IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier x/ACCESS.x.Doi Number

Breast lesion classification in ultrasound images using deep convolutional neural network

Bashir Zeimarani[1], Marly G. F. Costa[1], Nilufar Z. Nurani[2], Sabrina Ramos Bianco[2], Wagner C. A. Pereira[3] and Cicero F. F. Costa Filho[1]

[1]Federal University of Amazonas, Manaus, AM 69080-900, Brazil
[2]Fundação Centro de Controle de Oncologia do Estado do Amazonas– FCECON, Setor de Imagem, Manaus, Brazil
[3] Biomedical Engineering Program/COPPE/Federal University of Rio de Janeiro.

Corresponding author: Cicero F. F. Costa Filho (e-mail: ccosta@ ufam.edu.br).

ABSTRACT In recent years, convolutional neural networks (CNNs) have found many applications in medical image analysis. Having enough labeled data, CNNs could be trained to learn image features and used for object localization, classification, and segmentation. Although there are many interests in building and improving automated systems for medical image analysis, lack of reliable and publicly available biomedical datasets makes such a task difficult. In this work, the effectiveness of CNNs for the classification of breast lesions in ultrasound (US) images will be studied. First, due to a limited number of training data, we use a custom-built CNN with a few hidden layers and apply regularization techniques to improve the performance. Second, we use transfer learning and adapt some pre-trained models for our dataset. The dataset used in this work consists of a limited number of cases, 641 in total, histopathologically categorized (413 benign and 228 malignant lesions). To assess how the results of our classifier generalize on our data set, a 5-fold cross-validation were employed, where in each fold 80% of data were used for training and the 20% for testing. Accuracy and the area under the ROC curve (AUC) were used as the main performance metrics. Before applying any regularizations techniques, we achieved an overall accuracy of 85.98% for tumor classification, and the AUC equal to 0.94. After applying image augmentation and regularization, the accuracy and the AUC increased to 92.05% and 0.97, respectively. Using a pre-trained model, we achieved an overall accuracy of 87.07% and an AUC equal to 0.96. The obtained results demonstrated the effectiveness of our custom architecture for classification of tumors in this small US imaging dataset, surpassing some traditional learning algorithm based on manual feature selection.

INDEX TERMS Breast Tumor, Ultrasound Images, Convolutional Neural Network, Transfer Learning.

I. INTRODUCTION

According to World Cancer Report, 2014, Breast cancer has a very high incident rate, 43.3%, and is one of the main causes of cancer death among young women [1]. Despite the high death rate of 25.5 [2], a study showed that the early detection and treatment of malignant breast tumors accounted for 38% decrease in mortality rate from 1989 to 2014 [1].

Ultrasound (US) is one of the main procedures to diagnose breast lesion [3]. Although Digital Mammography (DM) is the most effective technique [3], US has the advantages of being safer, more cost effective, and sensitive to tumors located in dense areas [4]. In recent years, many attempts have been made to automate the diagnosis procedures and minimize the operator dependency of US imaging [5]. Scientists applied a variety of algorithms and employed Computer-Aided Diagnosis (CAD) tools for localization and classification of breast lesions. To name a few, in [6], the authors used the wavelet filters to reduce the noise in US images and applied the Adaptive Gradient Descent algorithm for classification of lesions. In [7], a set of features from US images, each rated by a radiologist, were selected to from a feature matrix, the matrix was then fed to a biclustring algorithm and a back-propagation neural network used to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3010863, IEEE Access

classify each lesions. In [8] and [9], the authors employed Convolutional Neural Network (CNN) and Transfer learning for breast lesion classification in US images, in each case the size of dataset was increased by applying image augmentation, then the dataset was split to form a training and evaluation set, the training set was used to fine-tune a set of well-known CNN architectures, and the evaluation set for measuring the performance. Using the CNN in these cases eliminated the need for manual feature selection, done in other works such as [6] or [7].

In [9] and [10] the authors employed CNN for localization of lesions in breast US images. In both, the performances of different CNN architectures were compared against other machine learning methods. Using CNN resulted in an overall improvement of tumor localization, compared to other machine learning algorithms.

According to [11], in medical image analysis, a very high sensitivity value and an AUC more than 0.95 are sought after, considering the vast popularity and success of CNNs in detection and classification of objects. Naturally, the following questions arise: Are CNNs effective when dealing with a relatively small dataset? Can a CNN architecture outperform traditional machine learning techniques in the classification of breast tumors in US imaging? To answer these questions, we propose a new CNN architecture for the classification of breast tumors in Ultrasound images. The results obtained by these networks will be compared to results from some traditional machine learning algorithms obtained from [12], which uses the same dataset and also compare with other well-known CNN architectures, using transfer learning.

As mentioned, the primary objective of this work is to propose a simple CNN architecture and test it effectiveness for classification of images in a US image dataset. As for the specific objects, the followings can be listed:

- Automatic feature selection using a custom build CNN architecture.
- Utilization of different optimizers and regularization techniques to increase the classifier performance.
- Using transfer learning to compare the performance of the custom-build network against other well-known architectures.
- Comparing the obtained results with other traditional machine learning methods, employing the same dataset.

The remainder of this paper is organized as follows: In Section II a review of some related work will be presented; Section III introduces some relevant theoretical backgrounds; Section IV introduces the dataset and the hardware specification used for implementation of the algorithm; Section V gives a detailed description of the methodology, including the image preprocessing techniques, network architecture, training parameters, evaluating metrics and comparison methods; Sections VI and VII present and evaluate the results obtained by the given methodology and in Section VIII the paper concludes.

II. Literature review

The Convolutional Neural Networks (CNNs) have evolved and been employed in diverse areas of research [13] such as computer vision applications, document analysis, study of natural phenomena, speech recognition, advertisement, etc.

