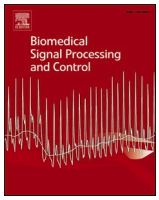
Biomedical Signal Processing and Control 84 (2023) 104819

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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



Mass segmentation and classification from film mammograms using cascaded deep transfer learning

Volkan Müjdat Tiryaki *

Department of Computer Engineering, School of Engineering, Siirt University, Siirt 56100 Turkey
Information Technology Application and Research Center, İstanbul Ticaret University, İstanbul 34840 Turkey

ARTICLE INFO

Keywords:
Mammography
Radiomics
Nodule
Malignant
Benign

ABSTRACT

Breast cancer is the most common type of cancer among women worldwide. Early breast cancers have a high chance of cure so early diagnosis is critical. Mammography screening allows early detection of breast cancer. There has been an increasing interest in the investigation of computer-aided breast cancer diagnosis recently due in part to the development of the novel high-performing deep learning models. In this study, cascaded deep transfer learning (DTL)-based segmentation methods were investigated to segment mass lesions using mammograms of Breast Cancer Digital Repository. In the first stage, the noise sources in the mammogram background were removed by deep learning-based breast segmentation. In the second stage, the mass segmentation performances of five-layer U-net and U-nets having pre-trained weights from VGG16, ResNet50, and Xception networks in the encoding path were investigated. The performances of attention U-net, residual U-net, Multi-ResUnet, DeepLabV3Plus, and Unet++ were also investigated. A Unet++ model that uses Xception network weights in the encoder region is proposed. The mass segmentation model predictions were used to estimate mass lesion characterization using DTL. On the test data, an AUC of 0.7829, Dice's similarity coefficient of 0.6356 and intersection over union of 0.5408 were obtained for mass segmentation using the proposed U-net++Xception model. An AUC of 0.8188 and accuracy of 0.7619 were obtained for mass classification into benign versus malignant. The results show that the proposed DTL pipeline can be used for automatic mass segmentation and classification without using clinical data and may reduce the workload of radiologists.

1. Introduction

Breast cancer is the most common type of cancer among women. Mammography is a standard 2D low-energy X-ray imaging method recommended for breast cancer screening for early detection as, breast cancer exhibits only subtle symptoms in the early stages. Early diagnosis is known to increase the probability of survival. Recently, machine learning and deep learning methods have been successfully applied for automatic detection and diagnosis of breast cancer from mammography [1–5]. Deep learning in particular has been shown to improve radiologists' performance in breast cancer screening [6]. However, research is needed to investigate and improve the performance of computer-aided methods for segmentation and classification of mass lesions.

Mass is a type of breast cancer tissue which appears as an abnormal region compared to the surrounding tissue on mammograms. Mass classification performance depends on the mass segmentation performance since mass shape has a correlation with malignancy. Mass segmentation from mammograms is a challenging task because mass shape, margin, texture, and density are all diverse. According to the mammography lexicon, mass shape can be oval, round, and irregular [7]. Mass margin can be circumscribed, obscured, microlobulated, indistinct, and speculated, and mass density can be fat-containing, low, equal, and high [7]. Additionally, mass segmentation performance depends on the mammogram density [8] because masses can be obscured by the fibroglandular tissue. Furthermore, mass and calcification abnormalities can overlap, which adds one more level diversity to the mass texture. These facts demonstrate the challenge in mass detection and segmentation.

Machine learning methods for mass segmentation require selection of features which can be time consuming. Deep learning methods allow automatic feature learning from the training data. Deep learning has been applied to the biomedical image semantic segmentation successfully by using fully convolutional neural networks called U-net [9], a supervised learning-based method that has encoder and decoder layers.

* Corresponding author at: School of Engineering, Block C, Room: C107, Siirt University Kezer Campus, Siirt 56100, Turkey.
E-mail addresses: vmtiryaki@ticaret.edu.tr, tiryakiv@siirt.edu.tr.

https://doi.org/10.1016/j.bspc.2023.104819
Received 14 March 2022; Received in revised form 10 January 2023; Accepted 5 March 2023
Available online 17 March 2023

Many variants of U-nets have been proposed to date including transfer learning methods, different skip connection designs, attention mechanisms, and visual transformer-based U-nets [10–14].

In this study, the Xception [15] transfer learning method is demonstrated to improve the performance of U-net++ for mass segmentation problem. The contributions of the present work can be summarized as: 1) A two-stage cascaded deep learning model was proposed to remove noises from mammogram and then to segment the mass lesion(s); 2) A first-time U-net++ implementation that uses weight coefficients of the Xception network in the encoder path is demonstrated as effective; 3) Mass classification is implemented and demonstrates the effectiveness of the proposed cascaded segmentation model without using clinical data. The flow diagram of the present study is shown in the graphical abstract.

2. Materials and methods

2.1. Mammogram data set selection, preprocessing, and mammogram annotation

The application of deep learning models requires training data. The data required for automated mass segmentation model should include mass lesions annotated by radiologists. Mass lesion annotation is a labor-intensive and a costly work and there are only a few datasets that include mass lesion annotations. The publicly available datasets that include mass lesion annotations are: the Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM), INbreast, and BCDR [16]. CBIS-DDSM is a large mammogram database that includes mass and calcification abnormalities from a total of 1,566 patients [8]. The mass lesion segmentation ground truth masks in CBIS-DDSM were obtained from a modified lesion segmentation algorithm, and the difference between the output of the algorithm and the radiologist assessment is quantified. The Dice's similarity coefficient (DSC) [17] between the mass outline from the segmentation algorithm and the mass outlines drawn by radiologist for types A, B, C, and D breast density mammograms are 0.904, 0.886, 0.749, and 0.808, respectively [8]. These values show the level of concordance between the radiologist's outline and the outlines that were provided in the dataset. INbreast data set includes 108 full-field digital mammograms that include mass lesions [18]. BCDR dataset includes 519 screen film mammograms with mass abnormalities and each mass lesion annotation was performed by a radiologist [19–22]. Therefore, the BCDR dataset was chosen for training, validation, and test data for mass segmentation and classification in this study. Craniocaudal (CC) view mammograms that belong to a total of 368 patients were used from the BCDR for mass segmentation and classification.

2.2. Deep learning-based breast segmentation

As a pre-processing step, the mammograms were automatically flipped to the left side by measuring the average pixel intensity in the right and left sides. Mammograms belonging to seven patients that were judged to be too dark and noisy were excluded from the study. The pectoral muscle can sometimes appear in the CC view, so the breast segmentation was handled as segmenting both the breast tissue and the pectoral muscle together from the mammogram background. Automated breast segmentation was required and challenging because: 1) Characters in the mammogram background had different intensity levels, font types, font sizes, and were sometimes clipped; 2) The white rectangular strip noises on the sides had different pixel intensities and had sometimes sharp and sometimes soft edges. Also, white rectangular strip noises were sometimes on the breast tissue and sometimes on the background; 3) The mammograms were rarely partially clipped. The detection of all of the noise sources could not be accomplished by using image processing techniques, therefore a U-net model was used to segment the breast tissue from the mammogram background. The noise removal step was handled as a separate operation to enable focus on only the mass segmentation in the next step.

