Biomedical Signal Processing and Control 76 (2022) 103696

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Breast cancer: Classification of suspicious regions in digital mammograms based on capsule network

Khaoula Belhaj Soulami [a,*], Naima Kaabouch [b], Mohamed Nabil Saidi [c]

[a] National Institute of Posts and Telecommunications (INPT), STRS Lab, Rabat, Morocco
[b] Electrical Engineering Department, University of North Dakota, Grand Forks, USA
[c] National Institute of Statistic and Applied Economy INSEA Laboratory of Information Systems, SI2M Rabat, Morocco

ARTICLE INFO

Keywords:
Breast cancer
Classification
Capsule
CapsNet
Mammograms
Mass
Lesion
Deep learning

ABSTRACT

Mammography screening is one of the common techniques that help identify suspicious masses' malignancy of breast cancer at an early stage. Yet, the early diagnosis of masses in mammograms is still challenging for dense and extremely dense breast categories and requires efficient and automated systems to assist radiologists in their diagnosis. Deep learning methods have widely been used for medical imaging applications, especially for breast masses classification. A novel capsule concept has recently been proposed to solve the drawbacks of deep learning models based on classic convolutional networks by learning the hierarchical structures in images. However, the originally proposed capsule network suffers from issues due to high number of parameters and high computational time. This paper proposes a new architecture of the capsule network that significantly reduces the computational time of the original capsule network by 6.5 times and makes the training of breast masses ROIs on affordable GPUs possible. The proposed architecture was fine-tuned according to the number of kernels and capsules and by using data augmentation. The results of the evaluation of the four breast density categories show that our capsule-based model outperforms existing techniques in classifying suspicious breast masses in one stage. The binary classification of masses into normal and abnormal achieves an accuracy of 96.03%, a F1-score of 96.27%, a precision of 94.28%, a recall of 98.38%, a specificity of 93.97%, and an AUC of 0.997. The multi-classification of breast masses into normal, benign, and malignant scores an accuracy of 77.78%, a F1-score of 77.45%, a precision of 71.54%, a recall of 84.54%, a specificity of 83.15%, and an AUC of 0.9.

1. Introduction

Breast cancer is the most frequent cancer type in women and the second most common cancer overall [1]. The fatality rate of breast cancer counts for 17% of cancer mortality [2]. Accurate detection of breast cancer at an early stage helps reduce the mortality rate [3]. Nowadays, mammography is the most useful and common screening tool for detecting breast lesions [4]. However, early detection using mammography is complex and challenging for the dense breast type where the normal breast tissue is often misdiagnosed as abnormal lesions or vice versa. This misleads to false positives and false negatives and, in some cases, to unnecessary additional examinations. The automatic identification of masses in the breast [5–10] and particularly the classification of their pathology presents crucial assistance to radiologists and experts in their decision-making process. Such classification can significantly minimize the number of false positives that lead to unnecessary clinical practices and discomforting biopsies.

During the last decade, machine learning models have been proposed for mammograms classification [5,11–13]. However, deep learning models based on convolutional neural network (CNN) have created a revolution in computer vision problems and are becoming frequently used for mammograms classification. Unlike machine learning, deep learning does not require structured feature data as input; instead, it directly extracts features from the input images based on its CNN layers. Nevertheless, CNN networks have limits and fundamental drawbacks. The orientation of the different components in an image and the relative spatial relationships between these components are not considered in a CNN-based network. In fact, the convolutional layer is the main layer in a CNN network. It helps detect and extract simple features from the images from deeper layers and combine them with more complex features from higher layers. These features are then fed to dense layers to make classification predictions. Max pooling is a solution that has been

[*] Corresponding author.
E-mail address: k.belhajsoulami@gmail.com (K. Belhaj Soulami).

https://doi.org/10.1016/j.bspc.2022.103696
Received 29 November 2021; Received in revised form 2 March 2022; Accepted 9 April 2022
Available online 18 April 2022

introduced into CNN networks to reduce the spatial size of the data throughout the different layers of the CNN network and to detect complex features of higher convolutional layers in large regions of the images. Indeed, max pooling achieves very good performance in many fields and surprisingly presents a good solution; however, as it decreases the data's spatial size, it loses valuable information from the data as well. The capsule is a new concept to deep learning models that considers the hierarchical and spatial relationships between the different internal components of an image without losing valuable information.

The capsule concept was first introduced by Geoffry Hinton et al. in 2011 [14]. Nevertheless, until 2017, Hinton and his team successfully implemented an algorithm called dynamic routing that allows the training of a capsule-based network [15]. Besides, even if the algorithm were discovered back in 2011, it would not have worked as it is today because the computers were not powerful enough in the pre-GPU era. The capsule network has decreased the training set sizes and the error rates when applied to the MNIST dataset. Results showed higher performance compared to the CNN-based model, particularly for overlapped digits. CapsNet has captured significant attention in the last two years and has been tested in different applications such as image classification, image segmentation, and semantic segmentation, which has shown breakthrough results [16–22]. Motivated by these promising results and the novelty of the concept, we propose a model based on the capsule concept to classify suspicious masses in the breast into normal, benign, and malignant. As mentioned before, these masses' structure is similar, which makes their classification a challenging task even for experts. In addition, capsules solved the issue of the hierarchical relationships between the different components in an image, which makes it recognize internal components of an image. Hence, we applied the capsule concept to the breast classification to test its performance in differentiating between breast masses and their surrounding normal tissue. Nonetheless, the training of the original model of CapsNet network is much slower than other existing deep learning models even for small size images as it contains a high number of parameters. In this context, LaLonde et al. [21] addressed the memory issue of CapsNet by dramatically reducing the number of parameters in the network and updating the dynamic routing algorithm for image segmentation. Therefore, in this paper, we propose a CapsNet model inspired by the work in [22] for the classification of suspicious masses in the breast. However, we maintain the original dynamic routing algorithm proposed in Hinton et al. paper [15]. In fact, the authors of [22] used a capsule-based model to segment lungs in CT images and they managed to reduce the number of parameters by modifying the original CapsNet architecture. Their outstanding results motivated us to modify and adapt the CapNet original architecture, but in our case, to solve the breast mass classification problem. The proposed CapsNet is evaluated on both validation and test data and according to the density rate of each of the used mammograms.