In this section we highlight a number of related works, applying machine learning algorithms in classification and segmentation of lesions in breast US images.

In [8], the authors employed Transfer Learning to adapt the ResNet50 network, pretrained with COCO image dataset, to segment and classify breast lesions in US images. Although the network is capable of identifying a variety of shapes and objects, the dataset used for fine tuning of the network was very small, 303 images in total, and could not obtain a satisfactory result. To overcome this limitation a set of different image augmentation techniques were employed, first, by applying image rotations and flips they doubled the number of training set , then they employed an image augmentation technique, called Multi-Scale Super-Pixel Elastic (MSSPE) to further increase the number of training set. Classification rate (CR), True Positive Rate (TPR), and True negative Rate (TNR) were used as performance metrices, obtaining 80.42%, 63.64% and 87.88%, respectively.

The authors in [14] employed a dataset containing 1043 images, each classified by a radiologist, and split the dataset for training, validation and testing a different set of known CNN architectures. The average recall rate (APR), F1 score and average precision rate (APR) were used as performance metrics. Between the different network architectures (such as VGG16, YOLO and SD300+ZFNet), the SD300+ZFNet had the best overall performance, obtaining 96.89, 67.23 and 79.38 for APR, ARR, and F1-score, respectively.

Hijab in [15] employed Transfer Learning for fine tuning of the VGG16 network and to classify breast tumors in an US image database, containing 1300 images. To overcome overfitting, the author used image augmentation to increase the dataset and built a new one containing 21,600 images, the new dataset was further divided to 15120 images for training and the rest for testing. The new training set was used to adjust the weights on the last convolutional layer of the VGG16 network and measure the performance of the system in classification of the images in the test set. The authors achieved 0.97 and 0.98 values for accuracy and AUC value, respectively. Although the obtaining results were satisfactory, only a portion of the dataset were used for testing, in addition some of the image augmentation techniques, such as shearing, are not recommended for this kind of images [16].

In [17], the authors used transfer learning to adapt and train a known CNN network for the classification of breast tumors in ultrasound images. The dataset used in their work consists of 882 images. The dataset was split into a training set and a test set. For the training process, a matching layer was used to rescale the pixel intensities of the grayscale images and convert them to three separate RGB channels.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3010863, IEEE Access

The VGG19 net was adapted for training and, after some fine-tuning, an AUC value equal to 0.936 was achieved. They argue that this performance surpassed the classification accuracy of a radiologist readings.

In [18] the authors used transfer learning to adapt an Inception-v3 CNN, the third generation of GoogLeNet, for the classification of breast tumors in ultrasound images. The proposed CNN was trained and evaluated on 316 breast lesions (135 malignant and 181 benign). The proposed CNN achieved an AUC of 0.9468 with five-folder cross validation. The values of sensitivity and specificity were 0.886 and 0.876, respectively.

The dataset in [19] is relatively bigger than other similar works (containing 2238 cases of breast lesions), but it is not histopathologically classified and is based on the BI-RADS classification. By using transfer learning, two neural networks were used first to detect the region of interest (ROI) and then classify this region (containing the lesion) into one of the five BI-RADS categories. The two-stage framework (as the authors call it) achieved the best accuracy value of 0.998 for category 3 BI-RADS and the lowest value of 0.734 for 4B category.

Although we focused on papers using automatic feature selection, it is of particular interest to note the results yielded by [20], which utilized the same data set used in our work. The authors proposed a feature selection technique based on mutual information technique and a statistical test for breast tumor classification in ultrasound images. As the first step of the algorithm, the authors used the watershed transformation to segment the tumor area. After tumor segmentation, the tumor region was used for computing 22 morphological features, quantifying some local characteristics of the lesions. The features were ranked with mutual information using the minimal-redundancy-maximal-relevance criterion. Employing the ranked feature space, several m-dimensional feature subsets were created and were used for training of the Fisher linear discriminant analysis classifier. The AUC value was used as the performance metric. The experiments showed a similar classification performance, using only the top seven ranked features versus the whole feature set, obtaining an AUC value of 0.952. The top seven ranked features used for classification were based on convex hull, equivalent ellipse, long axis to short axis ratio, geometric and shape morphological features.

Although CNNs have great success in automatic feature extraction and classification of objects, the authors in [21] state that they cannot generalize well in the lack of enough labeled data. As we mentioned earlier there are a few publicly available annotated datasets of breast lesions in ultrasound images. Therefore, in the majority of papers, the authors trained the network with a very limited number of labeled data. In some of these works, the obtained AUC value and sensitivity are not acceptable for clinical use (i.e. AUC < 0.95), and, in others, there is no justification in using a deep neural network for classification of images using a small dataset.

One could argue that in lack of enough data, it is more convenient to use a more traditional machine learning algorithm such as SVM [22], as they are much faster to train and, in some cases, generalize better, such as the work in [20].

In this work, as in the other previously reviewed works [8,9,14, and 17], using a small database (due to limitations of annotated medical image data bases), we study the performance of CNNs for the classification of breast lesions in US images. In our work, as will be presented in results section, through the use of some regularization techniques, we improve the CNN generalization, achieving and AUC >0.95 in the test set. The results of the proposed CNN model outperform the results obtained with some traditional algorithms obtained in [20], in which the authors achieved very satisfactory results using the same database, and the performance of some pre-trained networks (containing much more complexity) using transfer learning.

This work presents the results of a master's degree thesis [23] and extends the content of a preceding article [24].

III. Some Theoretical Background

In this section, a brief review of some regularization techniques used in this work and the philosophy behind their usage will be presented.

