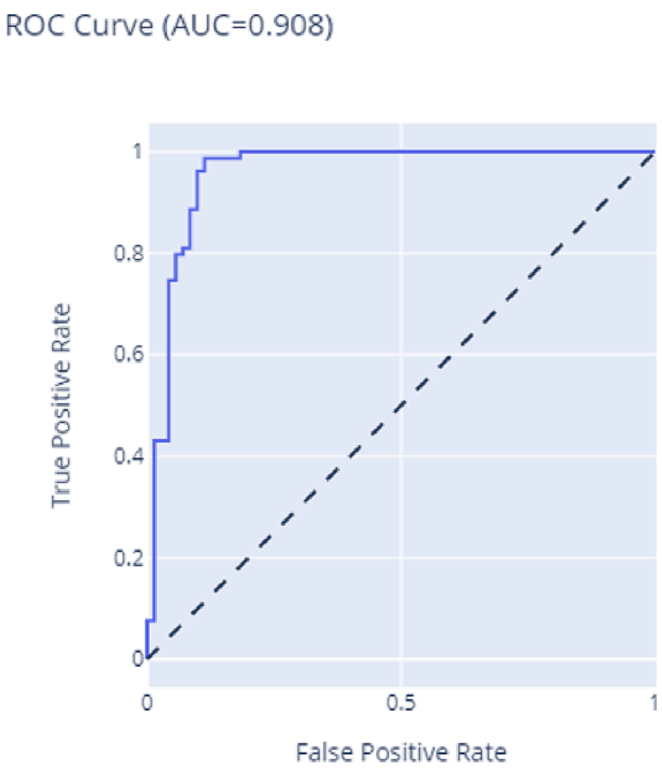

Fig. 4. ROC Curve and AUC value for the unseen data prediction.

model's, with unacceptable precision (0.48). Sensitivity is the ratio of true positive samples to the total positive samples (true positive + false negative). Therefore, regarding dramatically lower values of other metrics, this high sensitivity value shows that this model predicts most of the test samples as malignant, leading the number of false negatives in the sensitivity definition to zero.

All image labels used for the training and evaluation step, including public datasets [42–44] and the Tabesh dataset, were validated using a biopsy report. Testing a network that benefits from a biopsy report as the gold standard on a dataset labeled based on the radiologist's diagnosis provide a proper insight into the proposed model's value in identifying malignant tumors. Compared with other studies that train their networks based on the BI-RADS scores (indicated in the sonography report) as the gold standard, our approach achieved excellent performance. The results of the unseen dataset even outperformed training-validation sets, showing the value of validating the data using a biopsy report. One of the main advantages of Deep learning compared to radiomics or any method that needs segmentation (conventional CAD) would be eliminating the segmentation step. Segmentation elimination is one of the main steps toward feasible CAD in real-time ultrasound. Moreover, a radiology expert is needed to do the segmentation, and the segmentation process is time-consuming, labor-intensive, and prone to inter/intra-observer variability.

Several studies used the presented public dataset (BUSI dataset [42]) and achieved different results. For instance, authors in [53] achieved 95.6% accuracy at the best performance for classifying images into Normal, Benign, and Malignant. In [54], the BUSI dataset was used to classify breast ultrasound images into benign and malignant and gained 95.48% accuracy. The main reason these studies perform dramatically is that the input images are well-defined. This means they have been cropped around the breast lesion (removing all redundant information), making these networks' performance so well.

Morid et al. [45] presented a scoping review of transfer learning research on medical image analysis using the ImageNet dataset. They studied various CNN architectures and various imaging modalities tested on different anatomical sites. The results show that Inception models were the most frequently used for studies that analyzed ultrasound images, indicating that wide networks (rather than deep

