IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 19, 2021, accepted November 9, 2021, date of publication November 12, 2021,
date of current version November 18, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3127862

Comparison of Current Deep Convolutional Neural Networks for the Segmentation of Breast Masses in Mammograms

ANDRÉS ANAYA-ISAZA[1,2], LEONEL MERA-JIMÉNEZ[2,3], JOHAN MANUEL CABRERA-CHAVARRO[2,4], LORENA GUACHI-GUACHI[5], DIEGO PELUFFO-ORDÓÑEZ[6], AND JORGE IVAN RIOS-PATIÑO[7]

[1]Faculty of Engineering, Pontificia Universidad Javeriana, Bogota 410001, Colombia
[2]INDIGO Research, Bogota 410010, Colombia
[3]Faculty of Engineering, Universidad de Antioquia, Medellin 050010, Colombia
[4]Faculty of Engineering, Universidad Surcolombiana, Neiva 410010, Colombia
[5]Department of Mechatronics, Universidad Internacional del Ecuador, Quito 170411, Ecuador
[6]Modeling, Simulation and Data Analysis (MSDA) Research Program, Mohammed VI Polytechnic University, Ben Guerir 47963, Morocco
[7]Faculty of Engineering, Universidad Tecnológica de Pereíra, Pereira 666003, Colombia

Corresponding author: Leonel Mera-Jiménez (leonel.mera@udea.edu.co)

This work was supported by the Research Division of INDIGO Technologies (https://indigo.tech/).

ABSTRACT Breast cancer causes approximately 684,996 deaths worldwide, making it the leading cause of female cancer mortality. However, these figures can be reduced with early diagnosis through mammographic imaging, allowing for the timely and effective treatment of this disease. To establish the best tools for contributing to the automatic diagnosis of breast cancer, different deep learning (DL) architectures were compared in terms of breast lesion segmentation, lesion type classification, and degree of suspicion of malignancy tests. The tasks were completed with state-of-the-art architectures and backbones. Initially, during segmentation, the base UNet, Visual Geometry Group 19 (VGG19), InceptionResNetV2, EfficientNet, MobileNetv2, ResNet, ResNeXt, MultiResUNet, linkNet-VGG19, DenseNet, SEResNet and SeResNeXt architectures were compared, where "Res" denotes a residual network. In addition, training was performed with 5 of the most advanced loss functions and validated by the Dice coefficient, sensitivity, and specificity. The proposed models achieved Dice values above 90%, with the EfficientNet architecture achieving 94.75% and 99% accuracy on the two tasks. Subsequently, classification was addressed with the ResNet50V2, VGG19, InceptionResNetV2, DenseNet121, InceptionV3, Xception and EfficientNetB7 networks. The proposed models achieved 96.97% and 97.73% accuracy through the VGG19 and ResNet50V2 networks on the lesion classification and degree of suspicion tasks, respectively. All three tasks were addressed with open-access databases, including the Digital Database for Screening Mammography (DDSM), the Mammographic Image Analysis Society (MIAS) database, and INbreast.

INDEX TERMS Artificial intelligence, biomedical imaging, cancer, image segmentation, machine learning, mammography, medical diagnostic imaging.

I. INTRODUCTION

Breast cancer is a type of malignant tumor with the highest global incidence rate, accounting for approximately 10.4% of cancers [1]. It manifests as an excessive, disorganized, and invasive growth of breast cells [2]. The affected cells can spread through the blood or lymphatic system, generating new tumors and affecting other vital organs [2]. Currently, breast cancer is the leading cause of death among women between the ages of 20 and 50 years, and according to 2019 figures from the American Cancer Society, it estimated that there were approximately 268,600 new cases of invasive breast cancer, 48,100 cases of ductal carcinoma in situ (DCIS), and 41,740 deaths in the United States alone [1], [3]. By 2020, the figure reached a total of 684,996 deaths worldwide, making it the leading cause of female cancer mortality [4], [5]. Furthermore, at the beginning of 2021, the World Health Organization (WHO) reported this disease as

The associate editor coordinating the review of this manuscript and approving it for publication was Xiahai Zhuang.

the most common cancer worldwide, surpassing lung cancer [6]. Currently, these statistics continue to grow, and an increase of 50% is estimated over the next two decades as a consequence of increased life expectancy, unhealthy diets, insufficient physical activity, and the consumption of harmful substances such as alcohol [7]. This demonstrates the need for research at all stages related to breast cancer, from prevention to timely diagnosis and treatment [7], [8].

Breast cancer typically manifests as a mass or lump sensation, which can be detected by breast self-examination [9]. However, not all lumps are synonymous with cancer, i.e., there are benign and malignant lumps.

Various studies suggest that the incidence rates in low-income countries are lower than those in high-income countries. However, in the latter group, the mortality rate is lower, and the incidence rate (despite being higher) has been decreasing, while in the former, it has been progressively increasing [10]. These trends may be due to risk factors inherent to the socioeconomic positions of these countries, where one of the highest risk factors is the lack of access to early breast cancer detection [11]. In addition, this may be accompanied by other factors, such as age, ethnicity, breast characteristics, reproductive patterns, hormonal and environmental factors, and alcohol and tobacco consumption [12]. However, the probability of survival depends mainly on the stage and subtype of breast cancer. Detection at early stages can reduce the mortality rate from 40% to 15% [13], so it is vital to develop systems for the early and accurate detection of breast cancer.

There are many tools for diagnostic assistance in different areas of medicine [14]. Breast cancer is no exception to this rule, where technological evolution has allowed the integration of complex tools such as mammography, magnetic resonance (MR) imaging, positron emission tomography (PET), computed tomography (CT), and single-photon emission computed tomography (SPECT) [15]. These techniques have made it possible to evaluate and detect breast cancer with a high percentage of accuracy. However, the costs of these pieces of equipment prevent them from being integrated into health systems, especially in low-income countries or regions with difficult access [16], [17]. In this sense, conventional mammography is usually the most economical and viable solution. In addition, it is one of the most efficient tools for early breast cancer diagnosis [18]. In such mammographic images, benign masses appear as regular shapes, while those with irregular borders are usually malignant [19]. Furthermore, research has shown that annual mammograms can help detect abnormalities even before the patient or physician can perceive a significant change [20]. Consequently, mammography plays a primary role in the early detection of breast cancer, increasing the likelihood of curing the disease and the success of breast-conserving surgeries [20], [21]. In fact, this examination's effectiveness can decrease mortality from 40% 20% and increase the 5-year relative survival rate to 99% in screened women [22]–[24]. In this sense, mammography's potential has encouraged multiple applications ranging from interpretation, analysis, and visualization approaches for medical data [25]. Moreover, state-of-the-art artificial intelligence techniques have enabled the integration of complex tasks such as detection, segmentation, and classification with speeds that exceed human performance [26], [27]. However, much work is needed to develop and refine these systems, particularly in mammography cases, where the structure of the breast is quite complex. Additionally, traditional medical segmentation techniques are time-consuming and knowledge-intensive processes that can lead to errors or subjective diagnoses [28].

As mentioned above, lesions or masses are the main signals utilized for breast cancer diagnosis [8]. The boundary information in the affected regions reflects the growth pattern and biological characteristics of the disease [29]. Therefore, poor masses segmentation limits the classification of these masses (benign or malignant), making segmentation one of the most important processes in new diagnostic aid systems for breast cancer detection.

On the other hand, in the last decade, various artificial intelligence techniques for segmenting medical images or environments with objects of interest have been studied [30], [31]. The implementations include developments ranging from basic image processing techniques to current deep learning (DL) algorithms, where the latter has exhibited exponential growth in the areas of health informatics and medical imaging [31]–[33]. DL includes architectures such as a convolutional neural network (CNN), which is similar to the primary visual network [34]. In particular, the CNN design can extract complex features at the same level as humans, giving it a more efficient generalization capability than that of conventional machine learning methods. Furthermore, DL can be performed on raw data, i.e., it is unnecessary to perform preprocessing on the input images or to know the background of the problem in detail [35]. Moreover, the paradigm shift toward automatic diagnostic aid systems is a reality inherent to technological evolution due to the generation of large datasets and the development of state-of-the-art computers [36]. The implications of using artificial intelligence range from reducing radiologist workloads, aiding diagnosis, improving response times, and even providing information that is not perceptible to the naked eye during mass breast segmentation; therefore, AI systems are handy tools in daily-life medical practice [37], [38].

Following the above considerations and with the aim of improving the accuracy of automatic breast exam segmentation, we perform a comparative analysis of 12 DL networks with the latest backbones and architectures in this paper. The models are the most efficient and the most widely used methods in classification and segmentation tasks. Additionally, we propose studying these models under the five most-used loss functions to better compare them. The analysis includes architectures such as the original UNet, Visual Geometry Group 19 (VGG19), InceptionResNetV2, EfficientNet, MobileNetv2, ResNet, ResNeXt, MultiResUNet, linkNet-VGG19, DenseNet, SEResNet and SeResNeXt, where "Res" denotes a residual network. The model training process is performed on binary cross-entropy loss functions, including weighted binary cross-entropy, Dice, Tversky focal, and log-cosh Dice functions.

Mainly, the following elements are highlighted in the study.

- A model is found that achieves scores exceeding those reported for state-of-the-art methods in similar works regarding the segmentation breast lesions.
- A comparative analysis of the 12 state-of-the-art architectures with respect to the segmentation task is addressed.
- The state-of-the-art architectures and backbones presented before or during 2021 are included.
- Different loss functions are compared to determine which one has the best performance on the segmentation task.
- The importance of resolution for achieving strong in-network evaluation metrics under the segmentation scenario is highlighted.

Finally, the paper is organized as follows. Initially, work related to the segmentation and classification of breast lesions is addressed. The main DL techniques that have been used to address these problems are highlighted, followed by a brief literature review of the latest and most recent works. Next, details of the methodology used to explore the different networks are given, and the main characteristics of the utilized materials and methods are shown. Subsequently, the results are shown while each of the findings is discussed, leading to a general discussion that highlights the most relevant elements of the study. Finally, the main conclusions are presented.

II. RELATED WORK

Early concepts in mammographic image anomaly automation date back to the 1960s [39]. Originally, developments were focused on minimizing errors due to fatigue or inherent in human execution [39], and since that time, research and developments have included techniques ranging from the basic image processing methods to recent DL techniques [40].

DL has exhibited exponential growth in recent years, and there are even recent reviews highlighting the use of CNNs for different tasks and datasets. For example, Abdelrahman *et al.* [41] achieved advances with modern architectures such as ResNet, UNet, DenseNet, and attention mechanisms. The results demonstrated excellent performances on tasks such as classification, detection, and segmentation. However, the survey lacked heterogeneity between the utilized models and techniques, as all studies have different databases and evaluation metrics, limiting the ability to conduct an objective comparison between architectures [41]. On the other hand, diagnostic aid approaches can also be performed with other techniques.

For example, Zhou *et al.* defined a series of image intensity steps, consisting of background removal, pectoralis muscle removal, and a technique based on a regularized distance level to segment masses [42]. Similarly, Sadeghi *et al.* used image intensity with a new adaptive thresholding method based on variable-size windows. This method allows for the exact location of a mass to be calculated, reducing the possibility of generating false positives [43]. Salih and Kamil proposed mass segmentation through classical and diffuse morphological techniques. Their method processes the breast's internal structures generated from a thresholding process, allowing for highlighting and extracting the lesion of interest [44]. In other more novel approaches, Kamil and Salih implemented two clustering techniques as segmentation methods. In the first technique, they employed the K-means method, and in the second approach, they employed the fuzzy c-means (FCM) algorithm. In both cases, the techniques were integrated with the lazy snapping algorithm as an additional step, improving the segmentation of abnormal areas [45]. These techniques achieved accuracies of 91.18% and 94.12%, respectively.

Although the above methods are promising methods, most researchers focus on the versatility, performance, and advantages of recent DL algorithms. For example, Li, Abdelhafiz, De Moor, and Zhu *et al.* approached the problem of breast lesion segmentation with CNNs [28], [46]–[48]. Li *et al.* combined the densely connected UNet (DenseNet) with attention gates (AG). The model was trained under the cross-entropy loss function, and its performance reached 82.24%, 77.89%, and 78.38% in terms of the F1-score, sensitivity, and accuracy metrics, respectively [28]. Similarly, Abdelhafiz *et al.* used UNet as a base network in two different mass segmentation studies. In the first one, UNet was integrated with residual attention blocks (RUNet). The network was trained with the Dice loss function, and its segmentation ability was validated with the accuracy metric, reaching a value of 98.7% [46]. The second uses a version of UNet enhanced with batch normalization layers, dropout layers, and increasing convolutional layers. Again, the network was trained with Dice loss and achieved an accuracy of 92.6% [47]. De Moor *et al.* used the base UNet for segmentation and evaluated it through free receiver operating characteristics (FROCs). Moreover, in their study, they achieved a maximum sensitivity of 0.94 [48]. Finally, Zhu *et al.* implemented a fully convolutional network (FCN) to model a potential function, followed by the use of a conditional random field (CRF) to perform structured learning. The design was trained with the maximum likelihood loss function (distribution-based loss), achieving a Dice score of 91.30% [49].

Finally, Salama and Aly [50] segmented and classified mammograms by implementing DenseNet121, ResNet50, VGG16, and MobileNetV2 models for classification and a UNet model for breast region segmentation. The results achieved a maximum accuracy of 98.87% for the classification case.