The mammograms in the BCDR dataset had dimensions ranging from 1167 × 545 to 1469 × 1039. All of the mammograms were resized to 1024 × 768 using bicubic interpolation which resulted in the dimension reduction with minimal information loss. Breast boundary annotations were performed manually by representing mammogram background as 0 and the breast tissue and pectoral muscle regions as 255. Mammogram masks were resized using nearest-neighbor interpolation to avoid creating pixel intensities other than 0 and 255. Mammograms and the corresponding masks were saved as 8-bit grayscale in portable network

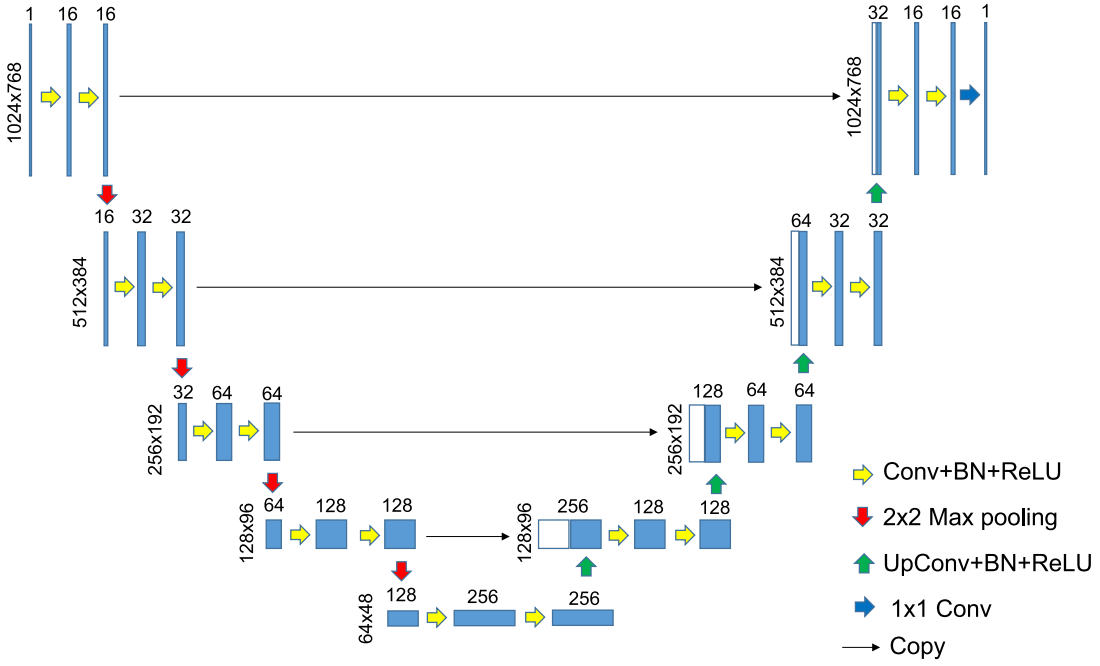


Fig. 1. Five layer U-net that was used for breast segmentation. Conv: Convolution, UpConv: Up-convolution, BN: batch normalization, ReLU: Rectified Linear Unit.

graphics (PNG) format. The ratio of breast tissue area in the cross-validation mammograms was 0.4821 ± 0.1602 (mean ± standard deviation). Breast tissue ground truth masks are accessible at https://github.com/tiryakiv/Mass-segmentation for interested researchers.

Mammograms were selected for the stratified five-fold cross-validation and test folders randomly based on the patient-level by keeping the same breast density distribution. 15% of all patients' mammograms were used for testing and the remaining mammograms were used for the cross-validation. The number of patients in the cross-validation folder are 15, 17, 26, and 4, with breast densities of type A, B, C, and D, respectively. The test folder had 14, 17, 23, and 4 mammograms with the same breast density order. The test mammograms were never used for training purposes. The total number of mammograms in the cross-validation and test folders are 352 and 60, respectively. The patient ids in the cross-validation and test folders are given in the Supplementary Materials section 1.

Deep learning models were implemented on a Dell T7610 workstation desktop computer with an NVIDIA GeForce RTX 3060 graphics processing unit (GPU), two Intel Xeon E5-2630 2.6 GHz CPUs, 16 GB RAM of system memory, and a 64-bit Windows 10 operating system. U-net model implementations were performed using Tensorflow 2.10 and Keras frameworks in the Python 3.9 environment [9,23,24]. Data augmentation was applied to input mammogram and label annotation images simultaneously by enabling vertical and horizontal flip, setting the rotation range to 45°, and setting the width shift range, height shift range, shear range, and zoom range to 10%. Image wrapping was used as a fill mode.

Early stopping criteria was used to avoid overfitting. Validation loss was monitored at each epoch and training was stopped when five consecutive losses did not decrease. Adam optimizer was used with an initial learning rate of 10^{-3} [25]. When three consecutive validation losses did not decrease, the learning rate was reduced by a factor of 0.1 while the minimum learning rate was 10^{-8}. Data generator was applied to efficiently use the system memory, and a batch size of two was used. Deep neural network weights were initialized by He normal method [26].

U-net breast segmentation models having two, three, four, and five encoding and decoding layers were implemented. Encoding layers have convolution, batch normalization [27], a ReLU activation function [28], and a max pooling layer. The number of filters is doubled at each down sampling step in the encoding layer. Decoding layers have upsampling, up-convolution, and concatenation. A dropout rate of 0.5 was used to avoid overfitting [29,30]. Dice loss was used to focus on the true positives [31]. The number of epochs per step was set to three times the training data divided by the batch size. In the final U-net layer, sigmoid activation function was used:

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (1)$$

where $\vec{z}$ is the input vector to the softmax, z_i are elements of the input vector, K is the number of classes. Different U-net models were implemented and trained by changing the number of network layers, and the model performances were compared to find the best performing one. The five-level U-net architecture is shown in Fig. 1.

2.3. Deep learning-based mass segmentation without transfer learning

Following the breast segmentation step, the breast tissue region was centered vertically and the remaining black regions were discarded. The mammograms were resized to 640 × 640 to use the memory efficiently while minimizing the information loss. Resizing from rectangular mammogram shape (1024 × 768) to square (640 × 640) slightly changed the aspect ratio since the black regions on the top and the bottom were discarded. The mammogram intensities were normalized in the range of [0, 255]. Mass annotations were already provided in the BCDR database and they were not changed. The same vertical centering and resizing steps were identically repeated for the mass masks so that the mammograms and the masks overlap. When there are multiple mass lesions, the masks were overlaid and all of the lesions were combined in a single mask. Four representative mammograms and their corresponding ground truth masks are given in Fig. 2.