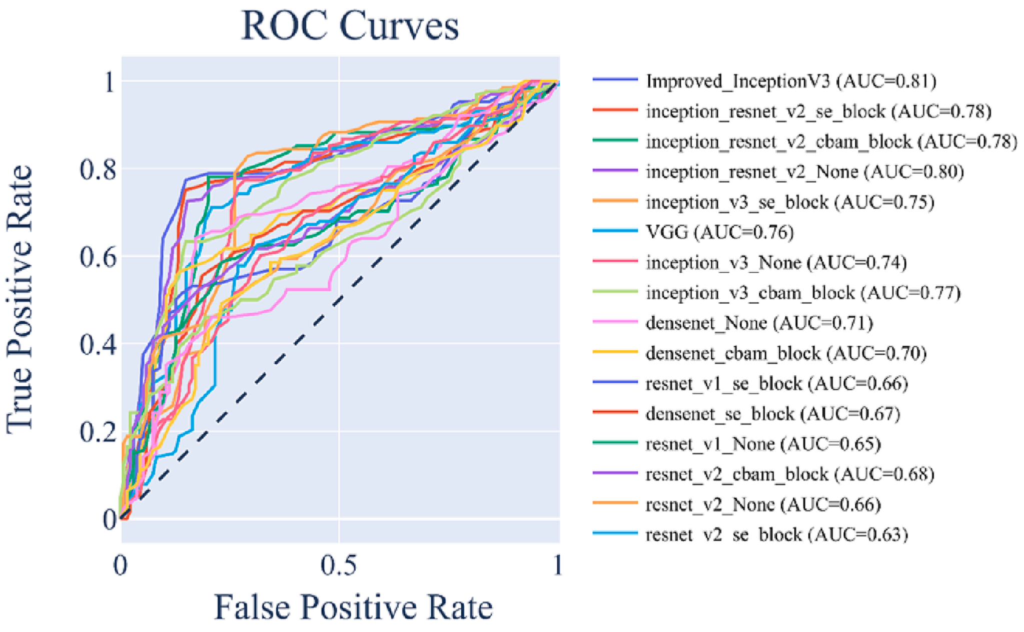

Fig. 5. ROC curves for all models.

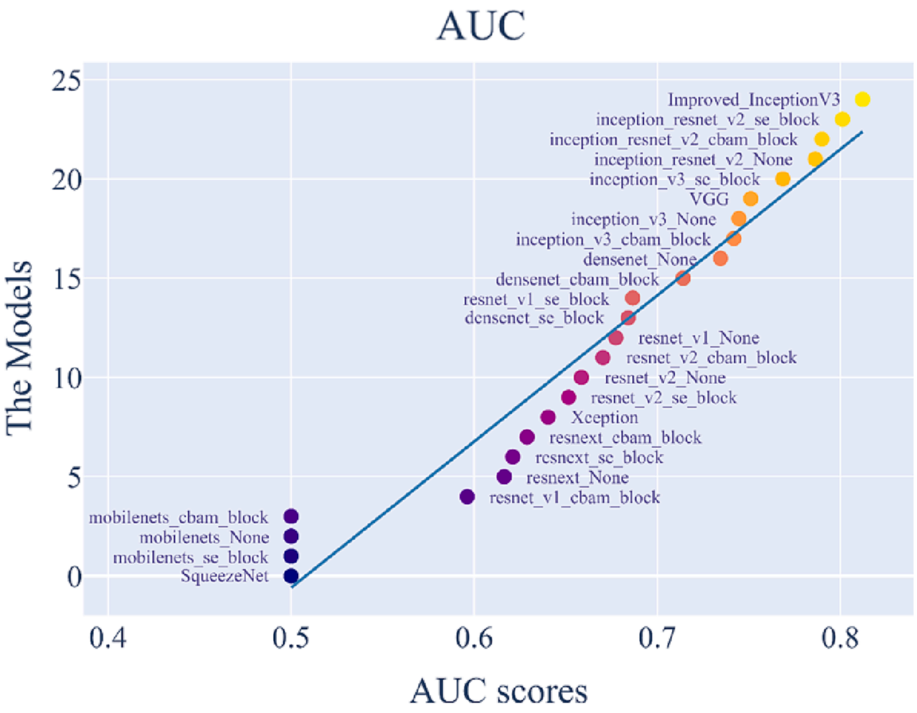

Fig. 6. AUC scores for all models.

networks) with inception modules could be more effective for these types of images. The authors concluded that the transfer learning approach improves the performance of deep CNNs, especially when the number of training images is limited.