III. MATERIALS AND METHODS

A. DATASET

Three sets of data were taken for the different tasks performed on the mammograms, which are described below.

1) SEGMENTATION

Segmentation was performed on the public database "Curated Breast Imaging Subset Digital Database for Screening Mammography" (CBIS-DDSM) [51]–[53]. The data contain the digital mammograms of several subjects with corresponding segmentation masks performed by expert radiologists. Only mammograms confirmed as normal, benign, and malignant cases, plus verified pathological information, were taken. Therefore, a total of 714 randomly distributed images were used for training, validation, and testing and divided into groups of 499, 72, and 143 images, respectively. It should be clarified that each model was run 20 times, and during each run, the training, validation, and test data were randomly selected to obtain more accurate descriptions of the architectures and ensure that the results did not depend on the splitting of the data. The process is similar to cross validation and is known as Monte Carlo cross validation.

2) CLASSIFICATION

The mammograms were classified in two different ways. The first method was based on the types of lesions, i.e., whether they were calcifications, well-defined or circumscribed masses, spiculated masses, other ill-defined masses, masses with architectural distortion, asymmetric masses or normal areas. In this case, 322 images from the mini-Mammographic Image Analysis Society (MIAS) database of mammograms were used [54]. Second, classification by degree of suspicion (BI-RADS) was performed using 410 images from INbreast, a full-field digital mammographic database [55]. The two datasets were divided into proportions of 60, 20, and 20% for training, validation, and test data, respectively, as shown in Table 1.

Similar to segmentation, during classification, the data were taken randomly in each run, ensuring that the results obtained did not depend on the split.

B. PREPROCESSING

The main feature of DL is that it can work with raw data [35]. For this reason, the mammograms were only subjected to two processes. First, they were normalized, converting the intensity values to a scale from 0 to 1. It should be clarified that the normalization process was performed because neural networks work more efficiently with these values and float data. However, this does not imply a reduction in or loss of information about the images.

Second, due to the large sizes of the images and the differences between them, the images were downsampled as follows: the segmentation images were changed to 512 × 512 size to reduce the computational load. Similarly, the images for classification were also reduced in size. However, their sizes were set to 256 × 256 to reduce the computational load and to increase the number of images for use with the new data augmentation method explained in the next section.

TABLE 1. Data used in the classification process and their respective splits for training, validation, and testing before data augmentation.

MIAS data				
Lesion	DC	TR	VA	TE
MISC: Other ill-defined masses	14	8	3	3
CIRC: Well-defined/circumscribed masses	23	13	5	5
CALC: Calcification	25	15	5	5
SPIC: Spiculated masses	19	11	4	4
ASYM: Asymmetry	15	9	3	3
ARCH: Architectural distortion	19	11	4	4
NORM: Normal	207	124	41	42

INbreast data				
BI-RADS	DC	TR	VA	TE
1: Negative	67	40	13	14
2: Benign	220	132	44	44
3: Most likely benign	23	13	5	5
4a: Low suspicion of malignancy (2–9%)	13	7	3	3
4b: Moderate suspicion of malignancy (10–49%)	8	4	2	2
4c: High suspicion of malignancy (50–94%)	22	13	4	5
5: Highly suggestive of malignancy	49	29	10	10
6: Known biopsy-proven malignancy	8	4	2	2

DA: Data by class, TR: training, VA: Validation and TE: test

C. DATA AUGMENTATION

Data augmentation techniques were integrated into this study to increase the size of the training set and avoid overfitting the model.

For the case with the segmentation images, a total of 7 additional images were generated for each image, yielding 3992 mammograms for training. The images were generated by inverting the pixels from right to left and rotating the images at 90-degree angles in all possible positions, as illustrated in Figure 1. Similarly, this process was performed for the true segmentation of the breast lesions.

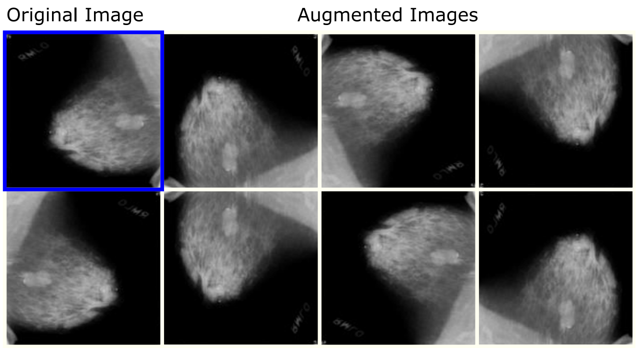

FIGURE 1. Example of images generated through the data augmentation method. Example of data augmentation for a single image. The original image (blue box) was rotated at angles of 90, 180, and 270 degrees (upper images). This same rotation was used to flip the images from left to right (lower images). The mammograms were taken from [51]–[53].

The two databases used in the classification process had unbalanced data, i.e., some classes had few subjects, while other classes had significant numbers of subjects. Therefore, data augmentation was performed in two different ways. In the first approach, eight images per subject were augmented regardless of the number of subjects per class.

In the second case, data augmentation was performed so that all classes had the same number of images, generating up to a maximum of 16 images per subject. The 16 possible images were generated by resizing the images to 1024 × 1024. Subsequently, each image was divided into 256*256 quadrants. Each quadrant had a size of 4 × 4. Finally, each image was generated by taking the pixels of each quadrant at the same position. The process is illustrated in Figure 2.

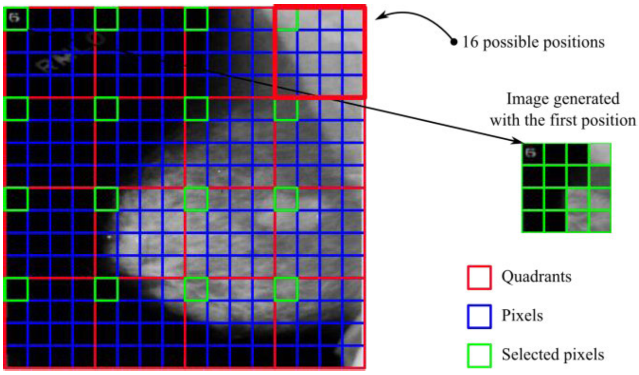

FIGURE 2. Reduction in image size and increase in number of data based on reduction. New images were created from the selected pixels in each quadrant. The mammograms were taken from [51]–[53].

D. SEGMENTATION WITH DL NETWORKS

In medical imaging, the segmentation process consists of classifying each pixel into all possible interest elements (e.g., background and affected tissues). UNet [49] is one of the most popular networks for this task and was originally created to focus on medical imaging. The structure consists of convolutional layers interspersed with clustering layers, forming an encoder-decoder design. The low-level layer features are combined with the high-level layers (see Figure 3), preserving some of the spatial information. The encoder extracts the features, and the decoder performs upsampling. The major difference between UNet and other segmentation networks is that UNet adopts splicing and fusion in the channel dimension. Additionally, the network can be trained with a low amount of data since its structure can converge quickly [56].

The design allows UNet to be highly efficient in the segmentation process. However, recent studies have presented variations of the network (e.g., backbones of other models), which could be more efficient for this task. In this sense, we proposed evaluating 11 of the most novel convolutional networks, including UNet, as a reference. The implemented networks are shown in the following table with their relevant characteristics.

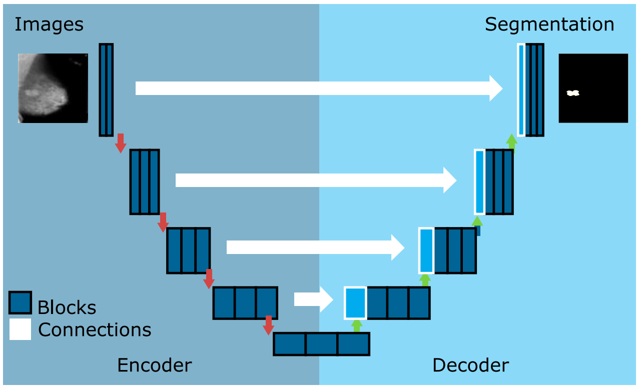

FIGURE 3. Base architecture of a deep neural network for segmentation. Based on [56].

The networks are made up of several layers, and each convolutional layer uses filters to extract the desired features once the model is trained. The following mathematical model governs the convolutional layers.

$$\boldsymbol{A}_j^{(l)} = \boldsymbol{f}^{(l)}\left(\sum_{i=1}^{M} \boldsymbol{A}_i^{(l-1)} * \boldsymbol{K}_{ij}^{(l)} + \boldsymbol{b}_j^{(l)}\right) \tag{1}$$

where $\boldsymbol{A}_j^{(l)}$ is the feature map (output) of the l-th convolutional layer associated with the j-th convolutional filter ($\boldsymbol{K}_{ij}^{(l)}$). $\boldsymbol{A}_i^{(l-1)}$ is the output of the previous layer or the input for the l-th layer. $\boldsymbol{b}_j^{(l)}$ is the bias, and $\boldsymbol{M}^{(l-1)}$ is the number of feature maps in the previous layer. Additionally, f denotes a nonlinear activation function, which usually consists of a rectified linear unit (ReLU). Despite the fact that all neural networks are based on convolutional layers, their behaviors can vary significantly due to their structural designs, i.e., the depth of the network (number of layers); the number of filters, connections or trajectories; and specific features, as described in Table 2.

The input image passes through the architecture, generating the output (the training parameters; see Figure 3). Finally, the training results are validated using the associated loss function, and the model is iteratively adjusted until the best model performance is obtained.

E. LOSS FUNCTIONS

As mentioned above, image segmentation consists of classifying pixels into different types of elements, usually those associated with the background and the object of interest (e.g., breast lesion). The difference between the regions spanned by the elements (data imbalance) usually causes the networks to be biased toward the larger element. However, some loss functions can circumvent this problem. These can be classified into four different types: distribution-based, region-based, boundary-based, and composite loss functions [67]. Therefore, we proposed using five of the most common loss functions to fit the segmentation models to the training data. The utilized loss functions are described in detail below.

TABLE 2. State-of-the-art CNN architectures.

Networks	Date of publication	Remarks
VGG19 [57]	2015	Sequential convolutional architecture with a depth of 19 weight layers. *
ResNet [58]	2016	First network with residual connections between convolutional layers. *
InceptionResNetV2 [59]	2017	Modified multipath convolutional network with residual connections. *
DenseNet [60]	2017	Architecture with direct-access connections (throughout the network) or a densely connected structure. *
LinkNet-VGG19 [61]	2017	Deep architecture that allows for greater model tuning without a significant increase in the number of parameters. *
ResNeXt [62]	2017	A network with residual connection blocks and building blocks that aggregate a set of transformations with the same topology. *
SEResNet [63]	2018	A network composed of "squeeze-and-excitation" blocks, which adaptively recalibrate channel feature responses by explicitly modeling the interdependencies between channels. *
SEResNeXt [63]	2018	A network composed of "squeeze-and-excitation" blocks, which adaptively recalibrate the responses of channel characteristics by explicitly modeling the interdependencies between channels. Additionally, it has blocks that aggregate a set of transformations with the same topology. *
EfficientNet [64]	2019	Architecture tuned through a composite coefficient that scales the width, depth, and resolution dimensions of the network. *
MobileNetv2 [65]	2019	A network based on an inverted residual structure. The residual block's input and output are thin bottleneck layers, allowing for the filtering of features in the intermediate expansion layer. *
MultiResUNet [66]	2020	Modern network with blocks and connections between blocks based on residual connections. Originally used in dermoscopy, endoscopy, fluorescence microscopy, electron microscopy, and MRI.

* Originally trained on natural images, i.e., not on medical images.

1) BINARY CROSS ENTROPY

Binary cross entropy is a loss function that is commonly used to measure the difference between two probability distributions. This principle can be applied to individual pixels in images, classifying elements into two possible values: the background and the object of interest [68]. The binary cross-entropy loss (L_{BCE}) is mathematically defined as:

$$L_{BCE}\left(y,\hat{y}\right) = -\left(y\log\left(\hat{y}\right) + (1-y)\log\left(1-\hat{y}\right)\right) \quad (2)$$

where y is the true value (label) and $\hat{y}$ is the predicted probability of the label for the same element in the dataset. It should be clarified that the binary cross entropy in a dataset is defined as the average of all the elements that compose the dataset.

2) WEIGHTED BINARY CROSS ENTROPY

As in the previous case, the weighted binary cross-entropy loss is used to measure the difference between two distributions. However, this variant weights the sets, eliminating the bias induced by imbalanced data [69]. The weighted binary cross-entropy loss is mathematically defined as:

$$L_{WBcE}\left(y,\hat{y}\right) = -\left(\beta y\log\left(\hat{y}\right) + (1-y)\log\left(1-\hat{y}\right)\right) \quad (3)$$

Here, y is the true value (label), $\hat{y}$ is the predicted probability of the label, and β is the weighting coefficient used to adjust for false positives or false negatives.

3) DICE LOSS

The Dice coefficient is a statistic used to calculate the similarity between two samples. Its use can be extended to images by comparing the similarity between spatially matched pixels [70]. The coefficient has also been included in a loss function, which is mathematically defined as:

$$DL\left(y,\hat{y}\right) = 1 - \frac{2y\hat{y}+1}{y+\hat{y}+1} \quad (4)$$

where y is the true value (label) and $\hat{y}$ is the predicted probability of the label. It should be noted that equation (4) is modified with a 1 in the numerator and denominator, ensuring that the function is defined even in extreme cases where y and $\hat{y}$ are equal to zero.