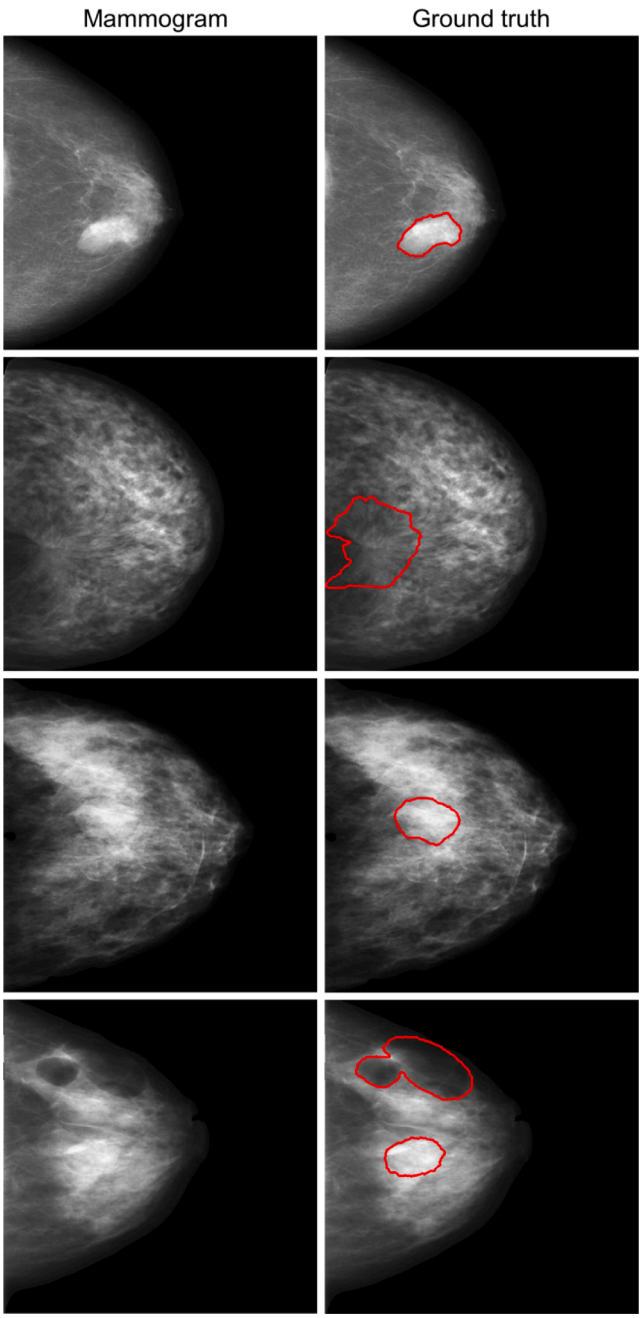


Fig. 2. BCDR film mammograms from the training data after the deep learning-based breast segmentation step are shown in the left column and the corresponding ground truth masks are shown in the right. Note the diversity of the mass densities in the mammograms. The mass lesion boundaries are shown in red. (For interpretation of the colors in the figure, the reader is referred to the web version of this article. Courtesy of MA Guevara and coauthors, Breast Cancer Digital Repository Consortium.).

The same cross-validation and test data were used for mass segmentation as in the previous deep learning-based breast segmentation. Data augmentation was applied to input and label annotation images simultaneously the same as breast segmentation settings except

increasing the rotation range to 90° and setting the width shift range, height shift range, shear range, and zoom range to 20%.

The ratio of mass pixels in the cross-validation mammograms is 0.0176 ± 0.0369 (mean ± standard deviation), which shows that the segmentation data is highly unbalanced. There is at least one mass lesion in each mammogram and the number of mass lesions in the cross-validation data is 1.0824 ± 0.3484 (mean ± standard deviation).

Five-layer U-net similar to the one shown in Fig. 1 was implemented and trained by modifying input size, and the model performance was evaluated on the cross-validation data. Recently, U-net variants having different connection types were proposed to improve the segmentation performance. U-net++ is a model where features were fused from variable scale by a new design of skip connections [10]. U-net++ was implemented with and without deep supervision by setting the number of input neurons to 8, and 12. The network with 12 input neurons with deep supervision yielded the highest DSC segmentation performance. Residual U-net was first proposed for road extraction from satellite images using rich skip connections [32]. Attention U-net (AU-net) has a dense up-sampling network and a channel attention function [33,34]. MultiResUnet is an improved version of U-net which has *MultiRes* block instead of convolution block and a *Res* path instead of skip connection [35]. Mass segmentation performances of residual U-net, AU-net, and MultiResUnet were investigated, using implementations of Nikhil Kumar Tomar [36] and Asma Baccouche [12,37].

2.4. Transfer learning for deep learning-based mass segmentation

Transfer learning is known to improve deep learning performance in breast cancer detection [1]. Transfer learning was implemented by importing the pre-trained network weights on ImageNet into the encoding path without the dense classification layer [38]. The five consecutive skip connections were identified by finding the last layer that has the same dimensions as the input. The decoding path consisted of five consecutive 2D transpose convolution, concatenation, followed by double convolution, batch normalization, and ReLU activation function [36,38].

Five variants of transfer learning U-net models were investigated for mass segmentation. U-VGG16 and U-ResNet50 has VGG16 [39] and ResNet50 [40] weights in the encoder path of U-net, respectively. The U-Xception model has Xception weights [15] in the encoder region and residual network in the decoder [41]. DeepLabV3Plus is a network that combines the semantic information in the encoder region and a decoder module that recovers object boundaries by Atrous convolutions [11]. The DeepLabV3Plus model has ResNet50 coefficients in the encoder. U-VGG16, U-ResNet50, and DeepLabV3Plus performances were investigated by using implementations of Nikhil Kumar Tomar [36]. Dice loss was used to focus on the mass region (true positive) [31]. Steps per epoch was set to three times the number of train steps and the batch size was two due to the memory limitation. The initial learning rate was 10^{-4}. Other settings were the same as mass segmentation models trained from scratch.

2.5. Transfer learning implementation of U-net++ Xception

U-net++Xception was implemented by using the Xception model weight coefficients in the encoder region. Skip connections of Xception model were determined as the 1st, 12th, 22nd, 32nd, and 122nd layers and these nodes were connected to the encoding path of the U-net++'s $X^{0,0}$, $X^{1,0}$, $X^{2,0}$, $X^{3,0}$, $X^{4,0}$ nodes, respectively [10]. Zero padding was applied to the input shape of the 12th and 22nd nodes to establish the same dimension during concatenation operation. The U-net++Xception network was trained by setting the initial learning rate to 10^{-4} and the number of filters to 8, 12, and 16. Fine tuning was applied by freezing the Xception weights and training the rest of the network with an initial learning rate of 10^{-2} followed by unfreezing all weights and training the whole network. The networks were trained with and without the deep-supervision. The best performing network was obtained when the number of filters was 12, the initial learning rate was 10^{-4}, and no supervision was used. Freezing the Xception network weights and training the rest of the weights did not improve the performance. The U-net++Xception model implementation is available at https://github.com/tiryakiv/Mass-segmentation.

2.6. Segmentation performance evaluation

The segmentation performances of model predictions were evaluated using precision, recall, accuracy, DSC [17], intersection over union (IoU) also known as Jaccard index or Jaccard similarity coefficient [42], Matthews correlation coefficient (MCC) [43], Cohen's kappa score (κ) [44,45], and classification success index (CSI) [46,47]:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$DSC = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{5}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{6}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{7}$$

$$\kappa = \frac{2 \times (TP \times TN - FP \times FN)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)} \tag{8}$$

$$CSI = Precision + Recall - 1 \tag{9}$$

where *TP* is true positive, *TN* is true negative, *FP* is false positive, and *FN* is false negative. For breast or mass tissue segmentation problem, tissue pixels that are correctly classified are *TP*, background or tissue other than mass that are correctly classified are *TN*, breast or mass tissue pixels that are incorrectly classified are *FN*, and background or tissue other than mass pixels that are incorrectly classified are *FP*. Area under the receiver operating curve (AUC) was used to evaluate the performances of mass segmentation and classification models. AUC was assumed to be zero when it was undefined for a mammogram due to misdetection of a mass lesion.