This work's most important limitation is that our classification system relies only on a single ultrasound image. In contrast, radiologists look at a breast lesion from different views to understand its morphology comprehensively. A possible improvement of the current study is to develop an AI system to analyze the breast lesions from a video captured using probe screening on the breast. In this way, the network's input would be an image sequence covering all breast lesion views from the observer's point of view. Moreover, our system does not consider the lesion's historical aspects and morphological changes over time. The other limitation is that this system ignores the demographic characteristics of the patients (such as age, family history of cancer, marital status, etc.). At the same time, this information undeniably impacts the radiologist's diagnosis in clinical practice. For instance, Polchai et al. [55] used age, postmenopausal status, estrogen receptor (ER) status, human epidermal growth factors receptor 2 (HER2) status, detection method, and generation of chemotherapy to validate the use of CncerMath and prediction as prognostic tools in Thai breast cancer patients. In another work, Fanizzi et al. [56] utilized age, histological subtype, ER status, progesterone receptor expression (PR) status, histological grade, and sentinel lymph node status. In addition, radiologists usually have access to patients' mammography images. This difference in the system information likely leads to the insufficiency of AI system deployment in clinical practice. There are many possible promotions that future works can focus on, for instance, boosting the datasets and enhancing the network structure. Despite the diversity of the trained-evaluated datasets and being from various sources, the number of images is insufficient to be asserted as a comprehensive dataset.

Our proposed network, a modified version of Inception V3, can extract meaningful features from different imaging sources containing medical or non-medical images. The original paper was evaluated on the ImageNet dataset, a large hand-labeled natural dataset [57]. On the other hand, recent studies have tested Inception V3 on various medical imaging modalities such as MRI [58], CT scan [59], X-ray [60], and OCT [61]. Since the backbone of the proposed network is the same as the Inception V3, and the feature extraction blocks did not change dramatically, we can conclude that the proposed network is suitable for feature extraction of other imaging modalities.

Conclusion

We proposed a novel deep CNN structure for breast tumor classification. We can label the lesion in breast ultrasound images using the network prediction as malignant or benign. Several imaging centers were used to gather the training dataset for this study, and several vendors were used. As a result of the dataset's general characteristics, our model is suitable as an AI-based predicting software for clinical routines. The results demonstrate a reassuring performance in the classification task. The comparison with 24 other CNN models makes our evaluation more demonstrative. Forthcoming work will concentrate on increasing data and designing a unique architecture for breast cancer detection.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Med Fanavaran Plus Co.

References

[1] Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. CA Cancer J Clin 2021;71:7–33. https://doi.org/10.3322/CAAC.21654.

[2] Yap Moi Hoon, Goyal Manu, Osman Fatima, Martí Robert, Denton Erika, Juette Arne, et al. Breast ultrasound region of interest detection and lesion localisation. Artif Intell Med 2020;107:101880.

[3] Ahmed M, Purushotham AD, Douek M. Novel techniques for sentinel lymph node biopsy in breast cancer: a systematic review. Lancet Oncol 2014;15:e351–62. https://doi.org/10.1016/S1470-2045(13)70590-4.

[4] Kim GR, Choi JS, Han BK, Lee JE, Nam SJ, Ko EY, et al. Preoperative axillary US in early-stage breast cancer: Potential to prevent unnecessary axillary lymph node

dissection. Radiology 2018;288:55–63. https://doi.org/10.1148/RADIOL.2018171987/ASSET/IMAGES/LARGE/RADIOL.2018171987.TBL4.JPEG.

[5] Xueyi Zheng Zhao Yao Yini Huang Yanyan Yu Yun Wang Yubo Liu et al. Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer Nat Commun 11 1.

[6] Samantha Bove Maria Colomba Comes Vito Lorusso Cristian Cristofaro Vittorio Didonna Gianluca Gatta et al. A ultrasound-based radiomic approach to predict the nodal status in clinically negative breast cancer patients Sci Rep 12 1.