The rest of the paper is organized as follows: Section 2 gives an overview of the most recent machine learning and deep learning works proposed for the classification of suspicious breast masses. Section 3 describes in detail the methodology of the proposed capsule-based architecture. Section 4 presents the experimental results and the datasets used for the classification performance of the proposed CapsNet model. Section 5 discusses and compares the obtained results with those of the most recent related works. Section 6 concludes the paper and gives the advantages and the future works of our proposed architecture.

2. Related works

The aim of regular breast screenings is the detection of malignant abnormalities at an early stage. The common screening modalities used for breast cancer detection, include, ultrasonography (US), contrast enhanced magnetic resonance imaging (MRI), and mammography. Many existing works classified suspicious lesions in the breast using US images [23–25] and MRI images [26–28]. However many studies showed that ultrasound fails to detect breast cancer at an early stage and can be used as a complement to mammography [29–31]. In fact, the US images miss small lumps and microcalcifications that are considered early signs of breast cancer. In addition, and like the ultrasound imaging, the MRI screening modality is often used as a complement to mammography [32,33] or in presurgical planning. In fact, compared to ultrasonography and mammography, the MRI screening is more expensive and has a significant positive rate [34]. Thus, the gold standard modality used for breast cancer detection is mammography, which is the most efficient and the less expensive screening option that helps detect breast cancer at an early stage. Currently, it has been proved that the usage of artificial intelligence systems as support for the classification of suspicious breast lesions has increased the radiologists' cancer detection accuracy without needing any further readings [35]. These artificial intelligence systems include machine learning-based and deep learning-based models.

Machine learning models have been widely used for the classification of lesions in digital mammograms as they give good results in the classification of small datasets. Most of recent works still use machine learning by improving existing methods, proposing new combinations of features or classifiers, or enhancing the input images beforehand in the preprocessing phase to increase the classification performance. For instance, Bhosle et al. [36] suggested a method based on machine learning tools for the classification of mammograms. Prior to the classification process, the authors applied a preprocessing technique to the mammograms. Then, they extracted regions of interest (ROI) of the suspicious regions from mammograms. Histogram equalization was applied to the ROIs for contrast enhancement purposes. This process was followed by the extraction of Gray Level Co-occurrence Matrix (GLCM) features from the enhanced ROIs. The authors used a pre-processing step to optimize the GLCM features that were then given to a proposed Hybrid KNN–RBFSVM for masses' ROIs classification into benign or malignant.

Vijayarajeswari et al. [37] proposed an approach-based on the support vector machine for the classification of mammograms into normal or abnormal. First, the authors preprocessed the mammograms by removing the pectoral muscle as it has similar characteristics as the abnormal areas in the breast image. Then, they applied a maximization estimation technique to segment the suspicious regions in the mammogram. Further, Hough transform was used for the extraction of features from the segmented mammograms. Finally, they applied a support vector machine classifier to these features to classify the mammograms into abnormal or normal.

Alqudah et al. [38] used a two-stages classification of breast masses in digital mammograms. The mammograms were preprocessed beforehand by removing the pectoral muscle from the mammograms. The ROIs of the suspicious lesions were then segmented from these mammograms. Further, GLCM features were extracted from the segmented ROIs and were fed to a two-stage classification process. The first stage includes the classification of breast lesions into seven classes, normal, calcification, circumscribed, spiculated, ill-defined, architectural distortion, and asymmetry. This was done based on a probabilistic neural network. The second classification stage defines the malignancy of the lesions, whether they are Benign or Malignant. This was performed using a support vector machine (SVM) classifier.

Jadoon et al. [39] suggested a technique for the classification of mammograms using a deep learning model. First, the authors used data augmentation to increase the number of training patches. Then, they applied the contrast limited adaptive histogram equalization filter to enhance the contrast of the used patches. Prior to the feature extraction part, these preprocessed patches were decomposed by means of two different methods, two-dimensional discrete wavelet transform (2D-DWT) and discrete curvelet transform (DCT). Further, the dense scale-invariant features (DSIFT) were extracted from the decomposed patches and then fed to a convolutional neural network (CNN). The Softmax and support vector machine layers are used to train CNN features to classify patches into three classes, normal, benign, and malignant. The results show that the CNN model with DCT decomposition achieves better classification results compared to the CNN model and 2D-DWT decomposition.

All these Machine learning methods achieve good results in classifying digital mammograms and are suitable for training small datasets. However, these methods involve multiple steps, including preprocessing, feature extraction, feature selection, and classification. In each step, the appropriate method and the fine-tuning of its corresponding parameters highly affect the classification results. In addition, these methods were not evaluated according to the breast density tissue. As mentioned in the previous section, the classification of the suspicious regions in mammograms remains challenging and difficult for the dense and extremely dense mammogram categories. Thus, the actual performance of the previously mentioned method still unknown.

On the other hand, deep learning CNN-based networks offer the possibility of classifying images "end-to-end" and does not require intermediate steps. For instance, Khan et al. [40] presented a multi-view feature fusion (MVFF) based four views model for Mammogram classification using a convolutional neural network (CNN). This proposed approach contains a three-stage classification of the mammograms using CNN. After data augmentation and in the first stage, the mammograms were classified into abnormal or normal, then in the second stage, into mass or calcification. In the final stage, they were classified into malignant or benign. In each of the three stages, classifiers perform independently on each view. Then, the extracted features were fused into one layer for a final prediction.

Li et al. [41] proposed a deep learning model for the classification of whole abnormal mammograms into benign and malignant. The proposed model, DenseNet-II, is an improved version of the DenseNet neural network model. The authors opted for a preprocessing step before the classification of the mammograms, in which the mammograms were normalized and enhanced to prevent the overfitting issue. The preprocessed mammograms were then fed to the proposed DenseNet-II and were classified into malignant and benign, and the training process was conducted based on a 10-fold cross-validation method. The proposed model's classification performance was compared with other deep learning models, including, AlexNet, VGGNet, GoogLeNet, and DenseNet.

Agnes et al. [42] introduced a multi-scale all-convolutional neural network for the classification of digital mammograms into normal, benign, and malignant. The images were preprocessed prior to the classification step by removing artifacts, noise, and pectoral muscle. Then, the preprocessed images were given to the proposed convolutional neural network classifier. The proposed model consisted of convolutional layers only, hence, it was named all CNN.