2.7. Mass classification

Mass classification was performed using U-net++Xception segmentation model predictions. Five-fold cross-validation was applied using the ground truth mass lesion grayscale 8-bit depth mammograms. The mass lesions in the cross-validation segmentation data were used for mass classification cross-validation data. When there are multiple mass lesions in a mammogram, they were separated and each mass lesion was set to the center of the 224 × 224 binary image. If the mass lesion is bigger than 224 × 224, it was resized to 224 × 224. The total number of benign masses was greater than the total number of malignant mass lesions. To have a balanced data set, 122 benign mass lesions were discarded randomly and the final mass classification cross-validation data included 114 benign and 114 malignant ground truth mass lesion patches. The mass segmentation model predictions were used in the test data. There are 60 mammograms in the test set and a total of 70 mass lesions since some of the mammograms has multiple lesions. The mass lesions that were detected with positive IoU were included for mass classification test. Ground truth lesion characterizations were taken

Table 1
Mean breast segmentation performances and total number of trainable parameters of U-nets with two. three, four, and five layers on the five-fold cross-validation data. Best result in each column is shown in bold. P: precision, R: recall, Acc: Accuracy, DSC: Dice's similarity coefficient, IoU: intersection over union, MCC: Matthews correlation coefficient, κ: Cohen's kappa, CSI: classification success index, #prm: total number of trainable parameters.

# U-net layers	P	R	Acc	DSC	IoU	MCC	κ	CSI	#prm
2	0.9734	0.9481	0.9655	0.9596	0.9237	0.9223	0.9203	0.9215	**6.55 K**
3	0.9783	0.9667	0.9760	0.9718	0.9461	0.9449	0.9436	0.9450	121.83 K
4	0.9767	**0.9741**	0.9781	0.9747	0.9515	0.9498	0.9485	0.9508	483.25 K
5	**0.9795**	0.9715	**0.9784**	**0.9748**	**0.9517**	**0.9501**	**0.9489**	**0.9510**	487.26 K

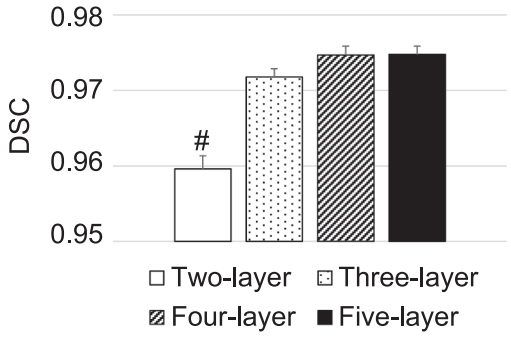


Fig. 3. Breast segmentation performances of two-, three-, four-, and five-layer U-nets in terms of DSC. Bars and error bars show the mean and SEM of n = 5 measurements. # denotes statistical significance from all other measurements in the graph. Lack of significance mark between measurements means non-significance. DSC: Dice's similarity coefficient.

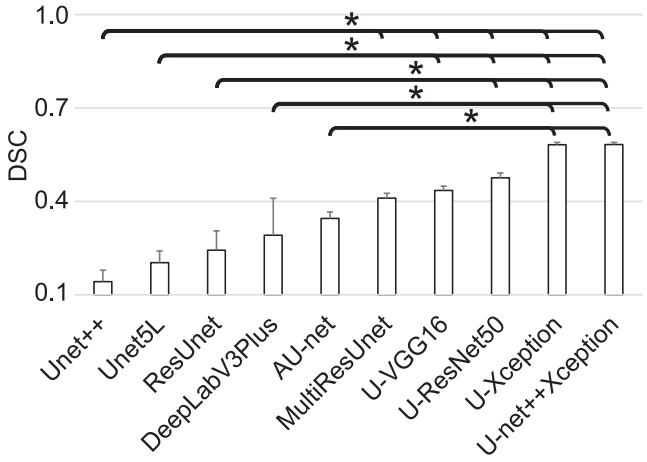


Fig. 4. Mass segmentation performances of U-net variants in the ascending order. Bars and error bars show the mean and SEM of n = 5 measurements. * denotes statistical significance between the two measurements. Lack of significance mark between measurements means nonsignificance. DSC: Dice's similarity coefficient.

from the BCDR where probably benign (P-benign) classes were assumed as benign, and probably malignant (P-malignant) classes were assumed as malignant. Mass classification was performed by deep transfer learning (DTL) from weight coefficients of ResNet50V2 [48], VGG16 [39], and InceptionV3 [49] as described in Shen *et al.* [1]. After the best model was determined on the cross-validation data, the test data was applied to the model and final performance was recorded. The classification performance of the DTL models were presented in terms of precision, recall, F1-score, accuracy, CSI, and AUC [50].

2.8. Statistical comparisons of model performances

Breast and mass segmentation and mass classification performances were presented as mean ± standard error of the mean (SEM). Segmentation performances were presented in DSC and mass classification performances were presented in accuracy. Variations in the performance of U-net variants for segmentation and DTL models were analyzed using one-way analysis of variance (ANOVA) followed by pairwise post-hoc comparisons with Tukey's highest significant difference test. Significance levels were set at $p < 0.05$. Statistical calculations were performed online at https://astatsa.com/OneWay_Anova_with_TukeyHSD/.

3. Results

3.1. Breast segmentation results

U-nets having two, three, four, and five layers were implemented using Dice loss to find the best performing breast segmentation model. Performance metrics for mammogram segmentation on five-fold cross-validation data is given in terms of mean precision, recall, accuracy, DSC, IoU, MCC, kappa, and CSI in Table 1. The segmentation performances increased as the number of layers were increased. Five-layer U-net trained from scratch outperformed other U-nets in terms of all metrics except recall. Therefore, five-layer U-net was used for breast segmentation. Since the segmentation performance was satisfactory, transfer learning options were not investigated for breast segmentation.

Table 2
Average mass segmentation performances of U-net variants on the validation data in the ascending DSC order. The best result in each column is shown in bold. P: precision, R: recall, Acc: Accuracy, DSC: Dice's similarity coefficient, IoU: intersection over union, MCC: Matthews correlation coefficient, AUC: area under the receiver operating curve, κ: Cohen's kappa, CSI: classification success index, #prm: total number of trainable parameters.

Model	P	R	Acc	DSC	IoU	MCC	AUC	κ	CSI	#prm
Unet++	0.3124	0.1183	0.9828	0.1421	0.0985	0.1630	0.4323	0.1391	−0.5693	**0.57 M**
Unet5L	0.3357	0.1932	0.9828	0.2033	0.1476	0.2243	0.4934	0.2009	−0.4711	1.94 M
ResUnet	0.3190	0.2625	0.9807	0.2431	0.1817	0.2563	0.5138	0.2383	−0.4185	8.22 M
DeepLab V3Plus	0.3390	0.3337	0.9826	0.2911	0.2356	0.3053	0.4738	0.2887	−0.3273	17.83 M
AU-net	0.3830	0.4269	0.9768	0.3453	0.2578	0.3622	0.7157	0.3384	−0.1901	11.02 M
MultiResUnet	0.4401	0.4907	0.9844	0.4107	0.3217	0.4293	0.6872	0.4056	−0.0692	7.24 M
U-VGG16	0.5287	0.4786	0.9855	0.4351	0.3469	0.4568	0.7510	0.4309	0.0073	25.86 M
U-ResNet50	0.5182	0.5594	0.9819	0.4757	0.3865	0.4949	0.7247	0.4710	0.0776	20.64 M
U-Xception	0.5869	**0.6576**	**0.9886**	0.5819	0.4858	**0.5960**	**0.7833**	0.5780	**0.2445**	23.10 M
U-net++ Xception	**0.5996**	0.6417	0.9881	**0.5826**	**0.4877**	0.5945	0.7641	**0.5785**	0.2413	17.18 M