[7] Kim S-Y, Choi Y, Kim E-K, Han B-K, Yoon JH, Choi JS, et al. Deep learning-based computer-aided diagnosis in screening breast ultrasound to reduce false-positive diagnoses. Sci Rep 2021;11:1–11.

[8] Takemura A, Shimizu A, Hamamoto K. Discrimination of breast tumors in ultrasonic images using an ensemble classifier based on the AdaBoost algorithm with feature selection. IEEE Trans Med Imaging 2010;29:598–609. https://doi.org/10.1109/TMI.2009.2022630.

[9] Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism 2017;69S: S36–40. https://doi.org/10.1016/J.METABOL.2017.01.011.

[10] Tagliafico AS, Piana M, Schenone D, Lai R, Massone AM, Houssami N. Overview of radiomics in breast cancer diagnosis and prognostication. Breast 2020;49:74–80. https://doi.org/10.1016/J.BREAST.2019.10.018.

[11] Rodrigues R, Pinheiro A, Braz R, Pereira M, Moutinho J. Towards Breast Ultrasound Image Segmentation using Multi-resolution Pixel Descriptors. In: Proc 21st Int Conf Pattern Recognit. IEEE; 2012. p. 2833–6.

[12] Sadek I, Elawady M, Stefanovski V. Automated Breast Lesion Segmentation in Ultrasound Images 2016. doi:10.48550/arxiv.1609.08364.

[13] Shi X, Cheng HD, Hu L. MASS DETECTION AND CLASSIFICATION IN BREAST ULTRASOUND IMAGES USING FUZZY SVM. Proc 9th Jt Conf Inf Sci JCIS 2006 2006;2006:253–6. doi:10.2991/JCIS.2006.257.

[14] Dhahri Habib, Al Maghayreh Eslam, Mahmood Awais, Elkilani Wail, Faisal Nagi Mohammed. Automated breast cancer diagnosis based on machine learning algorithms. J Healthc Eng 2019;2019:1–11.

[15] Zakeri FS, Behnam H, Ahmadinejad N. Classification of benign and malignant breast masses based on shape and texture features in sonography images. J Med Syst 2012;36:1621–7. https://doi.org/10.1007/S10916-010-9624-7.

[16] Hagiwara Yuki, Koh Joel En Wei, Tan Jen Hong, Bhandary Sulatha V, Laude Augustinus, Ciaccio Edward J, et al. Computer-aided diagnosis of glaucoma using fundus images: a review. Comput Methods Programs Biomed 2018;165:1–12.

[17] Paterakis NG, Mocanu E, Gibescu M, Stappers B, van Alst W. Deep learning versus traditional machine learning methods for aggregated energy demand prediction. IEEE PES Innov Smart Grid Technol Conf Eur 2017;2017:1–6.

[18] Moon WK, Lo CM, Chang JM, Huang CS, Chen JH, Chang RF. Computer-aided classification of breast masses using speckle features of automated breast ultrasound images. Med Phys 2012;39:6465–73. https://doi.org/10.1118/1.4754801.

[19] Gómez Flores W, Pereira WCDA, Infantosi AFC. Improving classification performance of breast lesions on ultrasonography. Pattern Recognit 2015;48: 1125–36. https://doi.org/10.1016/J.PATCOG.2014.06.006.

[20] Byra Michał, Nowicki Andrzej, Wróblewska-Piotrzkowska Hanna, Dobruch-Sobczak Katarzyna. Classification of breast lesions using segmented quantitative ultrasound maps of homodyned K distribution parameters. Med Phys 2016;43(10): 5561–9.

[21] Uniyal Nishant, Eskandari Hani, Abolmaesumi Purang, Sojoudi Samira, Gordon Paula, Warren Linda, et al. Ultrasound RF time series for classification of breast lesions. IEEE Trans Med Imaging 2015;34(2):652–61.