Falconi et al. [43] utilized the transfer learning from pre-trained deep learning models to classify digital mammograms into benign and malignant. First, ROIs were extracted from mammograms and were enhanced. Then, data augmentation was used to increase the number of training samples. The authors fine-tuned parameters of previously pre-trained models, including NasNet, MobileNet, Resnet, Xception, and Resnext. Then, these models were used to train breast abnormalities into malignant or benign. The fine-tuned pre-trained VGG16 model achieves the best performance compared to the other models.

Perre et al. [44] used a convolutional neural network for the classification of lesions in digital mammograms. The authors opted for a transfer learning approach in three different CNN models, CNN-F (Fast, imagenet-vgg-f) model, CNN-M (Medium, imagenet-vgg-m)), and Caffe reference model. These models were pre-trained on a large dataset, Imagenet, and were then fine-tuned using a smaller mammogram dataset. The three CNN models' performance was investigated in classifying patches of digital mammograms into benign or malignant. The Caffee model shows the best results and best AUC value without the normalization of input mammogram patches.

Zhang et al. [45] proposed a multi-scale attention DenseNet model for the classification of digital mammograms. This model is a two-stage classification method that classifies mammograms into normal and abnormal, then into benign and malignant. It consists of two independent branches that extract features from two mammograms from different views, which are then fused into a common prediction. According to the authors' results, the model archives good results for both types of classification, normal-abnormal, and benign-malignant.

Arora et al. [46] proposed an ensemble method for the classification of lesions in digital mammograms into benign or malignant. A preprocessing step that consists of a histogram equalization method was applied to the mammogram patches prior to the classification process. According to the authors, this step improves the mammograms' contrast and enhances the classification performance. The classification method consists of two steps. First, an ensemble classifier based on the transfer learning concept was used for the extraction of features from the preprocessed mammograms. This ensemble classifier contained pre-trained models, namely, AlexNet, VGG16, ResNet, GoogLeNet, and Inception-ResNet. The extracted features from each network of the ensemble classifier were combined into one vector that was fed to the next classification layer consisting of an artificial neural network unit for a final classification. The classification results show promising results.

Tsochatzidis et al. [47] investigated the use of deep learning models in classifying mammogram patches into benign and malignant. The models used in the authors comparative study are AlexNet, VGG-16, VGG-19, ResNet-50, ResNet-101, ResNet-152, GoogLeNet, Inception-BN (v2). The models went through two training scenarios: the first transfers the learning using pre-trained weights, while for the second, the weights are learned from scratch and are randomly initialized. Results show that AlexNet achieves the best performance when trained from scratch, while the ResNet-50 and ResNet-101 networks attain maximum performance by fine-tuning their corresponding pre-trained weights.

Aboutalib et al. [48] suggested a modified version of the deep learning model AlexNet for the classification of mammograms into recalled-benign, negative, and malignant. The authors opted for the transfer learning method to fine-tune their proposed deep learning model. The performance of the model classification was tested on both binary classification and three-class classification. The binary classification includes five scenarios: malignant versus (recalled-benign and negative), malignant versus negative, malignant versus recalled-benign, negative versus recalled-benign, and recalled-benign versus (malignant and negative). Whereas the triple classification consists of one scenario: malignant versus negative versus recalled-benign. Results show that the triple classification gave the lowest AUC score.

All the previously mentioned works, both the machine learning-based and the convolution-based deep learning methods still have a limitation when it comes to classifying breast masses in dense and extremely dense breast categories. In fact, the abnormal breast tissue can share similar characteristics with the normal dense tissue in these breast categories, which makes it misclassified. Hence, most of the proposed works used for the classification of breast masses are not evaluated according to the breast density as most of the time these masses are misclassified and negatively affect the overall performance of the proposed techniques. In this paper, we propose an effective deep learning model based on capsules that classifies the breast lesions into normal, benign, and malignant and it is evaluated according to all breast density categories. Our proposed model solves the limitations of the existing techniques as follows:

- Machine learning-based methods: as previously mentioned, these techniques require several steps and parameters than need to be optimized for better classification results. Hence, it is difficult to find the best combination of techniques in each step and their corresponding optimized parameters that best classify masses for all the breast tissue categories. However, we proposed a one-stage technique based on capsules that do not require any intermediate steps or stages and classifies the breast tissue into three classes normal, benign, and malignant including those in dense and extremely dense breast categories.

- Convolution-based deep learning methods: as previously mentioned in the introduction, these techniques use max pooling, which reduce the among of information extracted in the convolutional layers. However, capsule-based models keep all the features extracted in the lower convolutional layers. These features hold information about the masses' location, orientation, and its hierarchical relationship with the surrounding normal tissue that helps distinguish between the normal and the abnormal breast tissue boundaries. Thus, our capsule-based model outperforms the most popular CNN-based models in classifying breast masses in all breast density categories including the dense and extremely dense breast categories.

3. Methodology

3.1. Capsule concept

As mentioned previously, three years ago, Geoffrey Hinton and his team published two papers that introduced a neural network based on a new concept called capsules [15,16]. Traditionally, artificial neurons in models like CNN produce an activity scalar that represents the probability of detection. While artificial neurons in capsules collectively output detection probabilities in the form of an activity vector. In fact, the features produced in CNN models with max pooling lose valuable information to ensure invariance of activities. This invariance helps detect an object even if it shifts or changes its orientation or position in the image. As opposed to CNN with max pooling models, capsules' features encapsulate all important information of objects in images (such as size, position, orientation, deformation, texture, and hue) and then encode the probability of these features in the vector's length. This means that even though a detected object changes its orientation or its state in the image, its corresponding probability remains the same. This helps to keep the wanted invariance of activities and preserving all the information about the object in an image at the same time.

A capsule output vector v_j forms according to 2 main steps, the production of output vectors from lower-level capsules and the update of routing weights using a novel algorithm "dynamic routing" also known as routing-by-agreement. First, the output vectors $\hat{u}_{j/i}$ of lower-level capsules i are produced by multiplying input vectors u_i of these capsules by their corresponding weight matrices, W_{ij}, that carry spatial relationships information between lower-level and higher-level features. This is done by means of an affine transform as given by Eq. (1). In fact, the length of input vectors u_i hold probabilities of the detected objects in lower-level capsules according to their directions and internal state.

$$\hat{u}_{j/i} = W_{ij}u_i \tag{1}$$

Where $\hat{u}_{j/i}$ is the output vector of capsule j in layer $l+1$ according to capsule i in lower layer l, u_i is the output vector of capsule i in the layer below l, and W_{ij} is the weight matrix holing spatial relationships between the capsules i and j.