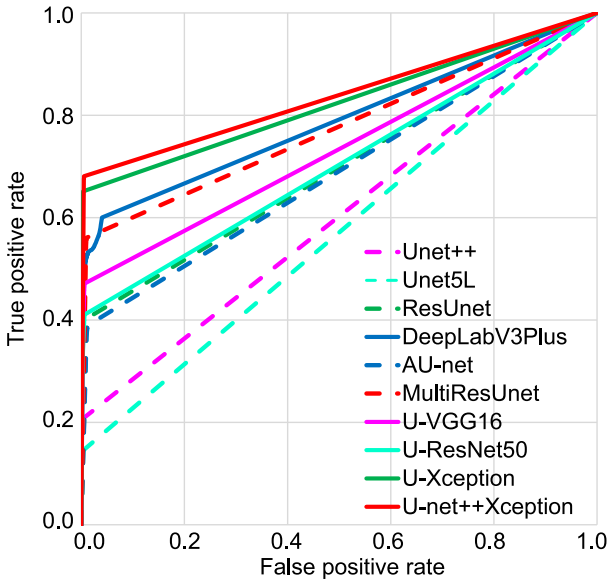


Fig. 5. Mass segmentation ROC of U-net variants on the validation data. Transfer learning U-net models are shown by solid lines and other models are shown by dashed lines. (For interpretation of the colors in the figure, the reader is referred to the web version of this article.).

Statistical comparisons of U-nets with two, three, four, and five layers are shown in Fig. 3. Statistical analysis of breast segmentation performance demonstrated that there were significant differences among U-nets ($p = 7.8498 \times 10^{-7}$). Fig. 3 shows that the five-layer U-net had the best segmentation performance. Post hoc pairwise comparisons demonstrated that the performance difference between two-layer U-net and three-, four-, five-layer U-nets were statistically significant ($p = 0.0010$). Four-layer U-net performed better than three-layer U-net but the performance difference was not statistically significant ($p = 0.4259$).

3.2. Mass segmentation results

A total of ten different variants of U-nets having different types of connections and transfer learning methods were investigated for mass segmentation. Performance evaluations of U-nets for mass segmentation on the five-fold cross-validation data are given in Table 2. The threshold levels were set 127 for each model.

Statistical comparisons of the performances of U-net variants are shown in Fig. 4. Statistical analysis of mass segmentation performance demonstrated that there were significant differences among U-nets ($p = 3.5592 \times 10^{-8}$). Fig. 4 shows that U-net++Xception had the highest segmentation performance. Post-hoc pairwise comparisons demonstrated that the performance difference between U-net++Xception or U-Xception and any element of {Unet++, Unet5L, ResUnet, DeepLabV3Plus, AU-net} were statistically significant. The U-net++Xception performance was slightly better than U-Xception but the performance difference was not statistically significant ($p = 0.9000$). p values for all of the comparisons are given in the Supplementary Materials section 2.

An ROC analysis was conducted to compare the performances of mass segmentation models and the results are given in Fig. 5. ROC was applied to the validation folder that is the closest to the mean fold performance. Among all models, U-Xception and the proposed U-net++Xception models have better mass segmentation performances than other models in terms of AUC. In general, transfer learning methods performed better than the models trained from scratch.

Mass segmentation quantitative and qualitative results on the validation data using U-net++Xception model is shown in Fig. 6. The results show that the some of the mass regions with low density caused false

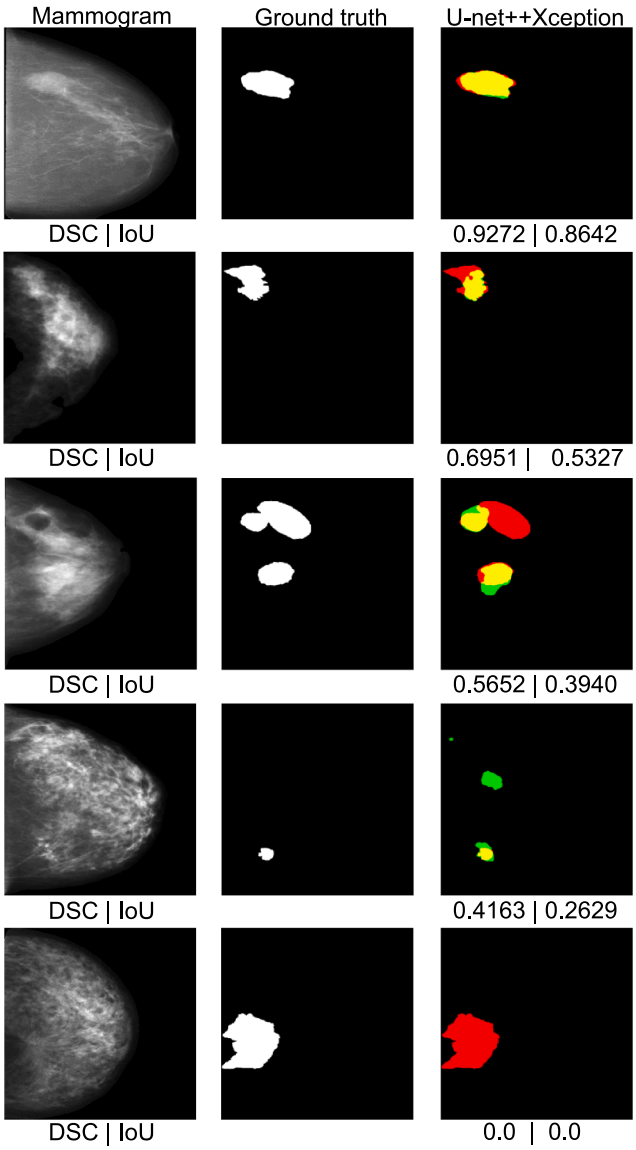


Fig. 6. Quantitative and qualitative mass segmentation results on the validation data. Left column: Mammograms after breast segmentation; middle column: ground truth binary mask; right column: U-net++Xception model output where yellow represents true positive, black represents true negative, red represents false negative, and green represents false positive. The top and bottom rows are examples for almost the best and the worst segmentation predictions respectively. DSC: Dice's similarity coefficient; IoU: intersection over union. (For interpretation of the colors in the figure, the reader is referred to the web version of this article. Courtesy of MA Guevara and coauthors, Breast Cancer Digital Repository Consortium).

Table 3
Confusion matrix analysis of mass segmentation on the validation data. The values represent the mean of 69 images and rows are normalized.

Radiologist	Mass region	0.6783	0.3217
	Mammogram other than mass	0.0052	0.9948
		Mass region	Mammogram other than mass
		U-net++Xception output	

Table 4
Comparison of mass segmentation test performance in the literature and this study by various performance metrics and datasets. Acc: Accuracy, DSC: Dice's similarity coefficient, IoU: intersection over union, κ: Cohen's kappa, CSI: classification success index, FFDM: full-field digital mammogram, FM: film mammogram, DETR: deformable transformer.

Publication	Method	Test dataset	Recall	Acc	DSC	IoU	AUC	κ	CSI
Abdelhafiz et al. [51]	Residual U-net	INbreast	–	0.9870	0.9830	0.9480	–	–	–
Baccouche et al. [10]	Connected-ResUnets	CBIS-DDSM	–	0.8691	0.8952	0.8002	0.83	–	–
Baccouche et al. [10]	Connected-ResUnets	INbreast	–	0.9303	0.9528	0.9103	0.96	–	–
Garrucho et al. [14]	DETR	BCDR FFDM	0.873	–	–	–	0.80	–	–
This study	U-net++ Xception	BCDR FM	0.6648	0.9918	0.6356	0.5408	0.7829	0.6326	0.4401

Table 5
The dependence of U-net++Xception model mass segmentation test performance on the breast density.