4) FOCAL TVERSKY LOSS

The Tversky index is a measure of asymmetric similarity between sets [71]. This function can be viewed as a generalization of Dice's coefficient, and it is mathematically expressed as follows:

$$TI\left(y,\hat{y}\right) = \frac{y\hat{y}}{y\hat{y} + \beta\left(1-y\right)\hat{y} + \left(1-\beta\right)y\left(1-\hat{y}\right)} \quad (5)$$

Equation (5) averages the false positive and false negative weights through the coefficient. Similar to Dice's coefficient, Tversky's index can also be fitted to a loss function as follows [72]:

$$TL = 1 - TI \quad (6)$$

The loss function can be modified to a focal loss by reducing the weights of individual examples and focusing the training process on hard negatives through a modulation factor γ [73], as shown below:

$$FTL = \sum_{c} (1 - TI_c)^{\gamma} \tag{7}$$

Here, the modulation factor must meet the condition of $\gamma > 0$.

5) DICE LOG-COSH LOSS

Dice's coefficient is widely used in computer vision for conventional images. However, due to its nonconvex nature, the smoothed version using a hyperbolic log-cosine has recently been proposed [67]. The loss function is defined as follows:

$$L_{DL} = \log(\cosh(DL)) \tag{8}$$

Here, DL is the loss with the Dice coefficient established in equation (4).

F. EVALUATION METRICS

As an important part of the objective model comparison d, our approach was based on five evaluation metrics: the Dice coefficient, sensitivity, specificity, accuracy, and F1-score. The five metrics are mathematically expressed as follows:

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{9}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{10}$$

$$specificity = \frac{TN}{TN + FP} \tag{11}$$

$$Accuracy = \frac{TP + TN}{TN + TP + FP + FN} \tag{12}$$

$$F1\_score = \frac{2TP}{2TP + FP + FN} \tag{13}$$

where the five metrics are established in terms of the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

G. EXPERIMENTAL DESIGN

1) SEGMENTATION

A comparative analysis among the models in terms of their automatic anomaly segmentation performance in mammographic images was proposed. The analysis addressed the study of the five loss functions and the 12 architectures described above. The performance of the networks was observed during training with the Dice coefficient to determine the architecture with the best performance. Similarly, validation was performed only through the Dice coefficient. Finally, all networks were evaluated (after training) with the test set under the Dice coefficient, sensitivity, specificity, and accuracy metrics.

Figure 4 shows a graphical summary of the experimental design. It started with the CBIS-DDSM dataset, and these values normalized and divided into three datasets at proportions of 70, 10, and 20% for training, validation, and testing, respectively. The training dataset was augmented with the proposed method, and subsequently, the models were trained on all possible combinations of the five loss functions (see Equations (2), (3), (4), (7), and (8)) and the 12 deep architectures (see Table 2). Each model was trained for 150 epochs with the training dataset, and their parameters were adjusted to the optimal values. At each epoch, the models were evaluated on the training and validation data. Finally, the trained models were used to generate predictions for the test data, and the resulting scores were calculated through the true segmentation and evaluation metrics.

Each network was run 20 times on average to generate different scores and obtain the approximate distribution of metrics. Besides, each network was run under the following hyperparameters:

- Number of epochs: 150
- Optimizer: Adam
- Batch size: 4
- Weights initialization: Uniform Glorot
- Bias initialization: Zeros

The performance metrics provide detailed descriptions of the implemented models. However, in discrete space, each metric's effectiveness is subject to the size of the element of interest. For example, in Equation 9, a difference of only one pixel between the actual and predicted segmentation results (false positives or false negatives) would not generate a low Dice score for a large region (many TPs). In contrast, small regions would generate low values based on a difference of only one pixel. Therefore, to obtain a more detailed description of the Dice score as a function of size, the best network scores were compared with scores generated from the dilated and eroded masks. These processes introduced minimal error in the real regions by increasing or decreasing the perimeter by one pixel (dilation and erosion).

2) CLASSIFICATION

Similar to the previous case, a comparative analysis of state-of-the-art CNNs was proposed for breast lesion classification. In this case, the most common classifications were used: classification by lesion type and by degree of suspicion. In the two classification processes, seven CNNs were addressed under the cross-entropy loss function. The performance of the networks was observed during training through accuracy and validated with the same metric (accuracy). Finally, all models were evaluated (after training) with the F1-score, accuracy, sensitivity, specificity, and precision metrics.

The process is shown in Figure 4. Initially, the datasets were split and preprocessed; data augmentation was performed using the proposed method and the different networks were trained. It is worth noting that each network was run an average of 40 to generate different scores and obtain the approximate distribution of metrics. Besides, each network was run under the following hyperparameters:

- Loss function: Cross entropy.
- Number of epochs: 40
- Optimizer: Adadelta
- Batch size: 10
- Weights initialization: Uniform Glorot
- Bias initialization: Zeros

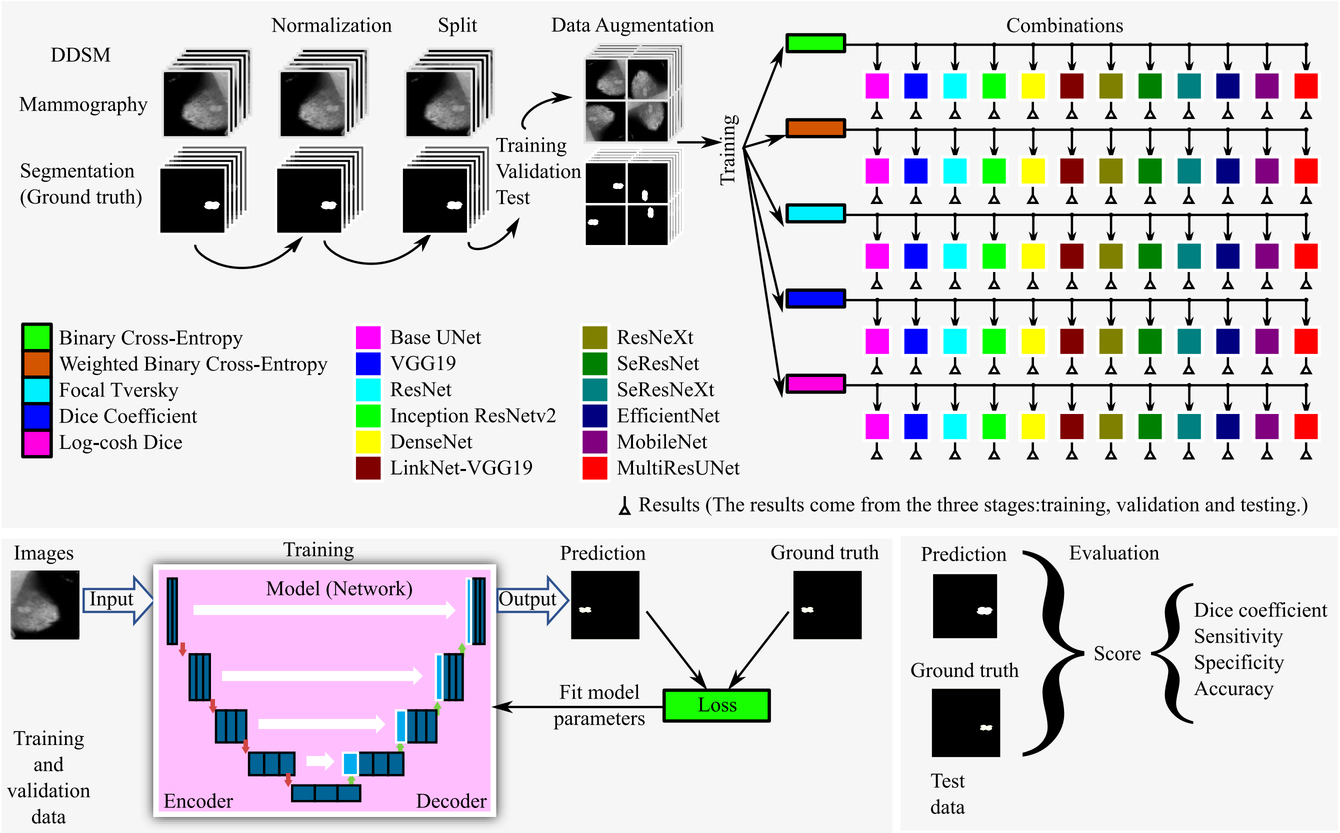

FIGURE 4. General scheme of the experimental design for the comparison between the twelve DL models and the 5 loss functions. The process started with the CBIS-DDSM dataset and continued with normalization, splitting, data augmentation, and training with all possible combinations of the networks and loss functions. The training process used the real data to fit the parameters with the training and validation data. The performance results were evaluated with the test Dice scores. The mammograms were taken from [51]–[53].

The architectures were modeled with the Python programming language by utilizing the main Keras and TensorFlow libraries. The models were executed on a Colab platform configured with 25 GB of RAM and a Tesla T4 GPU. The implemented codes are publicly available in the following GitHub repository: (https://github.com/Qsinap/Breast-cancer-segmentation).

IV. RESULTS AND DISCUSSIONS

A. SEGMENTATION

This section shows the results generated by the models under the different loss functions. The tables are presented with percentage values and graphs containing scores in their fractional form, i.e., with values ranging from 0 to 1 that are equivalent to values from 0 to 100%.

Table 3 shows the maximum metric values achieved by the 12 deep CNNs. In this case, the EfficientNet architecture delivered the best result, reaching a Dice score of 94.75%. Moreover, the model achieved the highest sensitivity with a value of 95.21%, ensuring a low negative false rate, i.e., a small loss of lesioned regions. Likewise, the specificity score was 99.99%, indicating a low probability of generating false regions as lesions.

Additionally, although EfficientNet was not the architecture with the highest number of convolutional layers, it had the highest number of training parameters, i.e., this architecture had more filters per convolutional layer, allowing it to obtain a higher number of feature maps per layer.

Similarly, the InceptionResNetV2 network exhibited similar behavior to that of EfficientNet. The results show that the same specificity and sensitivity values were achieved, with an almost 1% difference. The Dice score also yielded a difference of less than 1%, i.e., the network had high performance (slightly lower than that of EfficientNet) but with a smaller number of training parameters, which implies that a lower computational load was required when implementing this model.

In contrast, the UNet base network was the model with the lowest number of training parameters and the lowest number of convolutional layers. This explains the low performance obtained by this network compared to that of the other architectures to a large extent.

Similarly, Table 4 indicates the highest scores achieved by the loss functions, where binary cross entropy generated the highest Dice score, sensitivity, and specificity. Otherwise, the Dice, Tversky focal, and log-cosh Dice losses were assumed to be more efficient since they are region-based

TABLE 3. Best Dice scored achieved by the twelve deep CNNs and a summary of their structures.

Networks	Maximum Dice (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Training parameters (Millions)	Convolutional layers	Conv2D	Transposed convolution or upsampling	Pooling	Normalization	Concatenate	Merge/Add
EfficientNet	**94.75**	**95.21**	**99.99**	**99.96**	75.05	289	284	5	55	173	4	48
InceptionResNetV2	**94.06**	**94.23**	**99.99**	**99.95**	62.06	260	255	5	5	214	47	0
ResNet	93.60	93.25	99.99	99.95	67.30	223	166	5	1	163	4	50
ResNeXt	93.52	93.67	99.99	99.95	51.28	1178	1138	5	1	115	37	33
SEResNet	93.00	91.27	99.99	99.94	73.94	323	266	5	51	165	4	50
SEResNeXt	92.86	91.84	99.99	99.94	56.06	1244	1204	5	34	114	37	33
DenseNet	92.79	91.54	99.99	99.94	26.38	218	211	5	4	211	102	0
MobileNet	90.59	90.19	99.99	99.92	8.05	73	63	5	0	62	4	10
MultiResUNet	90.29	90.88	99.99	99.94	7.26	61	57	4	4	85	13	19
LinkNet-VGG19	89.93	87.46	100.00	99.91	25.63	39	34	5	5	17	0	4
VGG19	86.54	80.55	100.00	99.89	29.06	34	29	5	5	12	4	0
Base UNet	86.51	86.79	99.99	99.90	1.95	23	19	4	4	18	4	0

functions [67]. However, the results were more than 4% below those of the binary cross-entropy loss function and its weighted version. This finding implies that small-element segmentation is not performed well with region-based functions and would be performed better with distribution-based losses.

TABLE 4. Best scores by loss function.

Losses	Maximum Dice (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)
Binary CE	**94.75**	95.21	100	99.96
Weighted Binary CE	**94.47**	95.21	100	99.95
Focal Tversky	90.11	90.85	99.96	99.90
Dice coefficient	89.29	85.56	100	99.90
Log-cosh Dice	87.60	84.67	100	99.88

CE: cross entropy

Figure 5 shows two examples of the segmentations generated by the 12 models with the highest Dice scores (see Table 3). The process was performed on large and small lesion. The results were obtained with the respective loss functions that performed best for each network and were generated with the test data. Each example contained a mammogram, an enlarged image of the region of interest (ROI), the actual segmentation, and the probability or prediction map generated by the model. Among the predictions, it can be seen that EfficientNet presented a heat map that was similar to the real region, even with the small ramifications presented by the lesion. Additionally, the network approached the real region in the small lesion and presented more defined edges.