Second, the output vector of predictions v_j coming from higher-level capsule j is calculated based on the dynamic routing algorithm. This algorithm allows to train capsule-based networks by insuring the communication between capsules (cf. Algorithm. 1). Traditionally in neural network models, the weights are updated during back-propagation; but in capsule-based models (Eq. (3)). In the first step of the dynamic routing algorithm, all routing weights c_i between lower-level capsules i and all the capsules in the layer above are computed using a softmax function (Eq. (2)). The use of softmax function ensures that the routing weights c_i are positive numbers, and their sum equals one after the calculation of all the weights of the above layer. A linear combination of output vectors of lower layer capsule i and updated weights c_{ij}, determined in the previous step, produces output vector s_j (Eq. (3)). Then, this established vector is fed to the squash nonlinearity function, which preserves the direction of the vector and enforces its length to be no more than 1. This function produces the output vector v_j (Eq. (4)) for the higher-level capsule j. In the last step of the algorithm, the routing weights are updated by calculating the agreement between the current output vector, v_j, of the higher-level capsule,j, and the prediction of lower-level capsule i. This agreement is added to the initial weights, b_{ij}, to produce new values for all the routing weights linking the lower-level capsule i to the higher-level capsule j. After this step, the algorithm repeats the process r times, also called the routing number.

$$c_i = \frac{e^{b_{ij}}}{\sum_k e^{b_{ik}}}, \quad k \neq j \tag{2}$$

Where is b_{ij} is a temporary coefficient that is iteratively updated and then stored in the routing coefficient between the lower-level capsule i and the higher-level capsule j and k represents the higher-level capsules other than .

$$s_j = \sum_i c_{ij}\hat{u}_{j/i} \tag{3}$$

Where s_j is sum of weighted input vectors of capsule j, c_{ij} is the weight that multiplies the output vector of capsule, i from the lower layer, and computes input vectors of next level capsule j. $\hat{u}_{j/i}$ are the output vectors of capsules i.

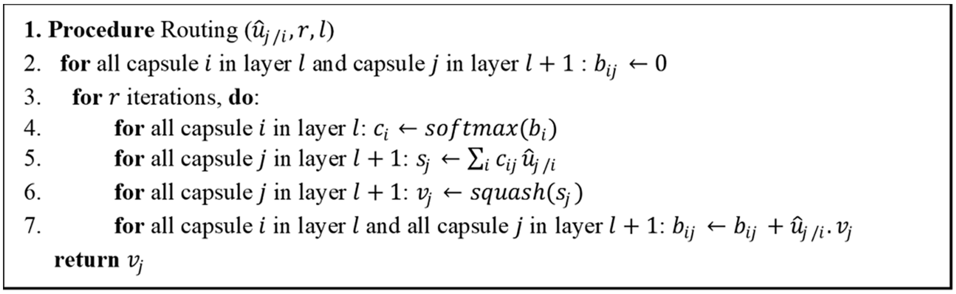
$$v_j = \frac{\|s_j\|^2 s_j}{1 + \|s_j\|^2 + \|s_j\|} \tag{4}$$

Where s_j is sum of weighted input vectors of higher-level capsule j.

$$b_{ijnew} = b_{ijold} + \hat{u}_{j/i}.v_j \tag{5}$$

Where b_{ijnew} and b_{ijold} are the new and the old routing weights,.$\hat{u}_{j/i}v_j$

Algorithm. 1. Dynamic routing algorithm, as published in the original paper [14].



```
Procedure Routing ($\hat{u}_{j/i}, r, l$)
for all capsule $i$ in layer $l$ and capsule $j$ in layer $l+1$ : $b_{ij} \leftarrow 0$
 for $r$ iterations, do:
     for all capsule $i$ in layer $l$: $c_i \leftarrow softmax(b_i)$
     for all capsule $j$ in layer $l+1$: $s_j \leftarrow \sum_i c_{ij}\,\hat{u}_{j/i}$
     for all capsule $j$ in layer $l+1$: $v_j \leftarrow squash(s_j)$
     for all capsule $i$ in layer $l$ and all capsule $j$ in layer $l+1$: $b_{ij} \leftarrow b_{ij} + \hat{u}_{j/i}.v_j$
 return $v_j$
```

3.2. Proposed CapsNet architecture

In this section, we explain the proposed architecture of the capsule network, CapsNet. We propose a simplified architecture of the original capsule network described in [15,16]. The three-layer capsule network introduced in Sabour et al. [16] is extremely computationally expensive as it has a significantly high number of parameters. In this work, we address this issue by significantly reducing the number of parameters in our modified CapsNet architecture. The first layer in the proposed CapsNet architecture is a two-dimensional convolutional layer with a kernel size of 5x5 and one stride followed by a ReLU nonlinear activation. This layer's role is to detect basic and initial features from the input image. The second layer in this CapsNet architecture is the primaryCaps layer, whose main role is to produce features from the lower 2D-convolutional layer's basic detected features. This layer contains capsules that have similar characteristics as the convolutional layer. Each capsule has a convolutional kernel size of 5x5 and a stride of 2. The next six layers consist of consecutive digitCaps that take as input the tensors produced by the primaryCaps layer. These digitCaps layers have two 16-dimensional capsules followed by four 16-dimensional capsules, four 32-dimensional capsules, eight 64-dimensional capsules, eight 32-dimensional capsules, and the last layer of three 16-dimensional capsules. We used three capsules in the last digitCaps layer because we aim to classify the ROIs of mammograms into three classes, normal, benign, and malignant. For the dynamic routing between capsules in both types of layers, primaryCaps, and digitCaps, we opted for routing with three iterations as it has shown good results in the original paper [16]. In this work and as presented in Fig. 1, we aim to test the performance of three CapsNet architectures in classifying suspicious lesions in mammograms. These three architectures differ in the number of kernels used in the convolutional layers and the number of capsules in the primaryCaps layer; however, they have the same consecutive six digitCaps layers with the same number and size of capsules and with the same routing number. We have changed the number of kernels in the convolutional layer and the number of capsules in the primaryCaps as they highly impact the basic and initial features extracted from these two layers. The three CapsNet architectures proposed and tested in this paper are given as follows:

- CapsNet16 has 16 kernels in the 2D-convolutional layer and two 8-dimensional capsules in the primaryCaps layer (cf. Fig. 1a).
- CapsNet32 has 32 kernels in the 2D-convolutional layer and four 8-dimensional capsules in the primaryCaps layer (cf. Fig. 1b).
- CapsNet64 has 64 kernels in the 2D-convolutional layer and eight 8-dimensional capsules in the primaryCaps layer (cf. Fig. 1c).