[22] Bellotti R, Bagnasco S, Bottigli U, Castellano M, Cataldo R, Catanzariti E, et al. The MAGIC-5 project: medical applications on a grid infrastructure connection. IEEE Nucl Sci Symp Conf Rec 2004;3:1902–6. https://doi.org/10.1109/NSSMIC.2004.1462616.

[23] Saw SN, Ng KH. Current challenges of implementing artificial intelligence in medical imaging. Phys Med 2022;100:12–7. https://doi.org/10.1016/j.ejmp.2022.06.003.

[24] Litjens Geert, Kooi Thijs, Bejnordi Babak Ehteshami, Setio Arnaud Arindra Adiyoso, Ciompi Francesco, Ghafoorian Mohsen, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88.

[25] Kermani Saeed, Ghelich Oghli Mostafa, Mohammadzadeh Ali, Kafieh Raheleh. NF-RCNN: Heart localization and right ventricle wall motion abnormality detection in cardiac MRI. Phys Med 2020;70:65–74.

[26] Comes Maria Colomba, Fanizzi Annarita, Bove Samantha, Didonna Vittorio, Diotaiuti Sergio, La Forgia Daniele, et al. Early prediction of neoadjuvant chemotherapy response by exploiting a transfer learning approach on breast DCE-MRIs. Sci Reports 2021;11(1):1–12. https://doi.org/10.1038/s41598-021-93592-z.

[27] Liu MZ, Mutasa S, Chang P, Siddique M, Jambawalikar S, Ha R. A novel CNN algorithm for pathological complete response prediction using an I-SPY TRIAL breast MRI database. Magn Reson Imaging 2020;73:148–51. https://doi.org/10.1016/J.MRI.2020.08.021.

[28] Ho P-S, Hwang Y-S, Tsai H-Y. Machine learning framework for automatic image quality evaluation involving a mammographic American College of Radiology phantom. Phys Med 2022;102:1–8.

[29] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. https://doi.org/10.1038/nature14539.

[30] Mohammed MA, Al-Khateeb B, Rashid AN, Ibrahim DA, Abd Ghani MK, Mostafa SA. Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images. Comput Electr Eng 2018;70:871–82. https://doi.org/10.1016/J.COMPELECENG.2018.01.033.

[31] Wang Y, Choi EJ, Choi Y, Zhang H, Jin GY, Ko SB. Breast cancer classification in automated breast ultrasound using multiview convolutional neural network with transfer learning. Ultrasound Med Biol 2020;46:1119–32. https://doi.org/10.1016/J.ULTRASMEDBIO.2020.01.001.

[32] Shi J, Zhou S, Liu X, Zhang Q, Lu M, Wang T. Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset. Neurocomputing 2016;194:87–94. https://doi.org/10.1016/J.NEUCOM.2016.01.074.

[33] Cao Z, Duan L, Yang G, Yue T, Chen Q. An experimental study on breast lesion detection and classification from ultrasound images using deep learning architectures. BMC Med Imaging 2019;19:1–9. https://doi.org/10.1186/S12880-019-0349-X/TABLES/2.

[34] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2015;2016-Decem:2818–26. doi:10.48550/arxiv.1512.00567.

[35] Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size 2016. doi:10.48550/arxiv.1602.07360.

[36] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017 2016;2017-Janua:2261–9. doi:10.48550/arxiv.1608.06993.

[37] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications 2017. doi:10.48550/arxiv.1704.04861.

[38] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv:14091556 2014.

[39] Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated Residual Transformations for Deep Neural Networks. Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017 2016;2017-Janua:5987–95. doi:10.48550/arxiv.1611.05431.

[40] Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. 31st AAAI Conf Artif Intell AAAI 2017 2016:4278–84. doi:10.48550/arxiv.1602.07261.

[41] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. IEEE Conf Comput Vis pattern Recognit 2016:770–8.

[42] Al-Dhabyani Walid, Gomaa Mohammed, Khaled Hussien, Fahmy Aly. Dataset of breast ultrasound images. Data Br 2020;28:104863.