Breast density	# Mammograms	IoU (mean ± std)
Type A	15	0.5019 ± 0.3341
Type B	17	0.6141 ± 0.3319
Type C	24	0.4768 ± 0.2989
Type D	4	0.7598 ± 0.2566

negative errors and some fibroglandular tissue regions caused false positive errors. To further analyze the percentage of false negatives and false positives, the confusion matrix is given in Table 3.

Because of its best performance on five-fold cross-validation data among other choices, the U-net++Xception model was used for mass segmentation on the test data. A comparison between the studies in the literature and the present study on the test data is given in Table 4.

To investigate the applicability of the proposed deep learning pipeline on another dataset, the mass segmentation performance of the proposed two-stage DTL based segmentation pipeline was tested on the whole 170 CBIS-DDSM test mammograms directly without any training. The segmentation performance in terms of accuracy, DSC, and IoU are 0.9891, 0.3300, and 0.2660, respectively. The percentage of detected masses were %50.51.

The mass segmentation performance of the U-net++Xception model on test mammograms with different breast densities is shown in Table 5. The highest and lowest mass segmentation IoU was obtained when the breast density is Type D and C respectively.

3.3. Mass classification results

The U-net++Xception mass segmentation model was used to segment the masses in the test mammograms. Mass segmentation results on test mammograms were further analyzed for classification of the mass lesions as benign or malignant. There are 60 mammograms in the test set and a total of 70 mass lesions. The U-net++Xception segmentation model was able to detect 54 of the 70 lesions (IoU > 0), making a detection rate of 0.7714 at 0.5167 false positive per mammogram. 12 of the lesions had no ground truth lesion characterization in the BCDR. Mass classification test performance results including 42 mammograms (25 benign and 17 malignant mass) are shown in Table 6.

The table 6 results showed that the best performing mass classification model was VGG16. The classification performance statistical analysis was performed on the accuracies obtained on the five-fold cross-validation data and the results are shown in Fig. 7. There was no statistical significance in performance comparison between the three DTL models (min $p = 0.5107$). Since VGG16 had the best performance, the test results were obtained by that model.

4. Discussion

The present study shows that U-net++Xception-based cascaded segmentation followed by DTL-based mass classification can be used for automated breast cancer lesion characterization. The reported results form a baseline for mass segmentation and classification performance for the BCDR dataset. The information loss during pre-processing was kept minimal by setting the mass segmentation mammogram resolution to 640 × 640. The proposed framework demonstrates the performance of the U-net++ rich skip connection architecture with the Xception transfer learning for mass segmentation on the unseen test data. The proposed deep learning mass segmentation model is able to discriminate the hidden texture pattern in the mass tissue from the surrounding tissue. The deep learning classification model is also able to classify the malignant versus benign mass tissue texture at the reported performances. Collectively, these results show that the proposed cascaded DTL pipeline would be beneficial to radiomics for breast cancer.

The proposed U-net++Xception method outperformed U-net, U-net++, and U-Xception methods on the mass segmentation problem. The U-net++Xception was able to segment the mass texture from the

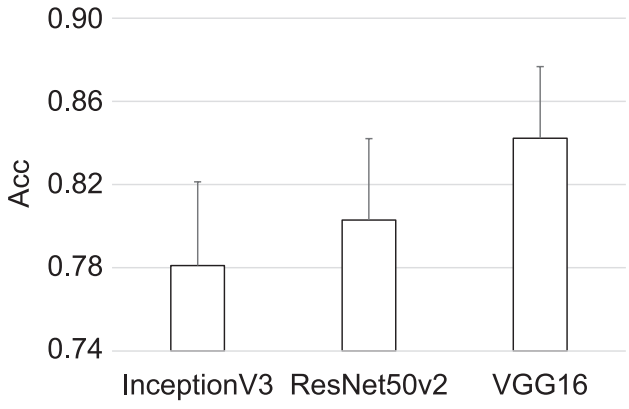


Fig. 7. Mass classification performances of deep transfer learning methods on the validation data in the ascending order. Bars and error bars show the mean and SEM of n = 5 measurements. There is no statistical significance between accuracy measurements. Acc: accuracy.

Table 6
Mean mass classification performances of DTL models on five-fold cross-validation and test data. Best result in each column is shown in bold. P: precision, R: recall, Acc: Accuracy, AUC: area under the receiver operating curve, CSI: classification success index.

Data	Classifier	P	R	Acc	F1-score	AUC	CSI
Cross-validation	ResNet50V2	0.8529	0.7549	0.8030	0.7921	0.8851	0.6078
Cross-validation	VGG16	**0.8659**	**0.8079**	**0.8423**	**0.8337**	**0.8979**	**0.6738**
Cross-validation	InceptionV3	0.7953	0.7640	0.7810	0.7776	0.8845	0.5593
Test	VGG16	0.8182	0.5294	0.7619	0.6429	0.8188	0.3476

surrounding breast tissue texture with a detection rate of 0.7714 and DSC of 0.6356 on the unseen test data. The U-net++Xception model had the least number of trainable parameters among all of the transfer learning-based segmentation models in the present study. Thus, U-net++Xception has higher segmentation performance and less computational cost than other transfer learning methods.

Mass segmentation from mammograms is challenging due to the high pixel resolution of mammograms, low percentage of mass region (mean = 1.76% in the present study), and similar appearance of fibroglandular and mass tissues in the mammograms. In the present study, the noise sources in the BCDR mammograms had large diversities so the breast segmentation was performed separately in the first step. The mammogram noise removal enabled to focus only on the mass segmentation in the second step. The best mass segmentation performance in terms of DSC was obtained by using the proposed U-net++Xception. U-Xception yielded the second highest mass segmentation performance. The two high performances show the capability of both U-net architecture and the depthwise and pointwise convolutions of the Xception network. The details of successful transfer learning from Xception network to the U-net++ architecture for mass segmentation is demonstrated. The segmentation predictions were classified with a classification accuracy of 0.7619 on the test mammograms without using clinical data. These results show the potential of the proposed model as an independent evaluation system. The segmentation and classification performances of the U-net variants were evaluated by CSI, which is useful when the data is unbalanced such as the present study. Finally, the challenge of segmenting masses with low and equal densities was presented and solved with the DTL-based pipeline.

The mass segmentation and detection performances of the proposed U-net++Xception model on CBIS-DDSM test mammograms without any training show the potential of the proposed method for breast cancer radiomics between different datasets. Breast segmentation results on CBIS-DDSM were good but since the dataset does not have ground truth, no quantitative segmentation performance was presented. The mass segmentation test performance results on the CBIS-DDSM dataset was acceptable even though the U-net++Xception was not trained on the CBIS-DDSM database, which is a collection of different scanners, resolution, and bit-depths than the ones of BCDR. Additional training is indicated when training a model with a dataset and testing the model on another dataset. A universal mammogram dataset that has a diverse set of scanners would be beneficial, which may be accomplished by the release of new datasets resulting from big collaborations in the future.

Mass segmentation has attracted numerous researchers in the past. The interest has recently increased because of the high performance of deep learning models on the image classification and segmentation tasks. Abdelhafiz *et al.* investigated residual attention U-nets for mass segmentation and classification [51]. They trained their model on CBIS-DDSM and BCDR-D01, and tested on INbreast dataset. The majority of the training data comes from the CBIS-DDSM (~92%). While the method and the data is useful, the focus in the present study is on the annotations that were directly made by radiologists.