3.3. CapsNet loss function

The loss function of a capsule network, also known as margin loss, is similar to the Support vector machine loss. The length output vectors from higher-level capsules present a class's probability of existence in a capsule k. Hence, the class k should have a long output vector if only the objects presented by that class exist in the image. The last digitCaps layer produces k 16-dimensional vectors. During the training process, a loss value is computed for each of the k vectors; then, these losses are summated together into one final loss (Eq. (6)), which will be represented with a one-hot encoded vector. The margin loss function is composed of two L2 norm terms; the term $T_k \max\left(0, m^+ - \|V_k\|^2\right)$ is calculated for correct digitCaps. T_k equals one if the digitCaps is correct and $\max\left(0, m^+ - \|V_k\|^2\right)$ equals zero when given correct predictions with

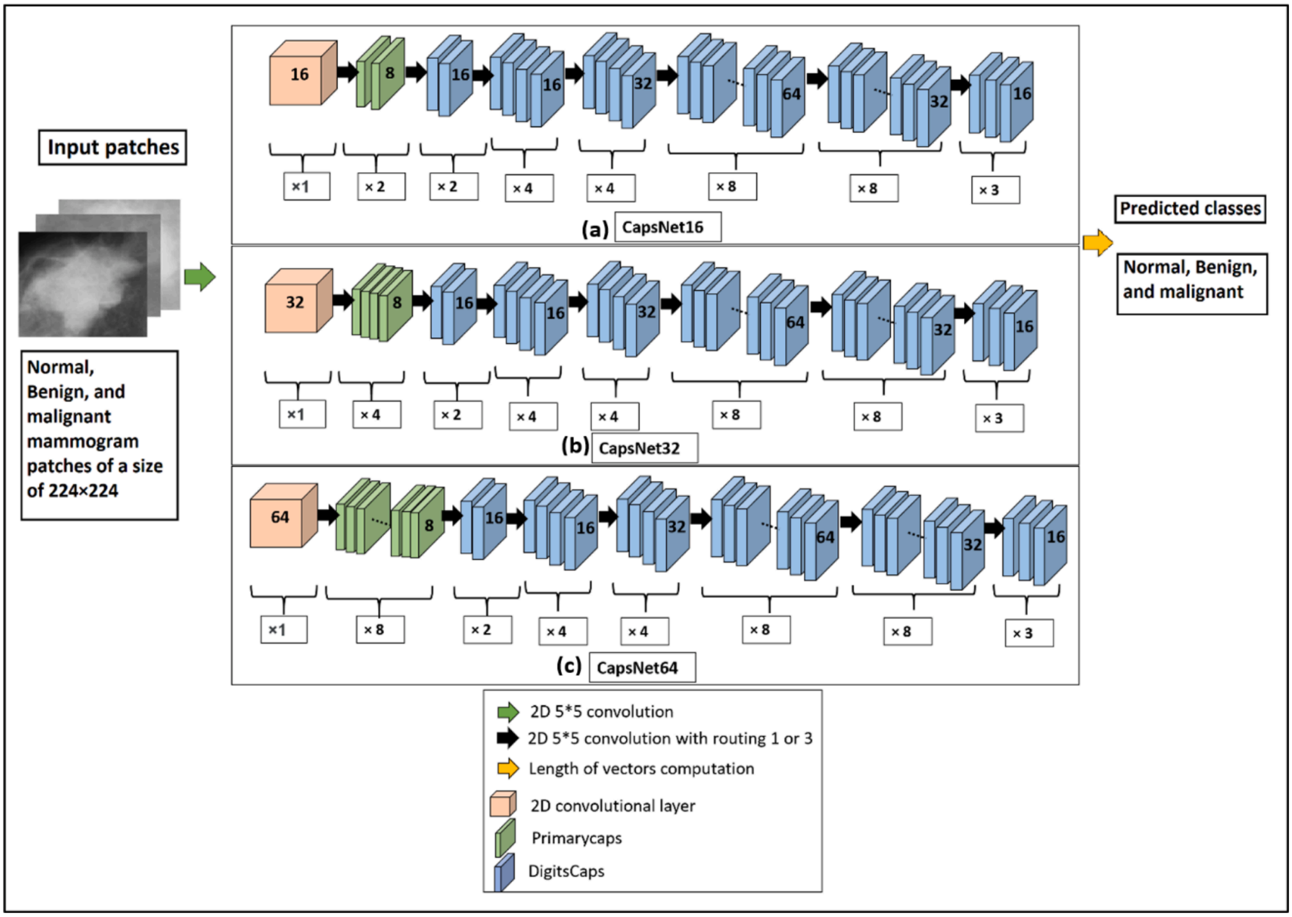


Fig. 1. Flowchart of the proposed method.

a probability greater than m^+ value, which is set by default to 0.9. The second term of the margin loss, $\lambda(1-T_k)\max(0,\|V_k\|-m^-)^2$ corresponds to the incorrect digitCaps predictions. In contrast to the first term, the $(1-T_k)$ term equals 1 in case of incorrect digitCaps predictions. Whereas $\max(0,\|V_k\|-m^-)^2$ equals zero given incorrect predictions with a probability value less than the m^- value, which is set by default to 0.1. The λ parameter decreases the weight of the loss in the case of absent classes, which prevents the initial learning weights from shrinking the lengths of the output vectors of all digitCaps capsules. In this work, we used $\lambda = 0.5$.

$$L_k = T_k \max\left(0, m^+ - \|V_k\|^2\right) + \lambda(1-T_k)\max(0,\|V_k\|-m^-)^2 \tag{6}$$

Where L_k is the margin loss of digitCaps k, T_k equals 1 in case of a correct prediction, V_k is the output vector of digitCaps k, $m^+ = 0.9$, and $m^- = 0.1$ parameters are set to prevent the length of digitCaps output vectors from maximizing out or collapsing, and λ is a parameter used for down-weighting the initial weights in cases of absent classes.

3.4. Evaluation metrics

In this work, classification evaluation metrics are the accuracy, F1-score, precision, recall, specificity, and Area Under the Receiver Operating Characteristics (AUC). These metrics are briefly defined below.

Accuracy metric is one the most intuitive metrics used to measure the classification performance. It is defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Where TP, TN, FN, and FP are the true positive, true negative, false negative, and false-positive counts, respectively.