Likewise, the InceptionResNetV2 network generated a probability map similar to that of the EfficientNet network. The central region appeared with high probability values (close to 1), guaranteeing low uncertainty in this region. In particular, the differences occurred at the edge, where the probability decreased because the network could not classify the pixels as lesions or nonlesions. It should be noted that this same effect was present in EfficientNet, making it difficult to discern the visual differences in between two results.

In contrast, the basic UNet displayed diffuse edges in both cases, generating a poor segmentation result for the element of interest and producing the lowest metric scores, as shown in Table 3.

As mentioned above, EfficientNet achieved the best performance; therefore, Figure 6 shows the training (Figure 6a) and validation (Figure 6b) results of this model only. The graphs are the training averages for all loss functions, where the translucent bands are the 95% confidence intervals. Figure 6a shows that the Dice coefficient increased, indicating better performance in each successive training epoch. Similarly, Figure 6b presents the same behavior, with the same values reached at the end of the 150 epochs. This behavior reveals that the model did not overtrain (overfitting) for any of the 5 loss functions, and again, it can be seen that the binary cross-entropy and binary-weighted cross-entropy loss functions achieved the best performance. It should be noted that all models presented the same behavior, i.e., they did not overfit the day. However, the results of the other models were made available to the public in the GitHub repository (https://github.com/Qsinap/Breast-cancer-segmentation).

Figure 7 presents the overall results of all the training processes. Figure 7a shows the distribution of the Dice scores generated by all the training sessions. The figure shows that EfficientNet had the most homogeneous distribution with the highest values. The behavior assumes a probability of conducting training with the highest score through this network. Similarly, EfficientNet network a more homogeneous sensitivity distribution than those of the other networks (see Figure 7b).

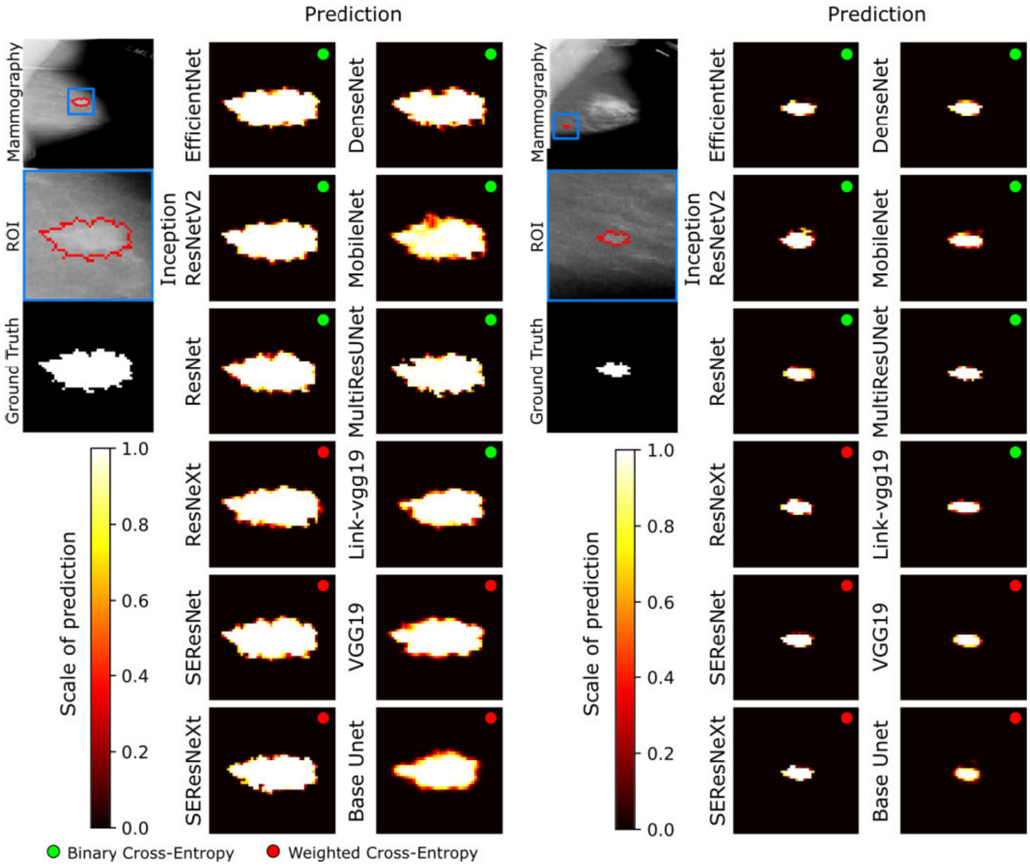

FIGURE 5. Examples of the segmentations generated by the deep neural networks with their respective mammograms, the region of interest, the real region, and the heat map associated with each model's prediction. The mammograms were taken from [51]–[53].

In the case of specificity, all the networks had distributions above 99.8%, with values generated by the background's large presence and low probabilities of generating false positives, i.e., pixels classified as injuries. Finally, Figure 7d shows that the 5 loss functions exhibited same behaviors regarding the metrics. The specificity had homogeneous distributions close to 1, while the Dice score and sensitivity distributions displayed more heterogeneous behaviors. Additionally, in this last graph, the crossed and weighted binary entropy distributions of the losses are above the other losses. In fact, the first quartile of the first loss function is above the last three, showing a marked difference between these functions.

The test set consisted of 143 images of patients with different lesion sizes. Consequently, Figure 8 shows the real breast lesion area and the areas generated by the best and worst models, i.e., by EfficientNet and the base UNet (see Table 3). The results show high agreement for EfficientNet, even for small lesions (see Figure 8a). Likewise, the InceptionResNetV2 network exhibited similar behavior to that of the EfficientNet network (see Figure 8c); however, there were some divergences for small lesions, limiting the segmentation ability of this model. UNet presented large differences with small lesions; however, the area was closer to the real values when the lesion was larger (see Figure 8b).

Although EfficientNet presented an average Dice score of 94.75% (see Table 3), Figure 8c shows that lesions with larger areas yielded scores above 95%. In contrast, smaller lesions led to values below 95% and even 90%.

Additionally, Figure 8c introduces the Dice score generated from the dilated and eroded masks, i.e., the score between the real mask and the mask with introduced morphological transformation error. As expected, the introduced error affected the behavior of the Dice score. The coefficient decreased by up to 40% for smaller lesions. Otherwise, the Dice score had high values for larger lesions even though they were generated with the same error type. On the other hand, theoretically, erosion and dilatation affect the internal and external perimeters, respectively. Consequently, the outer perimeter was expected to be larger than the inner perimeter, generating a larger error and affecting the Dice score to a greater extent. However, Figure 8c shows that erosion created a greater reduction in the Dice score. The metric's behavior versus the induced error reveals the dependence of the models on the sizes of the segmented regions. In other words, even if the segmentation effects are good,

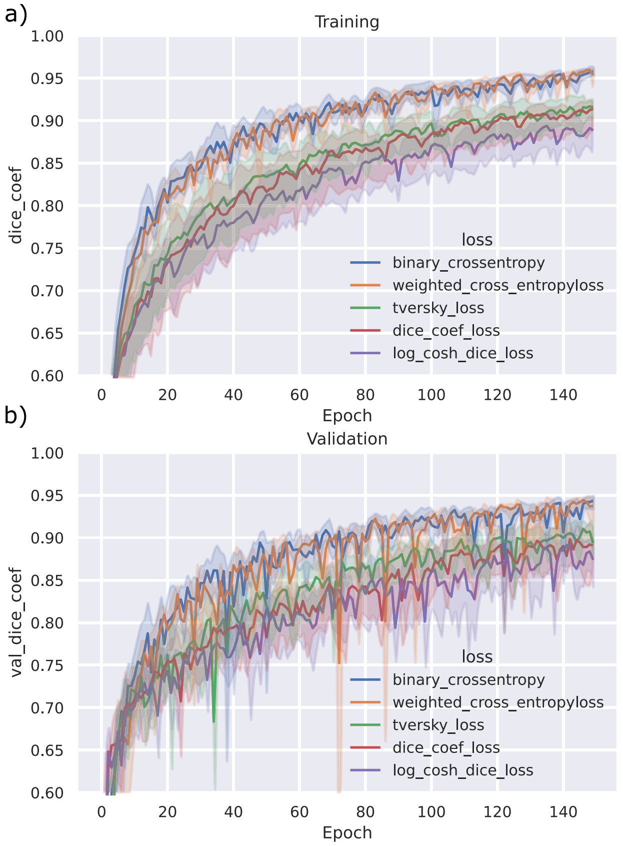

FIGURE 6. Training and validation of EfficientNet as a function of the number of epochs. The plots show the performance of the network with the Dice coefficient metric for the five loss functions in the a) training and b) validation processes.

the corresponding scores can decline significantly for small elements.

Finally, EfficientNet maintained high values despite the inherent failure of the Dice coefficient in discrete space.

Figure 9 shows the segmentation performed by EfficientNet on two lesions: a large lesion and a small lesion. The network prediction map closely resembles the actual segmentation; however, the Dice coefficient varies drastically between these two examples, confirming the impact that the size of the element of interest has on the Dice coefficient.

Figure 10 shows the average time requirements of the 12 models for automatic segmentation. In this case, the UNet base network presented the shortest segmentation time; however, the other architectures had comparable times, with values below 15 milliseconds.

Finally, as shown in Table 5, the results showed that EfficientNet achieved better scores than other similar works, guaranteeing better segmentation results with respect to masses on mammographic images.

Although there are innovative state-of-the-art DL architectures, it makes little sense to evaluate the models from their structures since each is complex regarding the elements that are not directly comparable to each other. In this sense, the performance of the models is directly summarized by their evaluation metrics, i.e., a model is efficient if the evaluation metrics are high relative to other models. Consequently, the results in Table 5 clearly show that one of the proposed networks outperformed the results reported to date in terms of the segmentation of mammographic images. This is very useful for finding the affected regions (breast lesions) in short times and with high performance, making this network a handy tool in clinical settings.

The results show that the proposed network generated higher scores than related approaches. For example, regarding the accuracy metric, EfficientNet reached a score of 99.96%, surpassing the maximum score achieved by the method of Abdelhafiz *et al.* [46] across care models. In the case of sensitivity, EfficientNet scored 95.21, surpassing the maximum score reported by de Moor *et al.* [48] (94%). Similarly, the Dice coefficient reached 94.75%, exceeding the value reached by Zhu *et al.* [49] with multiscale networks by almost 4%.

Finally, the results clearly show that the new DL architectures are at the technological forefront in terms of screening breast masses. However, there are some inherent limitations relative to the problem at hand. For example, the Dice metric is the most reported measure in the literature. However, the size dependence of the element of interest creates a bias that limits the objective evaluation of this work and existing work that has been evaluated with the same metric. In this sense, an approach could be sought to adjust the Dice metric to avoid the drops in the coefficient due to small regions. On the other hand, from a methodological point of view, the study was based on an extensive dataset; however, this does not cover all possible considerations of breast examinations, as several protocols generate different types of mammographic images.

Additionally, it is necessary to include outlier images to obtain more detailed descriptions of the networks in the face of these drawbacks. These challenges could be overcome by searching for and including new databases with different characteristics, where transfer learning could be conducted from the current methods to the models with the new databases. This transfer would enable the models to avoid starting with random parameters and allow them to reach lower loss function values more quickly.

B. CLASSIFICATION

This section shows the results obtained for the two types of classification problems, i.e., classification by the types of lesions and classification by the degree of suspicion in BI-RADS. As in the previous case, the tables present the results in percentage values, while the graphs show the values in their fractional forms.