A. Overfitting

Despite the huge popularity of neural networks, considerable challenges remain in order to provide a good level of performance. These challenges refer to some practical problems, the most important being the overfitting [13, 25-27]. The overfitting problem occurs when there is a gap between the CNN performance in the training data and in the test/validation data. The CNN presents a good performance in the training data but has a poor performance on the test/validation data. This is caused by the paucity in the training data and by the complexity of the CNN model. Increasing the size of the training data, decreases overfitting, while increasing model complexity, increases overfitting. Some techniques, as data augmentation, dropout, L_2 regularization and batch normalization can be used to reduce overfitting. Next, we will describe these techniques.

B. Data Augmentation

The number of parameters of a CNN network is extremely large. To adjust all these parameters, we need a large dataset. As stated before, most annotated medical image data bases are small. To overcome this limitation, one solution is using data augmentation.

In this study we used data augmentation by increasing the image database. In all the images of the training database we applied some operations like rotation, crops, and flips. Some examples of rotation and flip are shown in Figure 1. it is noteworthy to mention that data augmentation was not applied in the evaluation or test dataset, but only in the training dataset.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3010863, IEEE Access

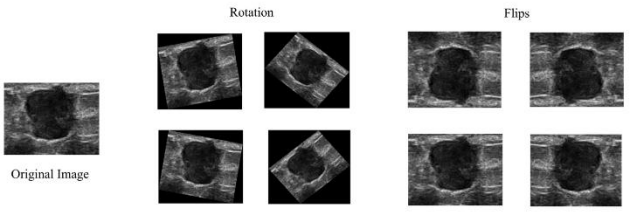

FIGURE 1. **Examples of Image rotation and Flips.**

C. Dropout

Dropout is a type of ensemble technique used in neural networks to reduce overfitting. This technique creates new neural networks by selective dropping some network nodes. and reduces the possibility that the network over-adapts to a given dataset. The predictions of different networks are combined to create the result. Dropout reduces overfitting and indirectly acts as a regularize [27].

D. Batch normalization

Batch normalization is a technique recent developed to cope with the problem of exploding and vanishing gradient problems, which accounts for the gradient increasing or decreasing in successive layers. Other important issue in deep is that of internal covariate shift. During training, the network parameters, as well as the hidden layer inputs changes. This continuing change causes the network training to slow down. The idea of batch normalization is to add "normalization layers" between hidden layers that oppose this type of behavior, creating features with rather similar variance.

Besides improve generalization, batch normalization also decreases the network training time and, adding some noise to each hidden layer, helps in regularization.

E. L_2 Regularization

L_2 regularization, also referred to as Tikhonov regularization, introduces a penalty in the loss function, defined as the sum of squares of the values of the parameters. According to Aggarwal [28] small weight values are penalized less than the large values, because small values do not affect the prediction significantly.

IV. Materials

A. Dataset

The database was shared by the Biomedical Engineering Graduate Program of Federal University of Rio de Janeiro – Brazil. The images were obtained at the Cancer National Institute (INCa, Rio de Janeiro, Brazil) from different patients and where acquired by several radiologist, during routine breast diagnosis exams.

The project from which this database was originated was approved by the INCa research ethics committee (38/2001). The dataset is comprised of 641 images (413 benign cases and 228 malignant). The gold standard classification of each image, as benign or malignant, was histopathologically obtained by biopsy. The images were acquired with a Sonoline Sienna ultrasound machine, in Tiff format, with a depth resolution of 8 bits (256 grayscale).

B. System Specification

The CNN training/validation and testing was done in a computer system with the following characteristics:

- Intel® Core™ i7 6700K @ 4.00 GHz processor;
- GTX 1080 8 GB with 2560 CUDA cores, GPU
- 16GB (2 x 8GB) DDR4 @ 2133MHz RAM memory.

We emphasize that the GPU and processor run under the native frequencies (overclocking was not performed).

V. Methods

In this study, the proposed method for US tumor classification comprises five steps: preprocessing, automatic feature selection using the first layers of a CNN, image classification using logistic regression with cross-entropy loss, hyperparameter tuning and results evaluation. Figure 2 shows a block diagram of these steps.

A. Pre-Processing

The pre-processing stages used for preparing the images to CNN submission were the following: image resize, database balancing, zero-centering and normalization. The first step adjust the images size to CNN architecture, 224 x 224 pixels. Due to the fact that the database consists of a different number of images per category (benign and malignant), in the second step we equalize those quantities. The third and fourth steps, zero-centering and normalization, were employed to increase the networj performance and decrease the training time.

Image Resizing

Different from traditional neural networks, the input data to a CNN is organized into a 2-dimensional grid structure. The value of each individual grid point is referred to as pixel. This grid structure is almost always fixed [25]. A priori, there are no defined rules for the choices of the grid size in a CNN input [27]. The grid size is a trade-off between the application, the training time and the amount of memory needed for processing. Large input grids enable deeper CNN architectures (when treating CNN at each convolutional layer the size of the output matrix decreases. Therefore, starting with a small-sized image limits the number of permitted layers). In recent and acquainted CNN architectures, grid sizes of 224x224 or 320x320 pixels are among some of the common choices [25]. In this study the images were resized to 224x224 pixels.