The F1 score is computed based on the precision and recall metrics, including both false positives and false negatives. It is given by:

$$\mathrm{F1} = 2* \frac{Precision*recall}{precision + recall} \tag{8}$$

Where precision and recall are given by the following equations:

$$\mathrm{precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\mathrm{recall} = \frac{TP}{TP + FN} \tag{10}$$

The specificity measures the classifier’s ability to correctly identify the true negatives. It is given as follows:

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

The area under this curve, AUC, measures the ability of a classifier in distinguishing between classes based on the Receiver Operating Characteristic (ROC) curves, which plot the tradeoffs between the recall also known as true positive rate (TPR) and 1-specificity known as the false positive rate (FPR). The AUC can be presented by:

$$AUC = \int_{x=0}^{1} TPR\left(FPR^{-1}(x)\right)dx \tag{12}$$

Where x represents the varying parameter of the AUC.

4. Performance evaluation

4.1. Datasets

In this work, we used three public databases of digital mammograms to train the proposed architecture of the novel CapsNet model, Digital Database for Screening Mammography (DDSM), Curated Breast Imaging Subset of DDSM (CBIS-DDSM), and (INbreast). DDSM [49,50] contains a set of mammograms belonging to normal, benign, or malignant cases. Each case belongs to one patient and contains a description of the case, such as the breast density rate (ACR), the location and type of the abnormal masses, and their corresponding ground truth annotations. CBIS-DDSM [51] is an updated version of the conventional DDSM database, which contains a subset of abnormal cases of the DDSM databases that were organized into training and testing sets. INbreast [52] is a set of digital mammograms acquired from the University Hospital, S. João [CHSJ] located in Porto. It includes both normal and abnormal cases provided with their corresponding density rate. In this work, normal ROI patches are extracted from the DDSM database while the abnormal ones are acquired from the CBIS-DDSM subset. Hence, we will refer to the ROIs extracted from these two datasets by DDSM images as they originally belong to the same database, DDSM.

As mentioned before, the normal cases in the DDSM and INbreast do not have pre-defined ROI regions. Thus, we extract the patches that contain the normal breast tissue information from both datasets using the process explained in Fig. 2. As shown in Fig. 2, first, a preprocessing step is conducted to remove the artifacts and the pectoral muscle from the acquired mammograms. This helps keep the breast region only and remove the areas that do not belong to the breast tissue. This step is

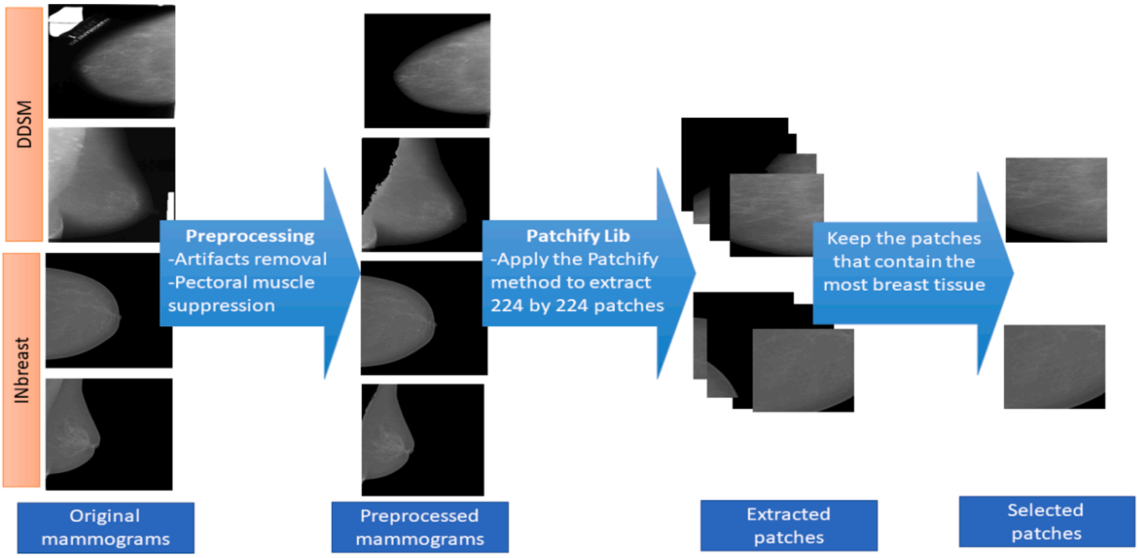

DDSM
INbreast
Preprocessing
-Artifacts removal
-Pectoral muscle suppression
Patchify Lib
-Apply the Patchify method to extract 224 by 224 patches
Keep the patches that contain the most breast tissue
Original mammograms
Preprocessed mammograms
Extracted patches
Selected patches

Fig. 2. Extraction process of patches from both DDSM and INbreast normal mammograms.

conducted using the preprocessing methods from our previous works [5,6,8]. Second, we use the Patchify library in Python to extract 224 by 224 patches from the previously selected breast regions. Finally, among the extracted patches, we select those that mostly contain the breast tissue and discard those that mainly have background pixels.

As shown in Table 1, in total, there are 2358 images that contain suspicious masses acquired from the DDSM database, from which 912 belong to the benign class, 784 to the malignant class, and 662 to the normal class. In addition, we ensured to select ROIs from each category of the breast density; thus, from the acquired DDSM ROIs, 496 belong to the D1 breast density, 924 to D2, '619 to D3, and 319 to D4. From the INbreast, 160 ROIs of suspicious masses are obtained, from which 51 are benign, 50 are malignant, and 59 are normal. From these INbreast ROIs, 42 belong to the D1 breast category, 33 to D2, 18 to D3, and 67 to D4.

4.2. Experimental results

The acquired mammograms from both DDSM and INbreast mammograms are split into a training set representing 90% of the data, and the remaining 10% are allocated to the test data. The training set is also divided into training and validation data during the 5-fold cross-validation training process. As mentioned before, in this paper, we propose a new architecture of the novel deep learning model CapsNet that significantly minimizes the number of parameters and improves the classification accuracy. The investigated models are CapsNet16, CapsNet32, and CapsNet64. To compare these models' performance in multi-classifying breast masses in mammograms with other deep learning models, we implemented the top 18 best models that give the highest accuracy values in classifying the conventional ImageNet images. These models are: ResNet50V2, ResNet101V2, ResNet152V2, NasNetLarge, VGG16, VGG19, MobiNetV2, Xception, EfficientNetB0, EfficientNetB1, EfficientNetB2, EfficientNetB3, EfficientNetB4, EfficientNetB5, EfficientNetB6, EfficientNetB7, InceptionResNetV2, and InceptionV3.