[43] Yap Moi Hoon, Pons Gerard, Marti Joan, Ganau Sergi, Sentis Melcior, Zwiggelaar Reyer, et al. Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J Biomed Heal Informatics 2018;22(4): 1218–26.

[44] Stanislav Makhanov. Ultrasound Images 2012. http://onlinemedicalimages.com/index.php/en/site-map.

[45] Morid Mohammad Amin, Borjali Alireza, Del Fiol Guilherme. A scoping review of transfer learning research on medical image analysis using ImageNet. Comput Biol Med 2021;128:104115.

[46] Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions. Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017 2016;2017-Janua: 1800–7. doi:10.48550/arxiv.1610.02357.

[47] Mada RO, Lysyansky P, Daraban AM, Duchenne J, Voigt JU. How to define end-diastole and end-systole?: Impact of timing on strain measurements. JACC Cardiovasc Imaging 2015;8:148–57. https://doi.org/10.1016/J.JCMG.2014.10.010.

[48] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. This paper is included in the Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16). Open access to the Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation is sponsored by US n.d.

[49] Ting KM. Confusion Matrix. Encycl Mach Learn Data Min 2017:260.

[50] Connelly LM. Cronbach's alpha. Medsurg Nurs 2014;20:1357–9. https://doi.org/10.1007/978-94-007-0753-5_622.

[51] Fürnkranz J, Chan PK, Craw S, Sammut C, Uther W, Ratnaparkhi A, et al. Mean Squared Error. Encycl Mach Learn 2011:653. doi:10.1007/978-0-387-30164-8_528.

[52] Breast Cancer: Statistics | Cancer.Net. CancerNet n.d. https://www.cancer.net/cancer-types/breast-cancer/statistics (accessed August 13, 2022).

[53] Eroğlu Y, Yildirim M, Çinar A. Convolutional Neural Networks based classification of breast ultrasonography images by hybrid method with respect to benign, malignant, and normal using mRMR. Comput Biol Med 2021;133:104407. https://doi.org/10.1016/J.COMPBIOMED.2021.104407.

[54] Zhuang Z, Yang Z, Raj ANJ, Wei C, Jin P, Zhuang S. Breast ultrasound tumor image classification using image decomposition and fusion based on adaptive multi-model spatial feature fusion. Comput Methods Programs Biomed 2021;208: 106221. https://doi.org/10.1016/J.CMPB.2021.106221.

[55] Polchai N, Sa-Nguanraksa D, Numprasit W, Thumrongtaradol T, O-Charoenrat E, O-Charoenrat P. P. A Comparison between the online prediction models cancermath and PREDICT as prognostic tools in thai breast cancer patients. Cancer Manag Res 2020;12:5549–59. https://doi.org/10.2147/CMAR.S258143.

[56] Annarita Fanizzi Domenico Pomarico Angelo Paradiso Samantha Bove Sergio Diotaiuti Vittorio Didonna et al. Predicting of Sentinel Lymph Node Status in Breast Cancer Patients with Clinically Negative Nodes: A Validation Study Cancers 13 2 352.

[57] Russakovsky Olga, Deng Jia, Su Hao, Krause Jonathan, Satheesh Sanjeev, Ma Sean, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis 2015;115 (3):211–52.

[58] Li J, Wang P, Li Y, Zhou Y, Liu X, Luan K. Transfer learning of pre-trained Inception-v3 model for colorectal cancer lymph node metastasis classification. IEEE Int Conf Mechatronics Autom 2018;2018:1650–4.

[59] Elmuogy Samir, Hikal Noha A, Hassan Esraa. An efficient technique for CT scan images classification of COVID-19. J Intell Fuzzy Syst 2021;40(3):5225–38.

[60] Shadin NS, Sanjana S, Lisa NJ. COVID-19 diagnosis from chest X-ray images using convolutional neural network (CNN) and InceptionV3. Int Conf Inf Technol 2021;2021:799–804.

[61] Vijayan T, Sangeetha M, Karthik B. Efficient analysis of diabetic retinopathy on retinal fundus images using deep learning techniques with inception v3 architecture. J Green Eng 2020;10:9615–25.