Initially, classification by lesion type was performed with the 7 CNNs: VGG19, ResNet50V2, DenseNet121, InceptionV3, InceptionResNetV2, EfficientNetB7 and Xception. Table 6 shows the overall results obtained for the test data in terms of the five different metrics. However, the results were organized from the highest to lowest F1-score metrics. The F1-score provides a better description of this

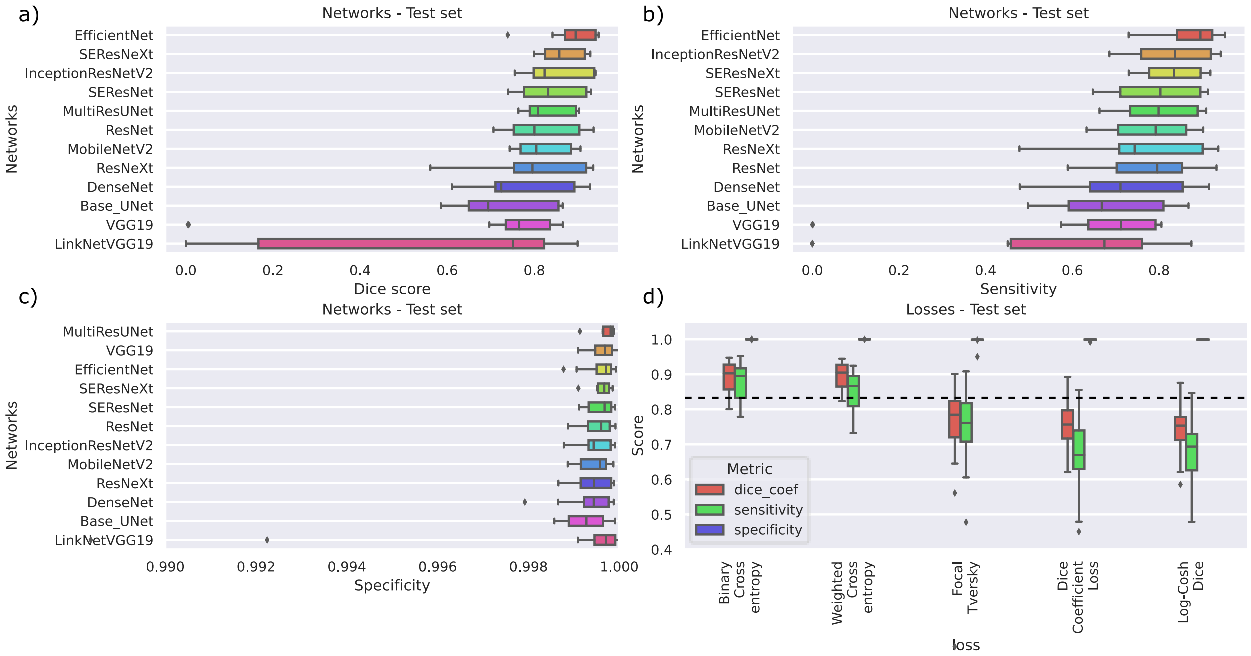

FIGURE 7. **Distributions of scores generated by training the 12 deep CNNs for the a) Dice coefficient; b) sensitivity; and c) specificity metrics; d) Distributions of scores as functions of the loss functions with the same three metrics. The results were derived from the test data.**

case, presenting the imbalances between the six classes' images.

In particular, the results show that the VGG19 network achieved the maximum F1-score on the test data. This network achieved the best performance even though the network was one of the worst networks in the segmentation task. The result confirms the need to search for a network carefully for each specific task. That is, if a network has the best performance on one task, this does not guarantee that this behavior will be maintained in other types of tasks.

On the other hand, Table 6 shows how misleading the accuracy metric can be. The metric was above 95% for all networks. However, the sensitivity dropped to 14% for the Xception network. That is, the network achieved 95% accuracy but had a low ability to identify true positives. Additionally, it is worth noting that the marked difference between accuracy and sensitivity is due to the class imbalance problem, where it is possible that there are true positives than true negatives.

The results in Table 6 also show the high effectiveness of the ResNet50V2 network. Although the network was 5% below VGG19, it remained among the most efficient networks for classification and segmentation.

Table 7 shows the maximum classification scores achieved for each of the lesion types. The results contain high F1-scores for normal lesions. That is, the models were highly efficient in discriminating mammograms without any abnormalities. However, it is also worth noting that the results could have been generated due to data imbalances. For example, the normal class yielded the highest probability percentage, but this class occurred more frequently than the others (see Table 1). Similarly, the MISC lesions (ill-defined masses, others) had the lowest frequency. Consequently, it was the class with the worst scores.

Table 8 shows the results for the case of classification by the degree of suspicion (BI-RADS). The networks presented similar behaviors to those in the previous case. The ResNet50 and VGG19 networks generated the best performances. However, in this case, the ResNet50 network outperformed the VGG19 network by more than 6% in terms of the F1-score.

Similarly, the accuracy metric did not present significant differences between the networks, and all of values exceeded 97%. Again, the differences were found among the sensitivities of the networks, where there was a difference of approximately 47%. In other words, the ResNet50 network had a better performance than EfficientNet in terms of discriminating between true positives. It should be noted that all the networks were excellent at determining true negatives (high specificity), which could be attributed to the greater probability of encountering a true negative.

Table 9 shows the scores achieved according to the different grades of suspicion regarding the test data. Class 1 yielded a marked difference relative to the other classes, i.e., negative mammograms were clearly distinguishable unlike the other classes. In fact, suspicions highly suggestive of malignancy (6) produced the second-best results, but with an almost 47% difference from the first-place results. In this case, the marked difference between the classes could not be directly attributed to class imbalance since the benign class (3) had the highest number of mammograms (see Table 1), but its F1-score reached 11.43%. In summary, Table 9 shows the high performance of the models in detecting true negatives

TABLE 5. Comparative summary of similar works and the two best architectures.

Author	Method	Loss Function	Advantages	Disadvantages	Metric	Score
Kamil	k-mean	Not applicable	ML methods are faster with lower computational burdens. This method is based on unsupervised algorithms, i.e., it does not need training data generated by a radiologist. Used under the MIAS database.	It has low generalizability, i.e., the results have low reproducibility if any changes are generated or if other images are used. The accuracy metric is an inferior metric that does not consider class imbalance.	Accuracy	91.2
Kamil	fuzzy c-mean	Not applicable	ML methods are faster with lower computational loads. This method is based on unsupervised algorithms, i.e., it does not need training data generated by a radiologist. Used under the MIAS database.	It has low generalizability, i.e., the results have low reproducibility if any changes are generated or if other images are used. The accuracy metric is an inferior metric that does not consider class imbalance.	Accuracy	94.1
Li	DenseNet with AG	Cross entropy	This model has relatively low parameters compared to those of the other models, and the architecture handles the vanishing gradient problem well, i.e., it does not encounter a problem when training the first deep layers. Attention models allow it to focus on more relevant regions, allowing it to preserve image attributes for better segmentation. Used under the CBIS- DDSM database.	It is trained only with the cross-entropy loss function, as it is a measure based on the input data distribution.	F1-score	82.2
Abdelhafiz	UNet with attention blocks	Dice	Attention models allow the model to focus on more relevant regions, allowing it to preserve image attributes for better segmentation. Used under the CBIS- DDSM and INbreast databases.	The UNet network is the most straightforward deep network among the fully convolutional models, and its segmentation capability is limited relative to that of modern models.	Accuracy	98.7
Abdelhafiz	UNet with normalization layers	Dice	Normalization layers prevent model overtraining. Used under the CBIS-DDSM and INbreast databases.	The UNet network is the most straightforward deep network among the fully convolutional models, and its segmentation capability is limited relative to that of modern models.	Accuracy	92.6
De Moor	UNet with FROC	Weighted logistic	The simplicity of the UNet network is generally reflected in shorter run times and lower computational loads. Used under own mammograms.	The UNet network is the most straightforward deep network among the fully convolutional models, and its segmentation capability is limited relative to that of modern models. It is trained only with the weighted logistic loss function, as it is a measure based on the input data distribution. It is evaluated through sensitivity, a metric that does not consider unbalanced data between the background and element of interest.	Sensitivity	94
Zhu	CNN with CRF	Maximum likelihood	This model uses multiscale fully convolutional neural networks to achieve good segmentation results for small masses. Used under the CBIS-DDSM database.	It uses the maximum likelihood function based on the distribution of the data without considering the imbalance between the background and element of interest, limiting the performance of the network.	Dice	91.3
Proposed	EfficientNet	Binary cross entropy			Accuracy	100
					Sensitivity	95,2
					Dice	94,8
Proposed	InceptionResNetV2	Binary cross entropy			Accuracy	100
					Sensitivity	94,2
					Dice	94,1

(specificity), but they have low abilities to determine true positives.

Based on the results shown in Tables 6 and 8, it is clear that the VGG19 and ResNet50 networks are the best architectures for classifying lesions and the degrees of suspicion regarding breast lesions. Therefore, Figures 11 and 12 show the average behaviors exhibited during the 40 different training runs.

Figure 11 shows the training and validation results of the VGG19 network as a function of the 40 epochs. In this case, the different training runs did not present significant differences since the error band (translucent color) was small. That is, the VGG19 network exhibited stability during training, guaranteeing convergence to similar training and validation scores. Moreover, the training and validation curves converged above 90%, guaranteeing a low degree of overtraining, which agreed with the test results shown in Table 6. On the other hand, both the accuracy and loss curves showed slight divergences between training and validation, which were generated near epoch 30, i.e., the models require approximately 30 epochs to reach the best performance without overfitting.

Similarly, Figure 12 shows the training and validation of the ResNet50 network in terms of the accuracy metric and

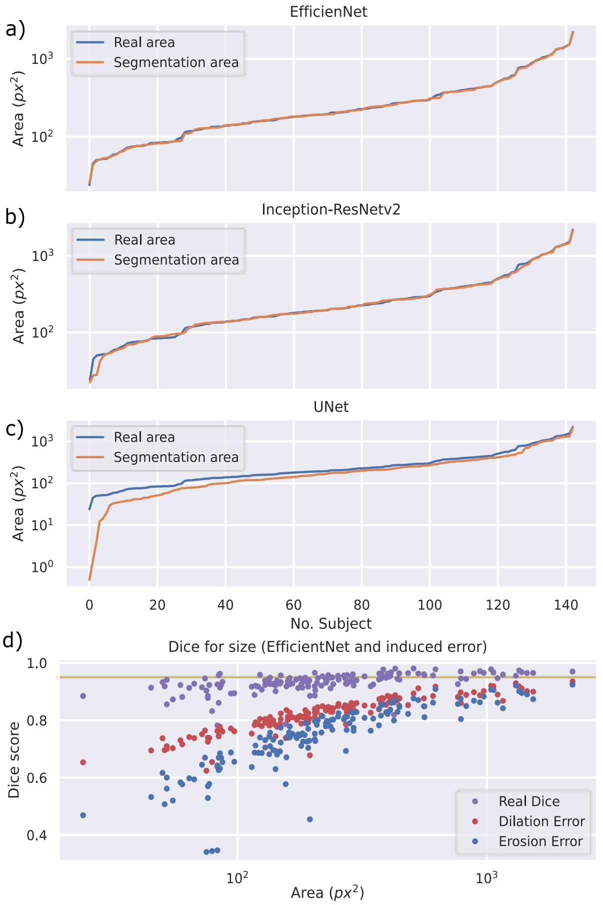

FIGURE 8. Results as a function of breast lesion size for the 143 test subjects. Comparison between the actual areas and the automatic segmentation of a) EfficientNet; b) InceptionResNetV2; and c) UNet (base); d) Dice scores as functions of the area generated by EfficientNet and by the induced dilation and erosion errors. A yellow line corresponding to a score of 0.95 is included in the graph.

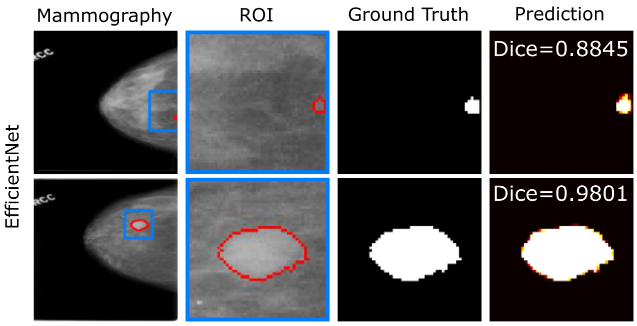

FIGURE 9. Comparison of the Dice scores obtained for two lesions of different sizes. The mammograms were taken from [51]–[53].

the loss of the model. The network exhibited stability between the different parts of training, generating a reduced error band and converging close to 0.8. The result guarantees that the model was not overtrained, and the results agree with those obtained in Table 8. On the other hand, the losses of training and validation exhibited similar behaviors, corroborating the fact that the model did not overfit. However, the curves do not show any apparent divergences, so it is possible to train the model over a greater number of epochs to obtain a better result.

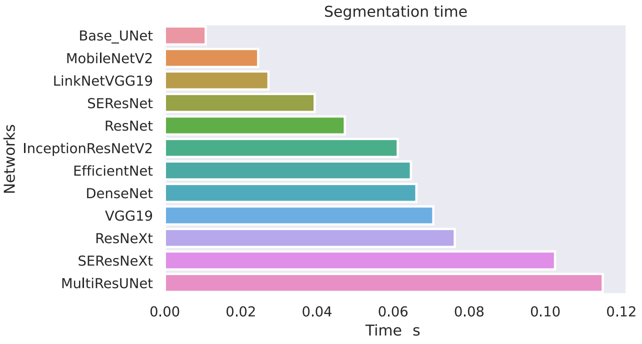

FIGURE 10. Segmentation times per subject for the 12 deep CNN models.

TABLE 6. Maximum clustering scores achieved by the different neural networks - MIAS database.

Network	F1_score (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
VGG19	79.12	96.97	92.86	100.00	100.00
ResNet50V2	74.47	95.45	83.33	100.00	100.00
DenseNet121	71.74	96.97	78.57	100.00	100.00
InceptionV3	54.55	95.45	50.00	100.00	100.00
InceptionResNetV2	50.00	96.97	33.33	100.00	100.00
EfficientNetB7	36.36	95.45	25.00	100.00	100.00
Xception	21.88	95.45	14.58	100.00	100.00

TABLE 7. Maximum classification scores achieved for the different types of lesions.