Table 1
Number of the used mammograms from each dataset per class.

Datasets/classes	Normal	Benign	Malignant	Total
DDSM	662	912	784	2358
INbreast	59	51	50	160
Total	721	936	834	2518

Table 2
Number of parameters and computational time per epoch of the investigated deep learning models.

	Number of parameters in millions	Training computational time per epoch
ResNet50V2	24 M	16 s 295 ms/step
ResNet101V2	43 M	29 s 518 ms/step
ResNet152V2	58 M	43 s 766 ms/step
NasNetLarge	85 M	51 s 902 ms/step
VGG16	14 M	14 s 243 ms/step
VGG19	20 M	16 s 283 ms/step
MobiNetV2	2 M	16 s 285 ms/step
Xception	21 M	32 s 572 ms/step
EfficientNetB0	4 M	19 s 334 ms/step
EfficientNetB1	7 M	40 s 719 ms/step
EfficientNetB2	8 M	42 s 758 s/step
EfficientNetB3	11 M	55 s 979 s/step
EfficientNetB4	18 M	72 s 1 s/step
EfficientNetB5	29 M	113 s 623 ms/step
EfficientNetB6	42 M	143 s 791 ms/step
EfficientNetB7	65 M	221 s 612 ms/step
InceptionResNetV2	54 M	33 s 588 ms/step
InceptionV3	22 M	16 s 288 ms/step
Original CapsNet model [13]	195 M	**485 s 268 ms/step**
Proposed CapsNet16	6 M	25 s 450 ms/step
Proposed CapsNet32	13 M	44 s 788 ms/step
Proposed CapsNet64	26 M	**86 s 2 s/step**

All the mentioned models, along with the three proposed CapsNet models, are trained using a 5-fold cross-validation for 60 epochs per fold, which is run two times using the Adam optimizer with a learning rate of 0.0001. The training is conducted based on the GPU Nvidia K80s for a batch size of 32 except for the original CapsNet, NasNetLarge, EfficientNetB5, and EfficientNetB6 models that require a small batch size of 10 and 5 for the EfficientNetB7 model to be trained on such GPU. Table 2 presents the number of parameters and the average training time per epoch for each of the previously mentioned models. As one can notice and using the mammogram's input size of (224, 224), the original CapsNet model [13] comes with the highest number of parameters of 195 M and a computational time of 485 s per epoch. Whereas our proposed CapsNet64 model having the maximum number of kernels investigated in this work has 26 M parameters only and a computational time of 86 s per epoch, 2 s per step. Thus, the proposed CapsNet64 architecture reduces by 7 times the number of parameters of the original CapsNet model and decreases by 6.5 times its corresponding training time per epoch.

The three proposed CapsNet models are trained from scratch. Their corresponding classification performance is investigated based on two training approaches: with data augmentation (CapsNet16_aug, CapsNet32_aug, and CapsNet64_aug) and without data augmentation (CapsNet16, CapsNet32, and CapsNet64). The used data augmentation operations include horizontal flip, vertical flip, and rotation range (0.15). Table 2 gives the classification average results of the six investigated CapNet models using 5-folds cross-validation training. As one can observe, the CapNet64_aug model shows the best validation results with an accuracy of 87.23% +/- 0.13, a F1-score of 88.93% +/- 0.11, precision and recall values of 83.33% +/- 0.15 and 95.34% +/- 0.04, respectively, and a specificity of 88.50% +/- 0.12. In comparison with the other considered five CapsNet models, the CapNet64_aug model has the largest kernel size of 64, a capsule dimension of 8, and a number of capsules of 8. This means the higher the numbers of kernels and capsules in a CapsNet model, the better its multi-classification performance. The models trained with the data augmentation approach give better performance as it expands the size of the training samples by creating variations of the used mammograms, which helps the models generalize their learned weights to new mammograms and hence improve their classification performance. Therefore, the 18 models included in this comparison are trained based on the same data augmentation process as the proposed CapsNet model.

Table 3
Multi-classification metrics' results for 5-folds cross-validation.

CapsNet models	Mean accuracy % (+/- std)	Mean f1% (+/- std)	Mean precision % (+/- std)	Mean recall% (+/- std)	Mean specificity % (+/- std)
CapsNet16	76.11% (+/- 0.13)	77.59% (+/- 0.12)	70.56% (+/- 0.17)	86.18% (+/- 0.04)	75.88% (+/-0.13)
CapsNet16_aug	78.92% (+/- 0.12)	81.53% (+/- 0.10)	72.97% (+/- 0.14)	92.38% (+/- 0.05)	80.53% (+/- 0.12)
CapsNet32	80.96% (+/- 0.15)	83.24% (+/- 0.13)	76.89% (+/- 0.18)	90.74% (+/- 0.04)	81.05% (+/- 0.15)
CapsNet32_aug	82.61% (+/- 0.13)	84.22% (+/- 0.12)	77.21% (+/- 0.16)	92.64% (+/- 0.07)	84.02% (+/- 0.12)
CapsNet64	85.96% (+/- 0.14)	87.55% (+/- 0.12)	83.37% (+/- 0.16)	92.17% (+/- 0.06)	86.48% (+/- 0.13)
CapsNet64_aug	**87.23% (+/- 0.13)**	**88.93% (+/- 0.11)**	**83.33% (+/- 0.15)**	**95.34% (+/- 0.04)**	**88.50% (+/- 0.12)**

Among the six proposed CapsNet models investigated in this paper, the CapsNet64_aug model has the best cross-validation performance in classifying breast masses into normal, benign, and malignant (cf. Table 3). Hence, in the presentation of the multi-classification test results, we consider only those of the CapsNet64 model when compared to other deep learning models. The 18 models that are compared to the proposed CapsNet model are trained from scratch using the same training parameters including the learning rate, optimizer, number of epochs, and data augmentation process. As Table 4 shows, the proposed CapsNet model takes the lead in classifying breast masses into normal, benign, and malignant compared to the other 18 implemented models. This model attains an accuracy of 77.78%, an F1-score of 77.45%, a precision and recall value of 71.54% and 84.54%, respectively, and a specificity of 83.15%. Then, the Inception model comes second with accuracy, F1-score, precision, recall, and specificity of 76.19%, 76.11%, 76.46%, 75.78%, and 88.31%, respectively. In the third rank comes the InceptionResNetV2 model with an accuracy of 73.41%, an F1-score of 73.21%, a precision value of 73.21%, and a recall score 73.21%, and a specificity value of 86.61%.