The most usual techniques for changing image size are interpolation and cropping. While interpolation preserves image information, image cropping does not.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3010863, IEEE Access

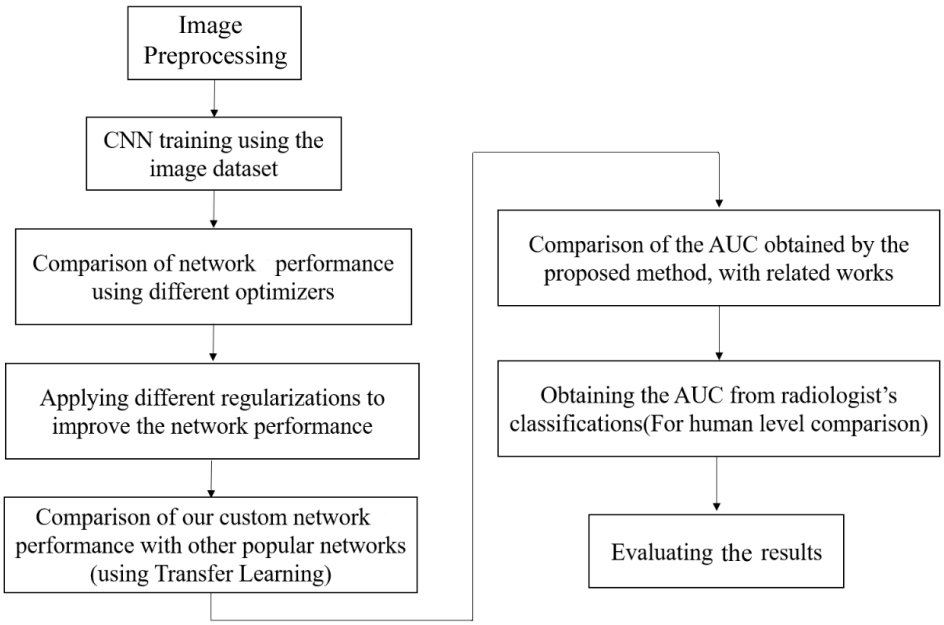

FIGURE 2. **Flowchart of our proposed method.**

In this study, to obtain a final image of 224 x224 pixels and do not lose image proportions we adopted the following procedure: first, we did a bilinear interpolation in the original image, size 159 x 182 pixels, and obtained an intermediate image size 201 x 224 pixels. After, zero-padding the vertical dimension, obtained a final image size of 224x224 pixels. Figure 3 shows the original and final image resulting from this pre-processing step.

Equalization, zero-centering and normalization

The original image database has 413 benign images and 228 malignant images. To equalize the number of benign and malignant images we used a data augmentation technique as described in the following. 185 malignant images were chosen randomly and after applying image flips, to this randomly chosen images, the total number of malignant cases were increased to 413. Therefore, the final image dataset was comprised of 826 images (413 benign and 413 malignant).

The application of zero centering and normalization was done using eq. 1 and eq. 2, respectively. With zero-centering and normalization we obtain zero mean and unit variance, favoring the existence of more uniform gradients, that accelerate the learning process [13].

$$x' = x - \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (1)$$

$$x'' = \frac{x'}{\sqrt{\frac{\sum_{i=1}^{N}(x-\bar{x})^2}{N-1}}} \qquad (2)$$

In eqs. (1) and (2), x represents the original image, x' the zero-centered image, N, the number of samples in the data set and x'' the normalized zero-centered image.

B. Network architecture

The CNN architecture proposed in this work is shown in Figure 4. It consists of four convolutional layers. Each convolutional layer is of a different size and number of filters. The size and number of filters of the convolutional layers are the following: In the first convolutional layer we used 32 filters size 3x3. In the second convolutional layer we used 64 filters size 7x7. In the third convolution layer, we used 128 filters size 5x5. In the last convolutional layer, we used 256 filters size 3x3. In all convolutional operations we used the stride of 1 and zero-padding of 1.

The activation function of all convolutional layers was the ReLU (Rectifier Linear Unit) function. In deep neural networks, the ReLU has many advantages over other non-linearity functions such as sigmoid [13], their application reduces the likelihood of vanishing gradient and also represent a sparse representation of each layer [26], which can improve the performance and accelarate the learning process . After the application of the ReLU function, we have a 2x2 max-pooling layer. The max-pooling operation aims at reducing CNN dimensionality and make the network more invariant to the position of input objects.

The last convolutional layer is followed by two fully connected layers. The first and second fully connected layers are followed by a ReLU and by a softmax activtion fucntion. The ReLU function adds nonlinearity. With the softmax activation function we obtain a binary logistic regression with cross-entropy loss, or a binary classification [26].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3010863, IEEE Access

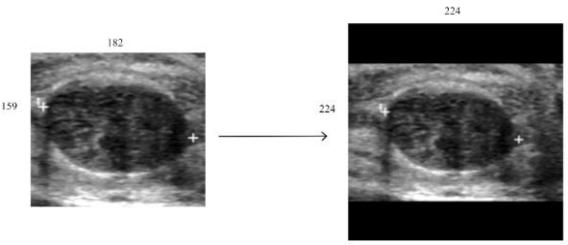

FIGURE 3. The effect of applying bilinear interpolation and zero padding on a sample image: (a) original image, (b) image obtained from the original one through bilinear interpolation and zero-padding.