Class	F1_score (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
Normal	79.12	88.39	92.86	93.75	73.47
Calcification	55.88	92.42	79.17	100.00	50.00
Asymmetry	46.15	95.45	43.75	100.00	52.63
Circumscribed masses*	40.00	93.94	43.75	100.00	100.00
Architectural distortion	40.00	95.45	47.92	100.00	100.00
Spiculated masses	39.18	93.94	39.58	100.00	50.00
Ill-defined masses**	36.54	95.45	39.58	100.00	50.00

*Well-defined. **Other

As mentioned above, the models were run by randomly selecting data. Therefore, the box-and-whisker plots corresponding to the different obtained scores are shown in Figures 13 and 14. In addition, the difference between the datasets with data augmentation and without the proposed data augmentation approach is also presented.

Figure 13a shows the distribution of the accuracy metric as a function of the seven CNNs. The results corroborate the

TABLE 8. Maximum classification scores achieved by the different neural networks - INbreast database.

Network	F1_score (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
ResNet50V2	73.68	97.73	79.55	100.00	100.00
VGG19	67.92	97.73	81.82	100.00	100.00
InceptionResNetV2	65.26	97.72	70.45	100.00	100.00
DenseNet121	60.00	97.70	56.82	100.00	100.00
InceptionV3	55.70	97.70	50.00	100.00	100.00
Xception	50.00	97.68	33.33	100.00	100.00
EfficientNetB7	42.09	97.68	32.66	100.00	100.00

TABLE 9. Maximum classification scores achieved for the different types of bi-rads abnormalities.

Class*	F1_score (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
1	73.68	90.63	79.55	100.00	100.00
5	27.03	97.65	15.63	100.00	100.00
3	26.67	96.47	18.75	100.00	100.00
6	26.67	88.67	18.75	100.00	100.00
4a	15.38	97.65	9.38	100.00	42.86
2	11.43	94.12	6.25	100.00	100.00
4b	11.43	94.12	6.25	100.00	100.00
4c	6.06	88.24	3.13	100.00	100.00

*BI-RADS

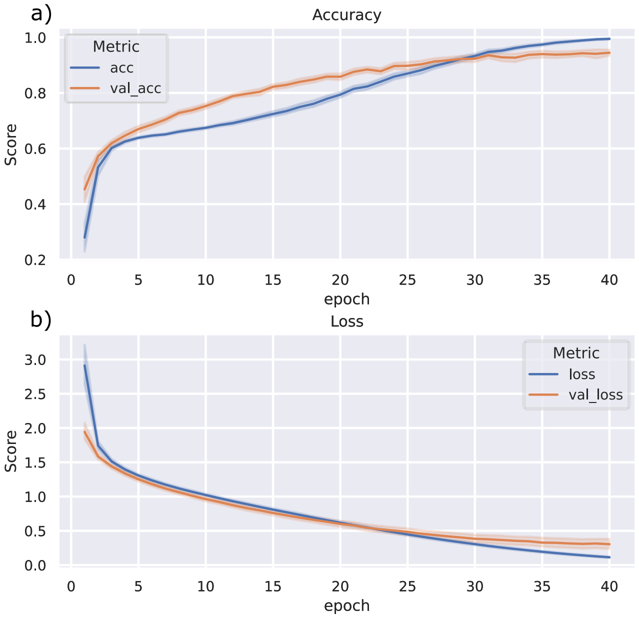

FIGURE 11. **Training and validation of the VGG19 network as a function of the number of epochs. The graphs show the network performance through the a) accuracy and b) loss. The network was trained to classify the types of lesions in the MIAS database.**

fact that there is was higher probability of arriving at a high-performance network through VGG19 than through other networks for the lesion type classification case. Moreover, it is more than evident that an increase in data contributes to better results for all networks. Indeed, this behavior can be seen in both the sensitivity (Figure 13b) and specificity (Figure 13c) metrics.

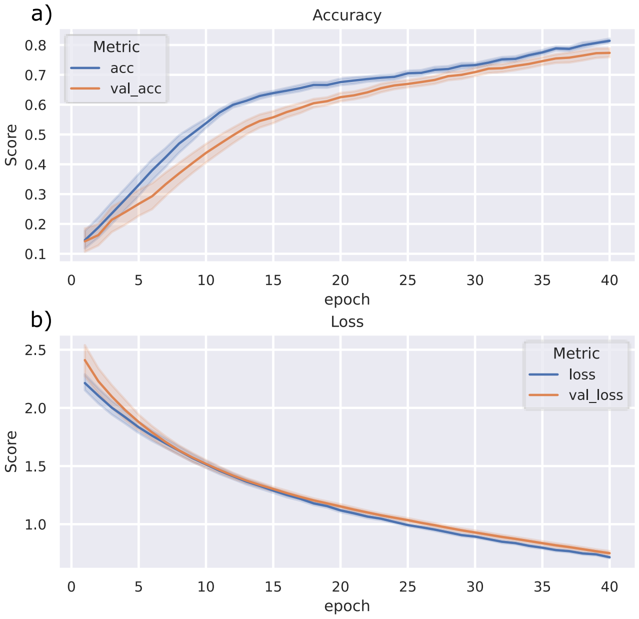

FIGURE 12. **Training and validation of the ResNet50 network as a function of the number of epochs. The graphs show the network performance through the a) accuracy and b) loss. The network was trained to classify the lesions in BI-RADS with the INbreast database.**

In Figure 13d, it can be seen once more that all classes exhibited high specificity, i.e., the models were highly efficient in identifying true negatives. In contrast, the detection rate of true positives declined significantly. Again, this trend can be attributed to the low chance of finding a true positive versus the high probability of finding a true negative. Furthermore, in this plot, it can be observed that the normal class yielded the three highest metrics, confirming that the class with the highest number of data (see Table 1) resulted in better discrimination for the networks.

The classification by degree of suspicion (BI-RADS) behavior similarly to the classification by the degree of the lesion. Figure 14a shows that the increase in data generated a better score distribution, reaching values close to 1, i.e., 100% probability. Likewise, the networks that achieved the best performance were the VGG19 network and the ResNet50 network, where the latter produced the highest scores for classifying the degree of suspicion. Again, specificity was observed to be the highest metric even without data augmentation, and it was difficult to see significant differences in most networks except for the VGG19 network (Figure 14c). In other words, the performance of the networks is subject to their sensitivity. That is, most of the networks managed to clearly identify the true negatives, as these were found in higher proportion, but were limited to incorrectly identifying the true positives (sensitivity), as shown in Figure 14b. We highlight that this effect could have been generated due to the distribution of the data.

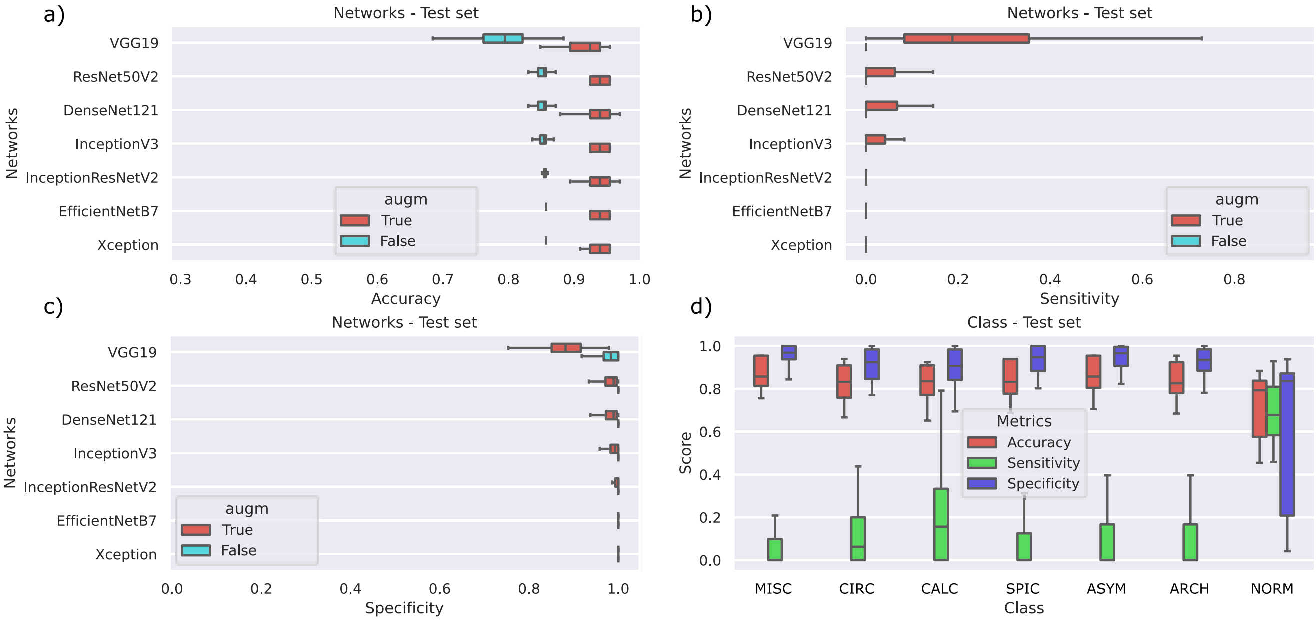

FIGURE 13. Distributions of scores generated by training all of the seven deep CNNs for the metrics of a) accuracy; b) sensitivity; and c) specificity; d) Distributions of scores as functions of lesions in terms of the same three metrics. The results were derived from the MIAS database and were generated with test data only.

Furthermore, since there were several classes, the possibility of finding a true positive of a class decayed in proportion to the number of classes being classified. That is, in this particular case, there were eight different classes. Therefore, the probability of finding a true positive of class 1 was 1/8. Additionally, this problem becomes more acute when the classes do not have the same numbers of images, i.e., when the data are unbalanced.

The accuracy, sensitivity, and specificity were maintained for the eight different classes. All classes yielded specificity distributions close to 100%, and accuracy produced high values. However, the distributions declined for sensitivity except for that of class 1.

In the classification by lesions scenario, although the class with the highest number of data presented the best results, in this case, this characteristic was not preserved. Table 1 shows that the benign class (3) had a higher number of images, but the distribution of this class was similar to those of the classes with less data (see Figure 14d).

Finally, to describe the performance of the running networks, Figure 15 shows the average classification time per subject for each of the seven utilized networks. In this case, the ResNet50 network required the shortest classification time; this was in agreement with the general descriptions of the residual connection networks, which improve their training times by forcing the learning process to follow a residual mapping $\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{x}$ and being easier to train if the ideal residual mapping is the identity function $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}$ [58]. Similarly, the other networks also had relatively short run times for classifying subjects, with most utilizing below 200 milliseconds of classification time.

It is also worth noting that all models had similar behaviors in the two classification cases. In fact, the ResNet50 network presented the best execution time, while the EfficientNet network generated higher execution times in the two cases.

Finally, as previously mentioned, this study focused on a comparative analysis of different CNNs implemented with conventional images. The approach sought to determine the behaviors of state-of-the-art DL networks in cases with medical images, specifically mammograms. The results showed the high effectiveness of EfficientNet regarding the segmentation of breast lesions, even in small lesions, despite the inherent constraints of small lesions. In addition, the study revealed the need to search for the network that best fits the specific task, i.e., although the MultiResUNet network is one of the newer architectures for segmentation, the performance metrics remained below those of EfficientNet. The review of the state-of-the-art approaches uncovered new elements for the case of breast lesions. However, most of the networks used with conventional images were shown to generate good results without the need for significant modifications except for the hyperparameters used during training. Even EfficientNet managed to the surpass state-of-the-art methods in terms of the segmentation of breast lesions. In the same sense, in the lesion classification and degree of suspicion tasks, the state-of-the-art networks did not generate the best performances. In fact, in this particular case with mammograms, the optimal classification results were obtained with the VGG19 network, even though this was one of the first deep CNNs to be developed.

Therefore, among the different types of available networks, it is necessary to test them to accurately establish the network

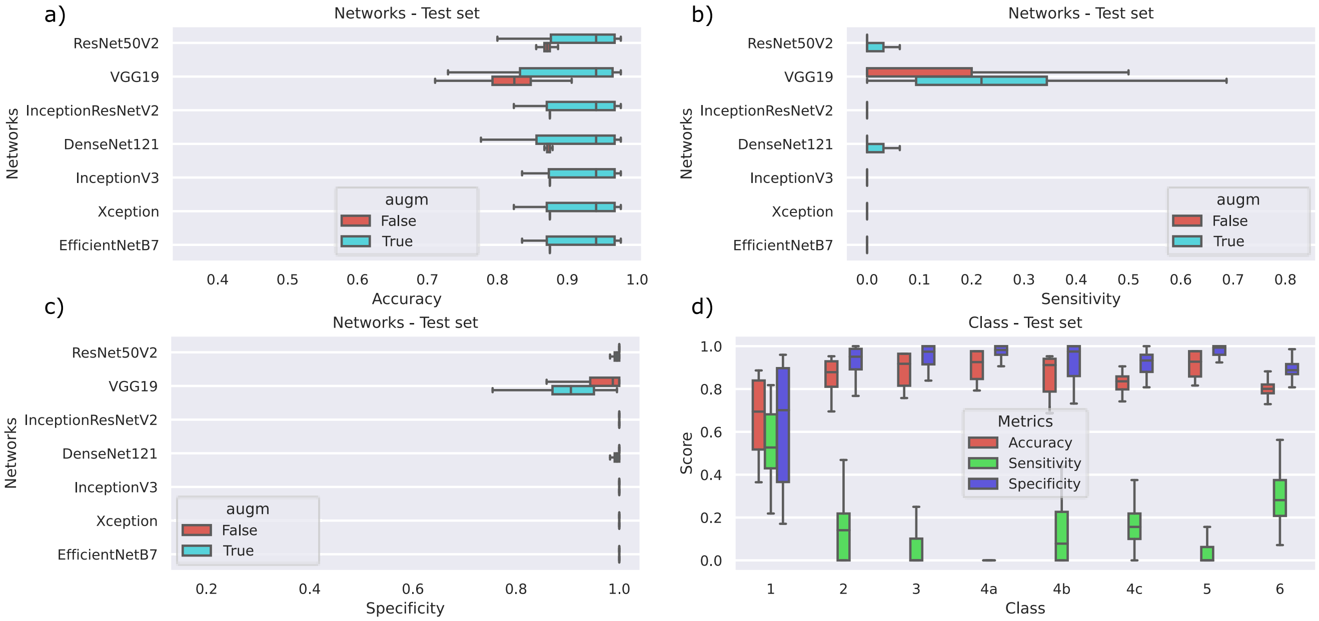

FIGURE 14. Distributions of scores generated by training all of the seven deep CNNs for the metrics of a) accuracy; b) sensitivity; and c) specificity; d) Distributions of scores as functions of BI-RADS in terms of the same three metrics. The results were derived from the INbreast database and were generated with test data only.