The mass segmentation performances obtained in the present study are lower than the ones in the literature by other datasets. The relatively low performance can be explained by: 1) The model performances in the present study were calculated by comparing directly to the radiologists' annotations. CBIS-DDSM dataset ground truth masks were created by an algorithm, and the concordance levels are within [0.749–0.904] range; 2) The INbreast dataset includes digital mammograms, which are known to yield better results than screen film mammography in breast cancer detection [52]; 3) The mass density diversity of the BCDR dataset is higher (see Fig. 2) than the diversity of the CBIS-DDSM train dataset since the dataset does not include any low-density masses. The INbreast dataset has only one low-density mass lesion, while BCDR cross-validation dataset of the present study have a total of fifteen low-density masses. 4) Recent literature findings about relatively low performance on the BCDR dataset are consistent with the results of this study [14].

The mass segmentation performance dependence on the breast density is almost as expected since the lowest segmentation was observed on Type C mammograms. The high segmentation performance on Type D can be due to the small number of mammograms in this breast density.

Mass segmentation has been achieved through a cascaded deep learning model. The limitations of the present study are: 1) A cascaded approach increases the requirement for both computational resources and the amount of training data; 2) BCDR and other datasets lack performance comparison between different radiologists so there is an uncertainty for setting an accomplishment level for the deep learning model; 3) The obtained performance results are valid for BCDR dataset. Application of the model on other datasets may require additional training.

Mass classification performance on the cross-validation data was higher than the test data (see Table 6). This is not due to the overfitting since the classification test data was the prediction of U-net++Xception segmentation model. The segmentation model lost some of the mass region information (IoU = 0.5408) and included some normal breast tissue that possibly caused a reduction in the classification performance.

There would be performance gains if more training data was available, so there is a need for bigger mammography datasets. Performance increase may be obtained if full field digital mammograms are used. The mass segmentation performances of the present study shows that there is still a need for performance improvement in the deep learning-based segmentation models. However, mass segmentation and classification system without using clinical data in the present study shows a successful application of deep learning in the computer-aided breast cancer research.

5. Conclusion

Breast cancer diagnosis can be done from film mammograms by using the proposed cascaded U-net++Xception deep learning pipeline without clinical data. The proposed model may be used to reduce the workload of radiologists for mass detection, segmentation, and classification which are all crucial mammogram interpretation steps. The proposed model is useful for breast cancer mass segmentation and classification on BCDR dataset and may be useful on additional mammogram datasets.

Funding Information

This work was supported by Siirt University Scientific Research Projects Directorate Grant No. 2021-SİÜMÜH-01 (VMT).

CRediT authorship contribution statement

Volkan Müjdat Tiryaki: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.

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.

Data availability

I do not have permission to share the raw mammogram data, but the mammogram dataset is accessible from https://bcdr.eu/. The codes are accessible at https://github.com/tiryakiv/Mass-segmentation

Acknowledgement and/or disclaimers, if any

The author thanks the Siirt University Scientific Research Projects Directorate for providing the NVIDIA GeForce RTX 3060 graphics processing unit under Grant No. 2021-SİÜMÜH-01. The author thanks MA Guevara and coauthors [16] for making the Breast Cancer Digital Repository available online. The author thanks Prof. Virginia M. Ayres and Dr. Veysel Kaplanoğlu for technical discussions. The author thanks the Google LLC for Chrome Remote Desktop and Google Drive applications. The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bspc.2023.104819.