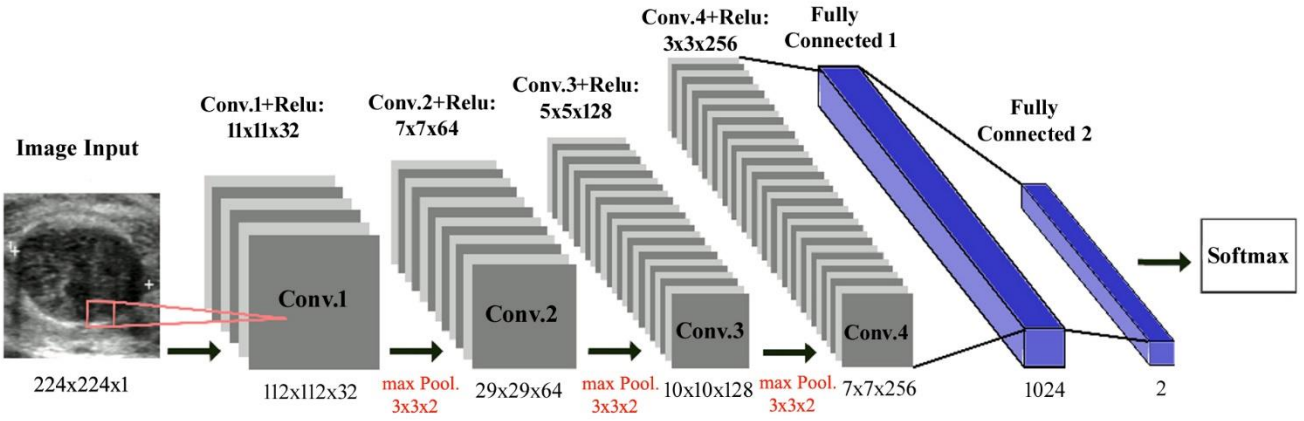

FIGURE 4 The proposed CNN architecture

Training Parameters

The following hyperparameter configuration was employed: For weights initialization we used the Gaussian/Uniform distribution. As optimizers in the backpropagation algorithm, we used. SGDM, ADAM, and RMSPROP. As a criterion for stopping CNN training we used 500 epochs. The Mini-batch size was set to 128. The same parameters were used for all optimizer simulations.

C. Improving CNN Performance

Deep neural networks require a large training set and generally perform better in the presence of more data [17]. Finding a reliable biomedical dataset is a difficult task [19]. Most of the available dataset, like the one used in this work, has a limited number of data.

To avoid overfitting, while maintaining good performance, we introduced image augmentation, L_2 regularization, and dropout. For image augmentation, various Image reflections, rotations, and translations were used to generate a new dataset. This new data set contains 41630 images. Batch normalization was applied after each convolutional layer (before the non-linearity). The dropout was employed after the first fully connected layer, with a probability of 0.5 and L_2 Regularization with a fixed regularization factor of 0.05.

D. Evaluation Metrics

In this work, accuracy (eq.3), specificity (eq.4), sensitivity (eq.5), precision (eq.6), false alarm (eq.7), and the Area Under the ROC curve (AUC) (eq.8) were used as the performance metrics.

$$Accuracy(\%) = \frac{TP + TN}{TP + TN + FN + FP} \times 100 \quad (3)$$

$$Specifity(\%) = \frac{TN}{TN + FP} \times 100 \quad (4)$$

$$\text{Sensitivity}(\%) = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \quad (5)$$

$$\text{Precision}(\%) = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100 \quad (6)$$

$$\text{False Alarm}(\%) = \frac{\text{FP}}{\text{TP} + \text{FN}} \times 100 \quad (7)$$

$$AUC = \int_0^1 f(x)dx \quad (8)$$

In these equations, TP, TN, FP, FN, f(x) represent the number of true positives, true negatives, false positives, false negatives, and the Receiver Operating Characteristic (ROC) curve, respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3010863, IEEE Access

E. Comparison Methods

In this work, three comparisons will be used to evaluate the performance of our proposed method.

First, the best results obtained from our method will be compared to similar work in [10]. The authors in [10] employed the same dataset and followed a similar objective, using different machine learning algorithms.

Second, to determine how well our custom network performs against other CNNs, three well-known networks will be chosen (VGG, ResNet, and GoogLeNet). Using transfer learning, the results obtained from each network will be compared with our architecture.

Finally, to have a human level comparison, all the images in our dataset were classified by two radiologists based on BI-RADS® (Breast Imaging Reporting System) characteristics. The beingin and malignant cases were mixed and shuffled, printed and passed to the radiologists for evaluation.To be able to compare the results of their findings, we need to establish a new method by applying a fixed number for each BI-RADS category, representing the probability of malignancy. To do so, we calculated the mean value of the probability of malignancy and attributed a fixed value to each BI-RADS category (see Table 1).

In addition, to calculate the accuracy, specify, sensitivity, precision and false alarm of radiologists' findings, we made an implicit assumption that the tumors classified as BI-RADS 2, 3, 4a and 4b (with probability of malignantly less than 50%) are benign and the ones classified as 4c, and 5 (with probability of malignancy more than 50%) are malignant.

TABLE 1. Fixed values assigned for each BI-RADS category.

Bi-RADS Category	Probability of Malignancy	Fixed Value
2	0 %	0%
3	0 - 2%	1%
4a	2 – 10%	6%
4b	10 – 50%	30%
4c	50 - 95%	75%
5	More than 95 %	97%

VI. Results

Table 2 summarizes the resultant performance metrics using different optimizers. Although some variations are present, the performance differences using these optimizers are minimal; using SGDM resulted in a slight improvement in AUC value and therefore selected as our candidate.

Table 3 demonstrate the resultant performance metrics after applying image augmentation and regularizations. Figure 5 compares the ROC curves for each case.

As these results show, image augmentation associated with appropriate regularization techniques increased both accuracy and AUC.

To better estimate the performance of our proposed method, some well-known pre-trained models were adapted, and the results were compared (Table 4). These networks are pre-trained on massive datasets, and although the types of data used for training were different, images exhibit similar characteristics, and, in many cases, a simple fine-tuning can adapt the pre-trained model for the new dataset.

Also, a comparison regarding the AUC with a different method was made. In [10], the same dataset was used, and the most important morphological and texture features attributes were selected. Table 5 summarizes the best AUC values achieved by each methodology.

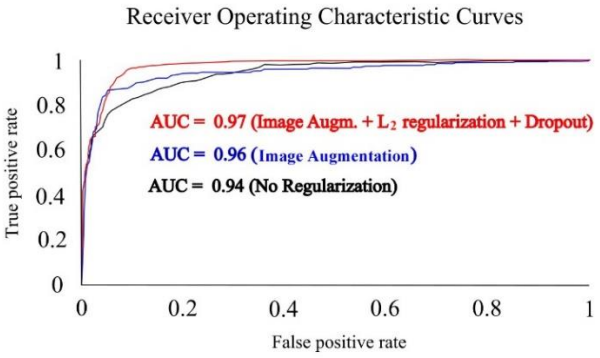

FIGURE 5. The ROC Curves, and the AUC value of our proposed method.