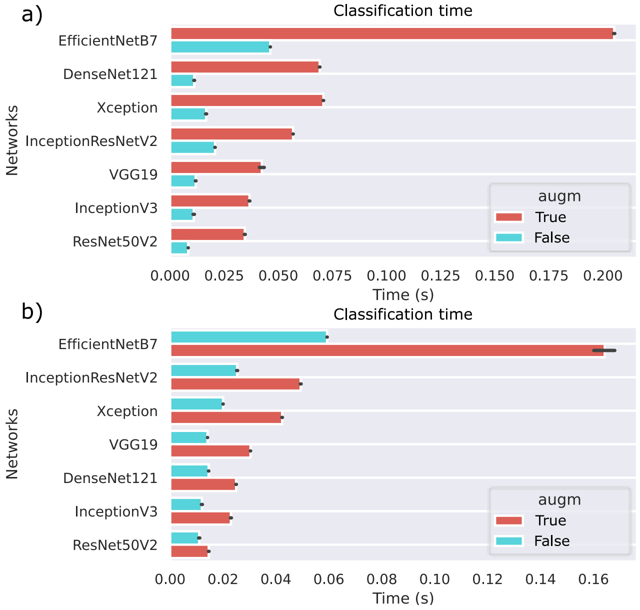

FIGURE 15. Classification time per subject for the seven deep CNN models in the a) classification by lesion and b) classification by BI-RADS tasks.

that is best suited for the specific task encountered in medical imaging.

The CNNs performed well in the segmentation and classification tasks, surpassing the state-of-the-art methods. However, the study presents some limitations that should be addressed in future studies. Initially, the databases remained one of the main limitations in the implementation of the DL algorithms. In this case, the CBIS-DDSM, MIAS, and INbreast databases, three of the main open-access databases in mammography, were used. However, all three databases lack heterogeneity, and each has features that address different problems. In other words, it is necessary to attempt to evaluate the results with external databases containing the same labels or segmentations of the same regions to reach an objective conclusion. For example, Salama and Aly [50] addressed the segmentation problem. However, their research focused on the segmentation of the breast and not lesions.

On the other hand, as mentioned above, the accuracy metric is not the most suitable for unbalanced data, as it can generate high values despite having very low sensitivity. To avoid this drawback, this study was performed with different evaluation metrics. However, a comparison with previous work evaluated with the accuracy metric might not reveal significant differences.

V. CONCLUSION

A comparative analysis of methods for the segmentation and classification of breast lesions on digital mammograms was proposed. Initially, we proposed a comparative analysis of 12 state-of-the-art DL networks under five loss functions to improve the automatic segmentation of breast examination images. The proposed convolutional models were built with the base UNet and the most recently developed networks with building blocks, squeeze-and-excitation blocks, residual connections, large numbers of deep layers, and novel architectures for segmentation or conventional classification, i.e., on problems other than medical imaging. The results showed that EfficientNet, together with the binary cross-entropy loss function, achieved an accuracy of 99.96%, outperforming the most recently developed approaches.

Additionally, this model presented the most homogeneous distribution with higher scores than those of the other architectures. EfficientNet generated training and validation curves that converged with a Dice score close to 95%, indicating that the model was not overtrained. The architecture was validated with test datasets of different sizes, where the generated segmentations had areas with sizes close to those of real areas, even for minor lesions. Similarly, in the segmentation tests, it was possible to observe the details generated at the edges of the lesions, demonstrating the high effectiveness of EfficientNet. The model's effectiveness provides a detailed view of the morphological characteristics of breast masses, allowing their structures to be compared with theoretical bases for an objective assessment of the morphological aggressiveness of the masses and, consequently, allowing for the pathological characterization of the masses as potentially malignant or benign masses.

On the other hand, the comparative method with Dice's coefficient and the morphological transformations of the real segmentation images allowed for observing the effect that the discretization of the images had on the segmentation results of small regions, i.e., the validation metrics lose objectivity when the segmented element is smaller.

In this sense, the results are promising with EfficientNet. However, new methods need to be explored to reduce the error associated with the metrics when lesions are small. Future work should consider this variable and focus on small lesions, increasing model performance and ensuring an efficient model that is applicable in the clinical setting.

Finally, a comparative analysis was performed for the classification process through two different databases: MIAS and INbreast. The images were classified by the types of lesions and the degree of suspicion of malignancy. In the first case, the classification by lesion type yielded 96.97% accuracy. However, the results were subject to the distribution of the input data because there was a greater probability of finding true negatives than true positives, which is a critical limitation of the classification process. Similarly, the lesion type classification case produced 97.73% accuracy but was limited by identical drawbacks due to the distribution of the data across classes. The results showed that although new developments and techniques have emerged in the architectures of DL models, it is necessary to explore different networks to arrive at the one that best fits the desired task. For example, in this case, we obtained the best results through the VGG19 and ResNet50 networks, where the former is one of the oldest DL networks available.

ACKNOWLEDGMENT

The authors would like to thank Laura Salazar (M.D. in radiology) for her comments, insightful suggestions, and careful reading of the manuscript.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA, Cancer J. Clinicians*, vol. 69, no. 1, pp. 7–34, 2019, doi: 10.3322/caac.21551.

[2] American Cancer Society. (2017). *Breast Cancer What is Breast Cancer*. Accessed: Oct. 19, 2021. [Online]. Available: http://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html

[3] American Cancer Society. (2020). *Breast Cancer Facts & Figures 2019–2020*. Atlanta. Accessed: Aug. 31, 2021. [Online]. Available: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2019-2020.pdf

[4] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, A Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[5] S. Anttila and P. Boffetta, "Occupational cancers," in *Occupational Cancers*. Cham, Switzerland: Springer, 2020, pp. 1–640, doi: 10.1007/978-3-030-30766-0.

[6] World Health Organization. (Dec. 2020). *Breast Cancer Now Most Common Form of Cancer: WHO Taking Action*. Accessed: Aug. 31, 2021. [Online]. Available: https://www.who.int/news/item/03-02-2021-breast-cancer-now-most-common-form-of-cancer-who-taking-action

[7] American Cancer Society. (2019). *Breast Cancer Risk and Prevention*. Accessed: Aug. 31, 2021. [Online]. Available: https://www.cancer.org/content/dam/CRC/PDF/Public/8578.00.pdf

[8] O. Akin, S. B. Brennan, D. D. Dershaw, M. S. Ginsberg, M. J. Gollub, H. Schöder, D. M. Panicek, and H. Hricak, "Advances in oncologic imaging," *CA, A Cancer J. Clinicians*, vol. 62, no. 6, pp. 364–393, Nov. 2012, doi: 10.3322/caac.21156.

[9] S. R. Shrivastava, P. S. Shrivastava, and J. Ramasamy, "Self breast examination: A tool for early diagnosis of breast cancer," *Amer. J. Public Health Res.*, vol. 1, no. 6, pp. 135–139, Jun. 2013, doi: 10.12691/ajphr-1-6-2.

[10] S. Vara, M. K. Karnena, and B. K. Dwarapureddi, "Epidemiology of cancers in women," in *A Theranostic and Precision Medicine Approach for Female-Specific Cancers*. Amsterdam, The Netherlands: Elsevier, 2021, pp. 71–90, doi: 10.1016/B978-0-12-822009-2.00004-2.

[11] ASCO. (2021). *Breast Cancer: Statistics*. American Society of Clinical Oncology. Accessed: Oct. 19, 2021. [Online]. Available: https://www.cancer.net/cancer-types/breast-cancer/statistics

[12] M. Kamińska, T. Ciszewski, K. Łopacka-Szatan, P. Miotła, and E. Starosławska, "Breast cancer risk factors," *Menopausal Rev.*, vol. 3, pp. 196–202, Sep. 2015, doi: 10.5114/pm.2015.54346.

[13] S. Winters, C. Martin, D. Murphy, and N. K. Shokar, "Breast cancer epidemiology, prevention, and screening," *Approaches to Understanding Breast Cancer*, vol. 151. New York, NY, USA: Academic, 2017, pp. 1–32, doi: 10.1016/bs.pmbts.2017.07.002.

[14] A. Barragán-Montero, U. Javaid, G. Valdés, D. Nguyen, P. Desbordes, B. Macq, S. Willems, L. Vandewinckele, M. Holmström, F. Löfman, S. Michiels, K. Souris, E. Sterpin, and J. A. Lee, "Artificial intelligence and machine learning for medical imaging: A technology review," *Phys. Medica*, vol. 83, pp. 242–256, Mar. 2021, doi: 10.1016/j.ejmp.2021.04.016.

[15] S. H. Jafari, Z. Saadatpour, A. Salmaninejad, F. Momeni, M. Mokhtari, J. S. Nahand, M. Rahmati, H. Mirzaei, M. Kianmehr, "Breast cancer diagnosis: Imaging techniques and biochemical markers," *J Cell Physiol*, vol. 233, no. 7, pp. 5200–5213, Jul. 2018, doi: 10.1002/jcp.26379.

[16] H. Kasban, M. El-Bendary, and D. Salama. (2015). *A Comparative Study of Medical Imaging Techniques*. [Online]. Available: https://www.semanticscholar.org/paper/A-Comparative-Study-of-Medical-Imaging-Techniques-Kasban-El-Bendary/699d2913a6e723ab9cf57bd46dc69865a6e937b1#citing-papers

[17] D. Paez, F. Giammarile, and P. Orellana, "Nuclear medicine: A global perspective," *Clin. Transl. Imag.*, vol. 8, no. 2, pp. 51–53, Apr. 2020, doi: 10.1007/s40336-020-00359-z.

[18] G. I. Balali, "Breast cancer: A review of mammography and clinical breast examination for early detection of cancer," *Open Access Library J.*, vol. 7, no. 10, p. 1, 2020, doi: 10.4236/oalib.1106866.

[19] A. L. W. Meisner, M. H. Fekrazad, and M. E. Royce, "Breast disease: Benign and malignant," *Med. Clinics North Amer.*, vol. 92, no. 5, pp. 1115–1141, Sep. 2008, doi: 10.1016/j.mcna.2008.04.003.

[20] American Cancer Society. (2019). *Breast Cancer Early Detection and Diagnosis*. Accessed: Oct. 19, 2021. [Online]. Available: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection.html

[21] T.-A. Moo, R. Sanford, C. Dang, and M. Morrow, "Overview of breast cancer therapy," *PET Clinics*, vol. 13, no. 3, pp. 339–354, Jul. 2018, doi: 10.1016/j.cpet.2018.02.006.

[22] S. W. Duffy, D. Vulkan, H. Cuckle, D. Parmar, S. Sheikh, R. A. Smith, A. Evans, O. Blyuss, L. Johns, I. O. Ellis, and J. Myles, "Effect of mammographic screening from age 40 years on breast cancer mortality (UK age trial): Final results of a randomised, controlled trial," *Lancet Oncol.*, vol. 21, no. 9, pp. 1165–1172, 2020, doi: 10.1016/S1470-2045(20)30398-3.

[23] A. Coldman, N. Phillips, C. Wilson, K. Decker, A. M. Chiarelli, J. Brisson, B. Zhang, J. Payne, G. Doyle, and R. Ahmad, "Pan-Canadian study of mammography screening and mortality from breast cancer," *JNCI, J. Nat. Cancer Inst.*, vol. 106, no. 11, Nov. 2014, Art. no. dju261, doi: 10.1093/jnci/dju261.

[24] C. E. DeSantis, J. Ma, A. G. Sauer, L. A. Newman, and A. Jemal, "Breast cancer statistics, 2017, racial disparity in mortality by state: Breast cancer statistics, 2017," *CA, A Cancer J. for Clinicians*, vol. 67, no. 6, pp. 439–448, Nov. 2017, doi: 10.3322/caac.21412.

[25] J. Katzen and K. Dodelzon, "A review of computer aided detection in mammography," *Clin. Imag.*, vol. 52, pp. 305–309, Nov. 2018, doi: 10.1016/j.clinimag.2018.08.014.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034, doi: 10.1109/ICCV.2015.123.