References

[1] L. Shen, L.R. Margoiles, J.H. Rothstein, E. Fluder, R. McBride, W. Sieh, Deep Learning to improve Breast cancer Detection on Screening Mammography, Sci. Rep. 9 (2019) 1–12.
[2] Y. Kaya, A new intelligent classifier for breast cancer diagnosis based on a rough set and extreme learning machine: Rs + elm, Turkish J. Electr. Eng. Comput. Sci. 21 (2013) 2079–2091, https://doi.org/10.3906/elk-1203-119.
[3] Z. Assari, A. Mahloojifar, N. Ahmadinejad, Discrimination of benign and malignant solid breast masses using deep residual learning-based bimodal computer-aided diagnosis system, Biomed. Signal Process. Control. 73 (2022), 103453, https://doi.org/10.1016/j.bspc.2021.103453.
[4] P.K. Chaudhary, R.B. Pachori, Automatic diagnosis of glaucoma using two-dimensional Fourier-Bessel series expansion based empirical wavelet transform, Biomed. Signal Process. Control. 64 (2021), https://doi.org/10.1016/J.BSPC.2020.102237.
[5] P.K. Chaudhary, R.B. Pachori, Automatic Diagnosis of Different Grades of Diabetic Retinopathy and Diabetic Macular Edema Using 2-D-FBSE-FAWT, IEEE Trans. Instrum. Meas. 71 (2022), https://doi.org/10.1109/TIM.2022.3140437.
[6] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzebski, T. Fevry, J. Katsnelson, E. Kim, S. Wolfson, U. Parikh, S. Gaddam, L.L.Y. Lin, K. Ho, J. D. Weinstein, B. Reig, Y. Gao, H. Toth, K. Pysarenko, A. Lewin, J. Lee, K. Airola, E. Mema, S. Chung, E. Hwang, N. Samreen, S.G. Kim, L. Heacock, L. Moy, K. Cho, K.J. Geras, Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening, IEEE Trans. Med. Imaging. 39 (2020) 1184–1194, https://doi.org/10.1109/TMI.2019.2945514.
[7] H. Zonderland, R. Smithuis, Bi-RADS for Mammography and Ultrasound 2013 (2014) 1–45. https://radiologyassistant.nl/breast/bi-rads/bi-rads-for-mammography-and-ultrasound-2013.
[8] R.S. Lee, F. Gimenez, A. Hoogi, K.K. Miyake, M. Gorovoy, D.L. Rubin, A curated mammography data set for use in computer-aided detection and diagnosis research, Sci. Data. 4 (2017) 1–9, https://doi.org/10.1038/sdata.2017.177.
[9] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 9351 (2015) 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.
[10] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation, IEEE Trans. Med. Imaging. 39 (2020) 1856–1867, https://doi.org/10.1109/TMI.2019.2959609.
[11] L.C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 11211 LNCS (2018) 833–851. https://doi.org/10.1007/978-3-030-01234-2_49.
[12] A. Baccouche, B. Garcia-Zapirain, C. Castillo Olea, A.S. Elmaghraby, Connected-UNets: a deep learning architecture for breast mass segmentation, Npj, Breast Cancer 7 (2021) 1–12, https://doi.org/10.1038/s41523-021-00358-x.
[13] N.K. Tomar, A. Shergill, B. Rieders, U. Bagci, D. Jha, TransResU-Net: Transformer based ResU-Net for Real-Time Colonoscopy Polyp Segmentation, ArXiv. (2022) 1–4. http://arxiv.org/abs/2206.08985.
[14] L. Garrucho, K. Kushibar, S. Jouide, O. Diaz, L. Igual, K. Lekadir, Domain generalization in deep learning based mass detection in mammography: A large-scale multi-center study, Artif. Intell. Med. 132 (2022), 102386, https://doi.org/10.1016/j.artmed.2022.102386.
[15] F. Chollet, Xception: Deep learning with depthwise separable convolutions, Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017. 2017-Janua (2017) 1800–1807. https://doi.org/10.1109/CVPR.2017.195.
[16] M.A.G. López, N.G. de Posada, D.C. Moura, R.R. Pollán, J.M.F. Valiente, C.S. Ortega, M.R. del Solar, G.D. Herrero, I.M.A.P. Ramos, J.P. Loureiro, T.C. Fernandes, B.M.F. de Araújo, BCDR : A BREAST CANCER DIGITAL REPOSITORY, in: 15th Int. Conf. Exp. Mech., Porto/Portugal, 2012: pp. 1–5.
[17] L.R. Dice, Measures of the Amount of Ecologic Association Between Species, Ecology 26 (1945) 297–302, https://doi.org/10.2307/1932409.
[18] I.C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M.J. Cardoso, J.S. Cardoso, INbreast: Toward a Full-field Digital Mammographic Database, Acad. Radiol. 19 (2012) 236–248, https://doi.org/10.1016/j.acra.2011.09.014.
[19] BCDR - Breast Cancer Digital Repository, (2012). http://bcdr.inegi.up.pt.
[20] D.C. Moura, M.A.G. López, P. Cunha, N.G. de Posada, R.R. Pollan, I. Ramos, J.P. Loureiro, I.C. Moreira, B.M.F. de Araújo, T.C. Fernandes, Benchmarking Datasets for Breast Cancer Computer-Aided Diagnosis (CADx), in: J. Ruiz-Shulcloper, G. di Baja (Eds.), Prog. Pattern Recognition, Image Anal. Comput. Vision, Appl., Springer Berlin Heidelberg, Berlin, Heidelberg, 2013: pp. 326–333.
[21] D.C. Moura, M.A.G. Lopez, An evaluation of image descriptors combined with clinical data for breast cancer diagnosis, Int. J. Comput. Assist. Radiol. Surg. 8 (2013) 561–574, https://doi.org/10.1007/s11548-013-0838-2.
[22] R. Ramos-Pollán, M.A. Guevara-López, C. Suárez-Ortega, G. Díaz-Herrero, J. M. Franco-Valiente, M. Rubio-del-Solar, N. González-de-Posada, M.A.P. Vaz, J. Loureiro, I. Ramos, Discovering Mammography-based Machine Learning Classifiers for Breast Cancer Diagnosis, J. Med. Syst. 36 (2012) 2259–2269, https://doi.org/10.1007/s10916-011-9693-2.
[23] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, in: Proc. 12th USENIX Conf. Oper. Syst. Des. Implement., USENIX Association, Savannah, GA, USA, 2016: pp. 265–283. http://arxiv.org/abs/1603.04467.
[24] F. and others Chollet, Keras, GitHub. (2015). https://keras.io (accessed October 30, 2021).
[25] D.P. Kingma, J.L. Ba, Adam: A method for stochastic optimization, 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc. (2015) 1–15.
[26] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in, IEEE Int. Conf. Comput. Vis. 2015 (2015) 1026–1034, https://doi.org/10.1109/ICCV.2015.123.
[27] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, 32nd Int. Conf. Mach. Learn. ICML. 1 (2015) 448–456.
[28] A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier nonlinearities improve neural network acoustic models, Proc. 30th Int. Conf. Mach. Learn. 30 (2013).
[29] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Salakhutdinov, R. Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfittin, J. Mach. Learn. Res. 15 (2014) 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html.
[30] Github. (2019). https://github.com/HZCTony/U-net-with-multiple-classification.
[31] C.H. Sudre, W. Li, T. Vercauteren, S. Ourselin, M. Jorge Cardoso, Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, Lect. Notes Comput. Sci. (2017) 240–248. https://doi.org/10.1007/978-3-319-67558-9_28.
[32] Z. Zhang, Q. Liu, Y. Wang, Road Extraction by Deep Residual U-Net, IEEE Geosci. Remote Sens. Lett. 15 (2018) 749–753, https://doi.org/10.1109/LGRS.2018.2802944.
[33] H. Sun, C. Li, B. Liu, Z. Liu, M. Wang, H. Zheng, D.D. Feng, S. Wang, {AUNet}: attention-guided dense-upsampling networks for breast mass segmentation in whole mammograms, Phys. Med. {\&} Biol. 65 (2020) 55005. https://doi.org/10.1088/1361-6560/ab5745.
[34] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, B. Glocker, D. Rueckert, Attention U-Net: Learning Where to Look for the Pancreas, in: 1st Conf. Med. Imaging with Deep Learn. (MIDL 2018), Amsterdam, 2018. http://arxiv.org/abs/1804.03999.
[35] N. Ibtehaz, M.S. Rahman, MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation, Neural Netw. 121 (2020) 74–87, https://doi.org/10.1016/j.neunet.2019.08.025.
[36] N. Tomar, Semantic-Segmentation-Architecture/TensorFlow/, Github. (2022). https://github.com/nikhilroxtomar/Semantic-Segmentation-Architecture/tree/main/TensorFlow (accessed February 1, 2022).
[37] S. Jadon, A survey of loss functions for semantic segmentation, in, IEEE Conf. Comput. Intell. Bioinforma. Comput. Biol. CIBCB 2020, Via del Mar, Chile 2020 (2020) 1–7, https://doi.org/10.1109/CIBCB48159.2020.9277638.
[38] J. Deng, R. Wei Dong, L.-J. Socher, K. Li, L.i. Fei-Fei, ImageNet: A large-scale hierarchical image database, in, IEEE Conf. Comput. Vis. Pattern Recognit, IEEE, Miami, FL, USA 2009 (2009) 248–255, https://doi.org/10.1109/cvprw.2009.5206848.
[39] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., 2015: pp. 1–14.
[40] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: pp. 770–778.
[41] Siddhartha, Unet Xception Keras for Pneumothorax Segmentation, Kaggle. (2019). https://www.kaggle.com/meaninglesslives/unet-xception-keras-for-pneumothorax-segmentation.
[42] P. Jaccard, The distribution of the flora in the Alpine zone, New Phytol. X I (1912) 37–50, https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

[43] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta - Protein Struct. 405 (1975) 442–451, https://doi.org/10.1016/0005-2795(75)90109-9.
[44] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1960) 37–46, https://doi.org/10.1177/001316446002000104.
[45] M.J. Warrens, On the Equivalence of Cohen's Kappa and the Hubert-Arabie Adjusted Rand Index, J. Classif. 25 (2008) 177–183, https://doi.org/10.1007/S00357-008-9023-7.
[46] S. Koukoulas, G.A. Blackburn, Introducing New Indices for Accuracy Evaluation of Classified Images Representing Semi-Nat ural Woodland Environments, Photogramm. Eng. Remote Sens. 67 (2001) 499–510.
[47] V. Labatut, H. Cherifi, Accuracy Measures for the Comparison of Classifiers, ArXiv. (2012). https://doi.org/10.48550/ARXIV.1207.3790.
[48] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 9908 LNCS (2016) 630–645. https://doi.org/10.1007/978-3-319-46493-0_38.
[49] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the Inception Architecture for Computer Vision, Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-Decem (2016) 2818–2826. https://doi.org/10.1109/CVPR.2016.308.
[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[51] D. Abdelhafiz, S. Nabavi, R. Ammar, C. Yang, J. Bi, Residual deep learning system for mass segmentation and classification in mammography, in: 10th ACM Int. Conf. Bioinformatics, Comput. Biol. Heal. Informatics (ACM-BCB '19), Niagara Falls, NY, 2019: pp. 475–484. https://doi.org/10.1145/3307339.3342157.
[52] A.M.J. Bluekens, R. Holland, N. Karssemeijer, M.J.M. Broeders, G.J. Den Heeten, Comparison of digital screening mammography and screen-film mammography in the early detection of clinically relevant cancers: A multicenter study, Radiology 265 (2012) 707–714, https://doi.org/10.1148/radiol.12111461.