Table 6 shows in a 2x2 contingency table, a comparison of hits and errors of the proposed CNN architecture with VGG19, GoogLeNet, and ResNet50 architectures, based on the accuracy values. To evaluate the statistical significance of the values presented in Table 6, the Chi-square test, χ^2 was applied. The null hypothesis is that there are no significant statistical differences between the accuracy values obtained with proposed CNN architecture and the other architectures. The adopted significance level was 99.0%, and degree of freedom was equal to 1, resulting in a critical value of χ^2 of $t_c = 6.63$. The values of the significance tests shown in Table 6 are all higher than t_c, so the null hypothesis must be rejected.

In this comparison, we applied a statistical significance test [29] to evaluate the differences between the AUC's obtained by the proposed CNN architecture, 0.971, and the other methods shown in table 5: texture feature selection [10] and morphological feature selection [10], 0.897 and 0.942, respectively. The comparison with first one shows that $p < 0.000001$. The comparison with the second one shows that $p = 0.009$. Therefore, the differences are statistically significant at a significance level of 99%.

Finally, to have a human level comparison, the resultant analysis of two radiologists in terms of accuracy, specificity, sensitivity, precision, false alarm, and the AUC value (table 7) were obtained and compared with our proposed method.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3010863, IEEE Access

TABLE 2. performance metrics of the network, using different optimizers.

Iteration	Accuracy	Specificity	Sensitivity	Precision	False Alarm	AUC	Training Time
SGDM	85.98%	88.82%	83.13%	88.42%	11.07%	0.94	31:06
ADAM	86.31%	89.07%	83.57%	88.51%	10.93%	0.93	31:05
RMSPROP	86.08%	85.94%	86.19%	86.34%	14.01%	0.93	31:58

TABLE 3. performance metrics after applying image augmentation and regularization.

Iteration	Accuracy	Specificity	Sensitivity	Precision	False Alarm	AUC	Training Time
Image Augmentation	91.91%	89.30%	94.49%	89.99%	10.58%	0.96	40:25
Image Augmentation + L2 regularization + Dropout	92.05%	89.81%	94.25%	90.51%	10.05%	0.97	40:05

TABLE 4. performance comparison of proposed method versus pre-trained models.

Iteration	Accuracy	Specificity	Sensitivity	Precision	False Alarm	AUC
VGG19	87.88%	92.93%	82.68%	92.68%	7.07%	0.96
GoogLeNet	87.07%	93.66%	80.48%	93.02%	6.34%	0.96
ResNet50	85.85%	79.51%	86.2%	83.44%	20.48%	0.96
Proposed Method	92.05%	89.81%	94.25%	90.51%	10.05%	0.97

TABLE 5. Comparison of the AUC values obtained using different methodologies.

Measurements	AUC
CNN approach	0.971
Texture feature selection [10]	0.897
Morphological Feature selection [10]	0.942

TABLE 6. Chi-square test applied to evaluate statistically significant differences between the accuracies obtained by the proposed architecture and by VGG19, GoogLeNet and ResNet 50 architectures.

Method 1 X Method 2	Hits	Errors	Chi-Square
VGG19 (Accuracy = 87.88%)	726	100	
x			15.5
Proposed Method (Accuracy = 92.05%)	760	66	
GoogLeNet (Accuracy = 87.07%)	719	107	
x			19.09
Proposed Method (Accuracy = 92.05%)	760	66	
ResNet50 (Accuracy = 85.50%)	706	120	
x			26.75
Proposed Method (Accuracy = 92.05%)	760	66	

TABLE 7. PERFORMANCE COMPARISON OF OUR METHOD VERSUS RADIOLOGISTS CLASSIFICATIONS.

Iteration	Accuracy	Specificity	Sensitivity	Precision	False Alarm	AUC
Radiologist 1	87.58%	99.73%	73.55%	99.58%	0.3%	0.97
Radiologist 2	81.76%	85.71%	74.44%	73.77%	26.45%	0.84
Our Proposed Method	92.05%	89.81%	94.25%	90.51%	10.05%	0.97

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3010863, IEEE Access

VII. Discussion

To summarize the obtained results, we categorize the findings into, first, the efforts to increase the network performance, and second, the comparison methods.

As for the efforts to increase the performance, various regularization techniques were used.

The given dataset is relatively small, which causes the system to suffer from overfitting problem. Data augmentation, hyperparameter tuning, and applying appropriate regularization, resulted in a significant increase both in terms of accuracy and the AUC.

After improving the performance of our proposed method, a set of comparisons were done to analyze better and understand the behavior of the system. In this work, three comparisons were made:

- Comparison of our CNN architecture with other three well know CNN architectures in the classification of tumors in our database.
- Comparison of our proposed method with some traditional machine learning techniques in the classification of the same dataset.
- Human-level comparison.

The objective of the first comparison was to evaluate the performance of our CNN against some other well-known network architectures. Using transfer learning, VGG19, GoogLeNet, and ResNet50 were used to classify the tumors in our dataset, and the results were compared to our proposed method. Between these three networks, GoogLeNet demonstrated the best performance.

Although GoogLeNet resulted in very satisfactory results, our network outperforms it in terms of accuracy, sensitivity, and AUC. The differences are statistically significant.

In the second comparison, the effectiveness of CNN versus some traditional machine learning algorithms, in the classification of breast tumors in our dataset, was evaluated. In [10], the same dataset was used. As table 5 summarized the results, the authors achieved an AUC equal to 0.897 and 0.942, using texture and morphological features, respectively, which is lower than 0.97 achieved by our CNN approach.