[27] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, "Artificial intelligence in radiology," *Nature Rev. Cancer*, vol. 18, no. 8, pp. 500–510, 2018, doi: 10.1038/s41568-018-0016-5.

[28] S. Li, M. Dong, G. Du, and X. Mu, "Attention dense-U-Net for automatic breast mass segmentation in digital mammogram," *IEEE Access*, vol. 7, pp. 59037–59047, 2019, doi: 10.1109/access.2019.2914873.

[29] Z. Monostori, P. G. Herman, D. P. Carmody, T. M. Eacobacci, N. R. Capece, V. M. Cruz, S. Gentin, and F. M. Vernace, "Limitations in distinguishing malignant from benign lesions of the breast by systematic review of mammograms," *Surg. Gynecol. Obstet.*, vol. 173, no. 6, pp. 438–442, Dec. 1991.

[30] J. Gill, A. Girdhar, and T. Singh, "A review of enhancement and segmentation techniques for digital images," *Int. J. Image Graph.*, vol. 19, no. 3, pp. 1–23, 2019, doi: 10.1142/S021946781950013X.

[31] A. Bali and S. N. Singh, "A review on the strategies and techniques of image segmentation," in *Proc. 5th Int. Conf. Adv. Comput. Commun. Technol.*, Feb. 2015, pp. 113–120, doi: 10.1109/ACCT.2015.63.

[32] D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G. Z. Yang, "Deep learning for health informatics," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 1, pp. 4–21, Jan. 2017, doi: 10.1109/JBHI.2016.2636665.

[33] S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: A review," *Artif. Intell. Rev.*, vol. 54, no. 1, pp. 137–178, Jan. 2021, doi: 10.1007/s10462-020-09854-1.

[34] G. W. Lindsay, "Convolutional neural networks as a model of the visual system: Past, present, and future," 2020, *arXiv:2001.07092*.

[35] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[36] L. D. Jones, D. Golan, S. A. Hanna, and M. Ramachandran, "Artificial intelligence, machine learning and the evolution of healthcare: A bright future or cause for concern?" *Bone Joint Res.*, vol. 7, no. 3, pp. 223–225, Mar. 2018, doi: 10.1302/2046-3758.73.BJR-2017-0147.R1.

[37] A. Ibrahim, P. Gamble, R. Jaroensri, M. M. Abdelsamea, C. H. Mermel, P.-H.-C. Chen, and E. A. Rakha, "Artificial intelligence in digital breast pathology: Techniques and applications," *Breast*, vol. 49, pp. 267–273, Feb. 2020, doi: 10.1016/j.breast.2019.12.007.

[38] B. Norgeot, B. S. Glicksberg, and A. J. Butte, "A call for deep-learning healthcare," *Nature Med.*, vol. 25, no. 1, pp. 14–15, Jan. 2019, doi: 10.1038/s41591-018-0320-3.

[39] F. Winsberg, M. Elkin, J. Macy, V. Bordaz, and W. Weymouth, "Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis," *Radiology*, vol. 89, no. 2, pp. 211–215, Aug. 1967, doi: 10.1148/89.2.211.

[40] Y. Gao, K. J. Geras, A. A. Lewin, and L. Moy, "New frontiers: An update on computer-aided diagnosis for breast imaging in the age of artificial intelligence," *Amer. J. Roentgenol.*, vol. 212, no. 2, pp. 300–307, Feb. 2019, doi: 10.2214/AJR.18.20392.

[41] L. Abdelrahman, M. Al Ghamdi, F. Collado-Mesa, and M. Abdel-Mottaleb, "Convolutional neural networks for breast cancer detection in mammography: A survey," *Comput. Biol. Med.*, vol. 131, Apr. 2021, Art. no. 104248, doi: 10.1016/j.compbiomed.2021.104248.

[42] W. Zhou, G. Lv, and L. Wang, "An automatic breast mass segmentation algorithm in digital mammography," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Oct. 2017, pp. 1–5, doi: 10.1109/ICSPCC.2017.8242445.

[43] B. Sadeghi, M. Karimi, and S. Mazaheri, "Automatic suspicions lesions segmentation based on variable-size Windows in mammography images," *Health Technol.*, vol. 11, no. 1, pp. 99–110, Jan. 2021, doi: 10.1007/s12553-020-00506-6.

[44] A. M. Salih and M. Y. Kamil, "Mammography image segmentation based on fuzzy morphological operations," in *Proc. 1st Annu. Int. Conf. Inf. Sci. (AiCIS)*, Nov. 2018, pp. 40–44, doi: 10.1109/AiCIS.2018.00020.

[45] M. Kamil and A. Salih, "Mammography images segmentation via fuzzy C-mean and K-mean," *Int. J. Intell. Eng. Syst.*, vol. 12, no. 1, pp. 22–29, Feb. 2019, doi: 10.22266/IJIES2019.0228.03.

[46] D. Abdelhafiz, S. Nabavi, R. Ammar, C. Yang, and J. Bi, "Residual deep learning system for mass segmentation and classification in mammography," in *Proc. 10th ACM Int. Conf. Bioinf., Comput. Biol. Health Informat.*, Sep. 2019, pp. 475–484, doi: 10.1145/3307339.3342157.

[47] D. Abdelhafiz, J. Bi, R. Ammar, C. Yang, and S. Nabavi, "Convolutional neural network for automated mass segmentation in mammography," *BMC Bioinf.*, vol. 21, no. S1, p. 192, Dec. 2020, doi: 10.1186/s12859-020-3521-y.

[48] T. de Moor, A. Rodriguez-Ruiz, R. Mann, A. Gubern Mérida, and J. Teuwen, "Automated lesion detection and segmentation in digital mammography using a U-Net deep learning network," in *Proc. 14th Int. Workshop Breast Imag. (IWBI)*, Jul. 2018, pp. 23–29, doi: 10.1117/12.2318326.

[49] W. Zhu, X. Xiang, T. D. Tran, G. D. Hager, and X. Xie, "Adversarial deep structured nets for mass segmentation from mammograms," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 847–850, doi: 10.1109/ISBI.2018.8363704.

[50] W. M. Salama and M. H. Aly, "Deep learning in mammography images segmentation and classification: Automated CNN approach," *Alexandria Eng. J.*, vol. 60, no. 5, pp. 4701–4709, Oct. 2021, doi: 10.1016/j.aej.2021.03.048.

[51] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, and L. Tarbox, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1045–1057, 2013, doi: 10.1007/s10278-013-9622-7.

[52] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Sci. Data*, vol. 4, no. 1, pp. 1–9, Dec. 2017, doi: 10.1038/sdata.2017.177.

[53] R. S. Lee, F. Gimenez, A. Hoogi, and D. Rubin, "Curated breast imaging subset of DDSM [dataset]," *Cancer Imag. Arch.*, 2016, doi: 10.7937/K9/TCIA.2016.7O02S9CY.

[54] J. Suckling. (1994). *The Mammographic Image Analysis Society Digital Mammogram Database*. [Online]. Available: http://peipa.essex.ac.uk/info/mias.html

[55] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "INbreast: Toward a full-field digital mammographic database," *Academic Radiol.*, vol. 19, no. 2, pp. 236–248, Feb. 2012, doi: 10.1016/j.acra.2011.09.014.

[56] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*.

[57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. ICLR Conf. Track*, Sep. 2014, pp. 1–14. [Online]. Available: http://arxiv.org/abs/1409.1556

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[59] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 4278–4284. [Online]. Available: http://arxiv.org/abs/1602.07261

[60] Y. Zhu and S. Newsam, "DenseNet for dense flow," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 790–794, doi: 10.1109/ICIP.2017.8296389.

[61] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," 2017, *arXiv:1707.03718*.

[62] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995, doi: 10.1109/CVPR.2017.634.

[63] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.

[64] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, Jun. 2019, pp. 10691–10700. [Online]. Available: https://arxiv.org/abs/1905.11946v5

[65] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[66] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, Jan. 2020, doi: 10.1016/j.neunet.2019.08.025.

[67] S. Jadon, "A survey of loss functions for semantic segmentation," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Oct. 2020, pp. 1–7, doi: 10.1109/CIBCB48159.2020.9277638.

[68] Y.-d. Ma, Q. Liu, and Z.-b. Quan, "Automated image segmentation using improved PCNN model based on cross-entropy," in *Proc. Int. Symp. Intell. Multimedia, Video Speech Process.*, Oct. 2004, pp. 743–746, doi: 10.1109/ISIMP.2004.1434171.

[69] V. Pihur, S. Datta, and S. Datta, "Weighted rank aggregation of cluster validation measures: A Monte Carlo cross-entropy approach," *Bioinformatics*, vol. 23, no. 13, pp. 1607–1615, Jul. 2007, doi: 10.1093/bioinformatics/btm158.

[70] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2017, pp. 240–248, doi: 10.1007/978-3-319-67558-9_28.

[71] A. Tversky, "Features of similarity," *Psychol. Rev.*, vol. 84, no. 4, pp. 327–352, 1977, doi: 10.1037/0033-295X.84.4.327.

[72] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, Jun. 2017, pp. 379–387, doi: 10.1007/978-3-319-67389-9_44.

[73] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.

ANDRÉS ANAYA-ISAZA received the degree in systems engineering from Universidad Cooperativa de Colombia, in 2009, the first master's degree in computer engineering from Atlantic International University, in 2012, with a focus on algorithmics, and the second master's degree in systems and computer engineering from the Technological University of Pereira, Colombia, in 2017, with a focus on machine learning. He is currently pursuing the Ph.D. degree in bioengineering and artificial intelligence with Pontificia Universidad Javeriana. He also works as the VP of research and innovation with the Indigo Technologies.

LEONEL MERA-JIMÉNEZ was born in El Tambo, Nariño, Colombia, in April 1990. He received the degree in physical engineering from Universidad del Cauca, Colombia, in 2015. He is currently pursuing the master's degree in engineering emphasizing bioengineering with the University of Antioquia, Colombia. He is also working as a Machine Learning Researcher with Indigo Technologies. His main interests are computer vision, machine learning techniques, deep learning, image processing, and biomedical imaging applications.

JOHAN MANUEL CABRERA-CHAVARRO was born in Neiva, Colombia, in 1995. He received the degree in software development engineering from Universidad Surcolombiana, Neiva, in 2020. He currently works as a Junior Developer with Indigo Technologies. His main research interests include software development, interface design, image processing, artificial intelligence, and medical applications.

LORENA GUACHI-GUACHI received the B.Sc. degree in systems engineering from the Escuela Superior Politécnica de Chimborazo, Ecuador, in 2009, and the Ph.D. degree in science and technologies of complex systems from the University of Calabria, Italy, in 2017. After her bachelor studies, she worked as a Software Developer with Smartwork S.A., Quito, Ecuador, and the Systems Manager with Produbanco S.A., Quito, until 2013. In 2016, she did her doctoral internship at the Computer Vision Research Group, University of Amsterdam, The Netherlands. From 2017 to 2020, she worked as a Researcher–Lecturer with the School of Mathematical and Computational Sciences, Yachay Tech University, Ecuador. She is currently a Researcher with The BioRobotics Institute, Scuola Superiore Sant'Anna, Italy, and a Lecturer with the Department of Mechatronics, Universidad Internacional del Ecuador.

DIEGO PELUFFO-ORDÓÑEZ was born in Pasto, Colombia, in 1986. He received the degree in electronic engineering and the M.Eng. and Ph.D. degrees from the Universidad Nacional de Colombia, Manizales, Colombia, in 2008, 2010, and 2013, respectively. In 2012, he undertook his doctoral internship at KU Leuven, Leuven, Belgium. From 2013 to 2014, he worked as a Postdoctoral Researcher with Université Catholique de Louvain, Louvain-la-Neuve, Belgium. From 2014 to 2015, he worked as an Assistant Teacher with the Universidad Cooperativa de Colombia, Pasto, Colombia. From 2015 to 2017, he worked as a Researcher/Professor with Universidad Técnica del Norte, Ecuador. From 2017 to 2020, he worked as a Professor with the School of Mathematical and Computational Sciences, Yachay Tech University, Ecuador. He is currently working as an Assistant Professor with the Modeling, Simulation and Data Analysis (MSDA) Research Program, Mohammed VI Polytechnic University, Morocco. Also, he works as a Consultant/Curriculum Author with DeepLearning.AI. He is also the Head and the Founder of the SDAS Research Group. He is also an External Collaborator with the Writing Lab, Tecnológico de Monterrey, Mexico. As well, he is an External Supervisor of Ph.D. programs with the Universidad de Granada, Spain; the Universitat Politécnica de Valéncia, Spain; and the Universidad Nacional de La Plata, Argentina. His main research interests include kernel-based and spectral methods for data clustering and dimensionality reduction. The scope of his topics of interest encompasses complex high-dimensional data, signal, image, and video analysis for medical and industry applications. He has served as an organizing committee member (the general chair, the session chair, and the competitions chair) and a keynote speaker for several conferences. Also, he has served as a Guest Editor for the *Computers and Electrical Engineering* journal.

JORGE IVAN RIOS-PATIÑO received the degree in industrial engineering from the Universidad Tecnológica de Pereira, Colombia, in 1976, and the master's degree in informatics and the master's degree in knowledge engineering from the Universidad Politécnica de Madrid, in 1988 and 1992, respectively. He is currently the Director of the master's program in systems and computer engineering with the Universidad Tecnológica de Pereira.