In the last comparison, the performance of our method was evaluated against the analysis of two radiologists. The radiologists were asked to classify the tumors based on the BI-RADS classification.

For a fair comparison, after the specialists' analysis, the tumors, categorized as 2, 3, 4a and 4b (with probability of malignancy less than 50 %) were classified as benign and the ones categorized as 4c and 5, as malignant (it is worth to mention that the neural networks follow a similar behavior in classification of objects). As can be seen (Table 7), our proposed method outperformed the radiologists' evaluations in terms of accuracy and sensitivity but falls below the radiologist 1 performance regarding specificity, precision, and false alarm.

Figure 6 demonstrates the results of these comparisons regarding the ROC curves and the AUC value.

VIII. Conclusion

In this work, we investigated the effectiveness of Deep Learning, in particular, CNNs, for the classification of abnormalities in breast ultrasound images. A network architecture with four convolutional layers was proposed capable of classifying US images as either Benign or Malignant. A variety of attempts were made to improve the performance of the proposed method.

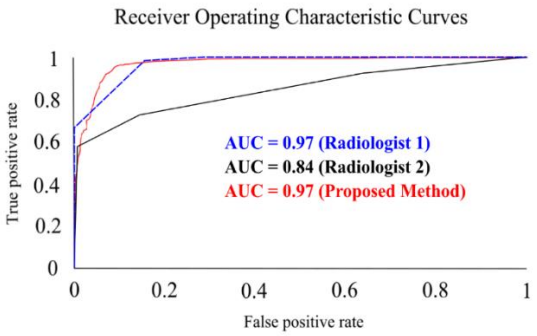

FIGURE 6. The ROC Curves, and the AUC value of our proposed method vs. radiologist's diagnosis.

We explored various hyperparameter tuning and regularization techniques such as image augmentation, L2 regularization, and dropout, to increase network performance and decrease the overfitting problem. The performance of both systems, with and without regularization, were evaluated both in terms of accuracy and the Area Under the ROC Curve (AUC). Our proposed method, without regularization, presented an overall accuracy of 85.98% and AUC equal to 0.94. After applying regularization and fine-tuning, the accuracy and the AUC were significantly improved: 92.05% for accuracy and 0.97 for the AUC. To verify the effect of overfitting on the network, the proposed method was compared to some pre-trained CNN architectures using transfer learning and fine-tuning. The comparison demonstrated the effectiveness of our proposed method against these well-known CNN architectures, for the given dataset. Although the pre-trained models had a similar performance, our network with fewer hidden layers is faster for testing and more suited for this specific application (considering the number of training data and characteristics of US images). Also, the results were compared to another CAD system, which considered to be state of the art for classification of breast tumors in US images, employing the same data set. The authors in [10], obtained their best results, using five morphological features, attaining an AUC equal to 0.942. The comparison showed that our proposed method, using automatic feature selection and CNN, outperformed the system using handcrafted morphological features. Finally, to have a human level comparison, the obtained results were compared to two radiologist's diagnoses, our proposed method outperformed the specialist's analysis in terms of accuracy but could not reach the same levels of precision and specificity obtained by one of the radiologists.

The main contribution of this work was proposing and fine tuning a CNN model (using data augmentation and appropriate regularization) that obtains a good generalization, with a few numbers of layers. The performance of this model was better than traditional machine learning approaches as well as pre-trained networks with much larger architectures.

Although the proposed method provided promising results and our study proved the effectiveness of CNNs for classification of breast lesion, even with a limited number of training data, our model can be improved in several ways. It is known that in the presence of more data, the performance of CNNs increases. In this work, the dataset was relatively small, and a limited number of hidden layers were used to prevent the overfitting problem (by preventing the system to adapt too much to the data). In future work, we plan to gather a bigger dataset and employ different CNN architectures with more hidden layers. Also, we plan to further study the tumors not classified correctly by our system, trying to find some similarities among these cases and the ones misclassified by the radiologists, adding more data with these specific characteristics to our dataset and build a more reliable system, closing the gap to Human-Level performance. Although we have obtained excellent results with the proposed architecture, in future work we also plan to combine the extracted characteristics from the three transfer leaning architectures, VGG, ResNet, and GoogLeNet, in a final classification layer, as shown in [30].

ACKNOWLEDGMENT

This research, according for in Article 48 of Decree nº 6.008/2006, was funded by Samsung Electronics of Amazonia Ltda, under the terms of Federal Law nº 8.387/1991, through agreement nº 004, signed with the Center for R&D in Electronics and Information from the Federal University of Amazonas - CETELI/UFAM, and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Funding Code 001. Academic English Solutions (https://www.academicenglishsolutions.com) revised this paper.

REFERENCES

[1] Siegel, R. L.; Miller, K. D.; Jemal, A.: Cancer Statics, 2017. CA Cancer J Clin, vol. 67, Issue 1, pp. 7-30, 2017.

[2] Stewart, B. W.; Wild, C. P.: World Cancer Report 2014. Edited by, WHO, World Health Organization, www.who.int/cancer/publications/WRC_2014/en/, 2014.

[3] Akin, O.; Brennan, S.; Dershaw, D.; Ginsberg, M.; Gollub, M.; Schoder, H.; Panicek, D.; Hricak, H.: Advances in oncologic imaging: Update on 5 common cancers. CA Cancer Journal for Clinicians, vol. 62, no. 6, pp. 364–393, 2012.

[4] Stavros, A.; Thickman, D.; Rapp, C.; Dennis, M.; Parker, S.; Sisney, G.: Solid breast nodules: Use of sonography to distinguish between benign and malignant lesions. Radiology, vol. 196, no. 1, pp. 123–134, 1995.

[5] Zhou, S. K.; Greenspan, H.; Shen, D.: Deep Learning for Medical Image Analysis. First Edition, Elsevier, USA, 2017.

[6] Singh, B. K.; Verma, K.; Thoke, A. S.: Adaptive gradient descent backpropagation for classification of breast tumors in ultrasound imaging. Proceedings of the International Conference on Information and Communication Technologies, Icict, vol. 46, pp. 1601-1609, 2015.

[7] Chen, Y.; Ling L.; Huang Q.: Classification of breast tumors in ultrasound using biclustering mining and neural network. 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Datong, pp. 1787-1791, 2016.

[8] Liang, Y.; He, R.; Li, Y.; Wang, Z.: Simultaneous segmentation and classification of breast lesions from ultrasound images using Mask R-CNN. *IEEE International Ultrasonics Symposium (IUS)*, Glasgow, Scotland pp. 6-9, 2019.

[9] Yap, M. H.; Pons, G.; Marti, J.; Ganau, S.; Sentis, M.; Zwiggelaar, R.; Davison, A. K.; Marti, R,: Automated Breast Ultrasound Lesions Detection using Convolutional Neural Networks. IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 4, pp. 1218-1226, 2018.

[10] Bakkouri, I.; Afdel, K.: Breast tumor classification based on deep convolutional neural networks. International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Fez, pp. 1-6, 2017.

[11] Rice, M. E.; Harris, G. T.: Comparison Effect Size in Follow-up Studies: ROC Area, Cohens d, and r, Lawan Human Behavior, vol.29. no.5, 2005.

[12] Flores, W. G.; Pereira, W. A.; Infantosi, A. F. C.: Improving classification performance of breast lesions on ultrasonography. Pattern Recognit., vol.48, no. 4, pp. 1125-1136, 2015.

[13] Goodfellow, I.; Bengio, Y.; Courville, A.: Deep Learning. First Edition, MIT Press, USA, 2016.

[14] Cao Z.; Duan L.; Yang G.; Yue T.; Chen Q.; Fu H.; Xu Y.: Breast tumor detection in ultrasound images using deep learning, Springer International Publishing, G. Wu et al. (Eds.): Patch-MI, LNCS 10530, pp. 121-128, 2017.

[15] Hijab, A.; Rushdi, M.; Gomaa, M.; Eldeib, A.: Breast cancer classification in ultrasound images using transfer learning, IEEE fifth inrenational conference on advances in biomedical engineering, 2019.

[16] Zhou, S. K.; Greenspan, H.; Shen, D.: Deep learning for medical image analysis. First Edition, Elsevier, USA, 2017.

[17] Byra, M.; Galperin, M.; Ojeda-Fournier. H.; Comstock, C.; Andre. M.: Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion, Med. Phys., vol. 46, no. 2, 2019.

[18] Wang, Y. I.; Choi, E. J.; Choi, Y.; Zhang, H.; Jin, G. Y. and Ko, S. B.: Breast cancer classification in automated breast ultrasound using multiview convolutional neural network with transfer learning, Ultrasound in Med. & Biol., vol. 46, no. 5, pp. 1119-1132, 2020.

[19] Huang, Y.; Han, L.; dou, H.; Luo, H.; Yuan, Z.; Liu, Q.; Zhang, J.; Yin, G.: Two-stage CNNs for Computerized BI-RADS categorization in breast ultrasound images, Huang et al. Biomed Eng online, 2019

[20] Gómez W.; Rodríguez A.; Pereira W. C. A.; Infantosi A. F. C.: Feature selection and classifier performance in computer-aided diagnosis for breast ultrasound, 10th International Conference and Expo on Emerging Technologies for a Smarter World (CEWIT), Melville, NY, pp. 1-5, 2013.

[21] Sakr, G. E.; Mokbel, M.; Darwich, A.; Khneisser, M. N.; Hadi, A.: Comparing deep learning and support vector machines for autonomous waste sorting, IEEE International Multidisciplinary Conference on Engineering Technology (IMCET), Beirut, pp. 207-212, 2016.

[22] Dey, N.: Classification Techniques for Medical Image Analysis and Computer Aided Diagnosis, Advances in ubiquitous sensing applications for healthcare (Book 4), 2019

[23] Zeimarani, B.: Breast Tumor Classification in Ultrasound Images using Deep Convolutional Neural Network. M. S. thesis, Amazonas Federal University, Manaus-AM, Brazil, 2019.

[24] Zeimarani, B.; Costa, M. G. F.; Nurani, N. Z.; Costa Filho, C. F. F.:

A novel breast tumor classification in ultrasound images using deep convolutional neural network, XXVI Brazilian Congress on Biomedical Engineering, Springer Singapore, pp.89-94, 2018.

[25] Pal, K. K.; Sudeep, K. S.: Preprocessing for Image Classification by Convolutional Neural Networks. International Conference on Trends in Electronics Information Communication Technology, pp. 1778–1781 (2016)..

[26] Khan S.; Rahmani H.; Shah, S. A. A.: A Guide to Convolutional Neural Networks for Computer Vision, Morgan & Claypool Publishers, Synthesis Lectures on Computer Vision, ISBN-10: 1681730219, 2018.

[27] Li, F. F.; Johnson, J. and Yeung, S.: CS231n: Convolutional Neural Networks for Visual Recognition, Course Notes, Available at : http://cs231n.github.io/..

[28] Aggarwal, C. C. Neural Netwroks and Deep Learning, Springer, Switzerland, 2018,

[29] Hanley, J.A. and McNeil, B. J. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. Radiology, vol. 143, pp. 29-36, 1982.

[30] Khan, S.; Islam, N.; Jan, Z.; Din, I. U. and Rodrigues, J. J. P. C.: A novel deep learning based framework for the detection and classification of breast cancer using transfer learning, Pattern Recognition Letters vol. 125, pp.1–6, 2019.