Fig. 3 contains the test ROC curves and their corresponding AUC values for the proposed CapsNet64_aug model and the other 18 implemented models. The ROC curves are obtained by computing the micro-average of the true positive and false positive rates considering all the three classes (normal, benign, and malignant). In fact, in a multi-class classification problem and unlike the macro-average, the micro-average aggregates the contribution of all samples in the final score so we get more accurate results. As one can see from Fig. 3, the ROC curve of the CapsNet model is above the other models' ROC curves, which means it has a better classification performance than the other models. As we are comparing 19 ROC curves in one plot, we have added a zoomed version of the 19 ROC curves for better visualization of the top left side of the curves (cf. Fig. 3). As it can be seen, the ROC curve of the CapsNet model is the closest to the top-left corner of the plot, which indicates once again its good classification performance comparing to the other 18 models. This performance can be also concluded from the AUC values where the CapsNet64_aug model has the top score of 0.9088 followed by the InceptionV3 model with a value of 0.8929.

The test results of the multi-classification performance of the top 3 best models, CapsNet64, InceptionV3, and InceptionResNetV2, using both DDSM and INbreast databases and the breast density categories, are presented in Table 5 and Table 6, respectively. For the DDSM test mammograms, the proposed CapsNet64 model scores an average accuracy of 78.81%, an F1-score of 80.12%, a precision value of 74.95%, a recall score of 86.33%, and a specificity value of 85.16%. Whereas, the InceptionV3 model scores an accuracy, an F1-score, a precision, a recall, and a specificity of 77.12%, 79.07%, 79.23%, 78.91%, and 89.65%, respectively. Lastly comes the InceptionResNetV2 model with an average accuracy of 75.85%, an F1-score, a precision, and recall values of 77.08%, and a specificity score of 88.54%. Regarding the INbreast mammograms, the three models score lower metrics compared to those of the DDSM mammograms, the CapsNet64 and InceptionV3 scores an accuracy of 62.5%, yet the first one has higher F1-score, precision, and recall of 64.86%, 57.14%, and 75.00%, respectively against 58.06%, 60.00%, and 56.25% for those of the InceptionV3 model. While the InceptionResNetV2 model fails to classify most of the INbreast mammograms that can be shown by its low accuracy, F1-score, precision, and

Table 4
Test results of breast masses' multi-classification using different deep learning models.

	Test accuracy (+/- std%)	Test f1 score (+/- std%)	Test precision (+/- std%)	Test recall (+/- std%)	Test specificity (+/- std%)
ResNet50V2	63.89% (+/- 0.48)	63.79% (+/- 0.48)	63.90% (+/- 0.48)	63.67% (+/- 0.48)	82.06% (+/- 0.24)
ResNet101V2	59.13% (+/- 0.49)	59.15% (+/- 0.49)	59.25% (+/- 0.49)	59.04% (+/- 0.49)	79.72% (+/- 0.25)
ResNet152V2	61.11% (+/- 0.49)	60.94% (+/- 0.49)	60.94% (+/- 0.49)	60.94% (+/- 0.49)	80.47% (+/- 0.24)
NasNetLarge	68.65% (+/- 0.47)	68.66% (+/- 0.47)	68.80% (+/- 0.47)	68.53% (+/- 0.47)	84.46% (+/- 0.24)
VGG16	67.86% (+/- 0.47)	67.77% (+/- 0.47)	68.03% (+/- 0.47)	67.52% (+/- 0.47)	84.18% (+/- 0.23)
VGG19	66.67% (+/- 0.47)	66.57% (+/- 0.47)	66.57% (+/- 0.47)	66.57% (+/- 0.47)	83.29% (+/- 0.24)
MobiNetV2	64.68% (+/- 0.48)	64.62% (+/- 0.48)	64.62% (+/- 0.48)	64.62% (+/- 0.48)	82.31% (+/- 0.24)
Xception	67.46% (+/- 0.47)	67.18% (+/- 0.47)	67.57% (+/- 0.47)	66.80% (+/- 0.47)	84.01% (+/- 0.23)
EfficientNetB0	66.67% (+/- 0.47)	66.41% (+/- 0.47)	66.41% (+/- 0.47)	66.41% (+/- 0.47)	83.20% (+/- 0.24)
EfficientNetB1	69.44% (+/- 0.47)	69.42% (+/- 0.47)	69.42% (+/- 0.47)	69.42% (+/- 0.47)	83.71% (+/- 0.24)
EfficientNetB2	67.46% (+/- 0.47)	67.41% (+/- 0.47)	67.41% (+/- 0.47)	67.41%	83.71% (+/- 0.24)
EfficientNetB3	63.49% (+/- 0.47)	63.34% (+/- 0.47)	63.34% (+/- 0.47)	63.34%	81.67% (+/- 0.24)
EfficientNetB4	66.27% (+/- 0.47)	66.29% (+/- 0.47)	66.29% (+/- 0.47)	66.29%	81.15% (+/- 0.24)
EfficientNetB5	71.03% (+/- 0.47)	71.39% (+/- 0.47)	71.69% (+/- 0.47)	71.09%	85.94% (+/- 0.23)
EfficientNetB6	70.64% (+/- 0.47)	70.42% (+/- 0.47)	70.71% (+/- 0.47)	70.15%	85.49% (+/- 0.23)
EfficientNetB7	68.65% (+/- 0.47)	68.61% (+/- 0.47)	68.76% (+/- 0.47)	68.47%	84.43% (+/- 0.23)
InceptionResNetV2	73.41% (+/- 0.44)	73.21% (+/- 0.44)	73.21% (+/- 0.44)	73.21% (+/- 0.44)	86.61% (+/- 0.22)
InceptionV3	76.19% (+/- 0.43)	76.11% (+/- 0.43)	76.46% (+/- 0.43)	75.78% (+/- 0.43)	88.31% (+/- 0.21)
Proposed CapsNet64_aug	77.78% (+/- 0.41)	77.45% (+/- 0.36)	71.54% (+/- 0.37)	84.54% (+/- 0.36)	83.15% (+/- 0.25)

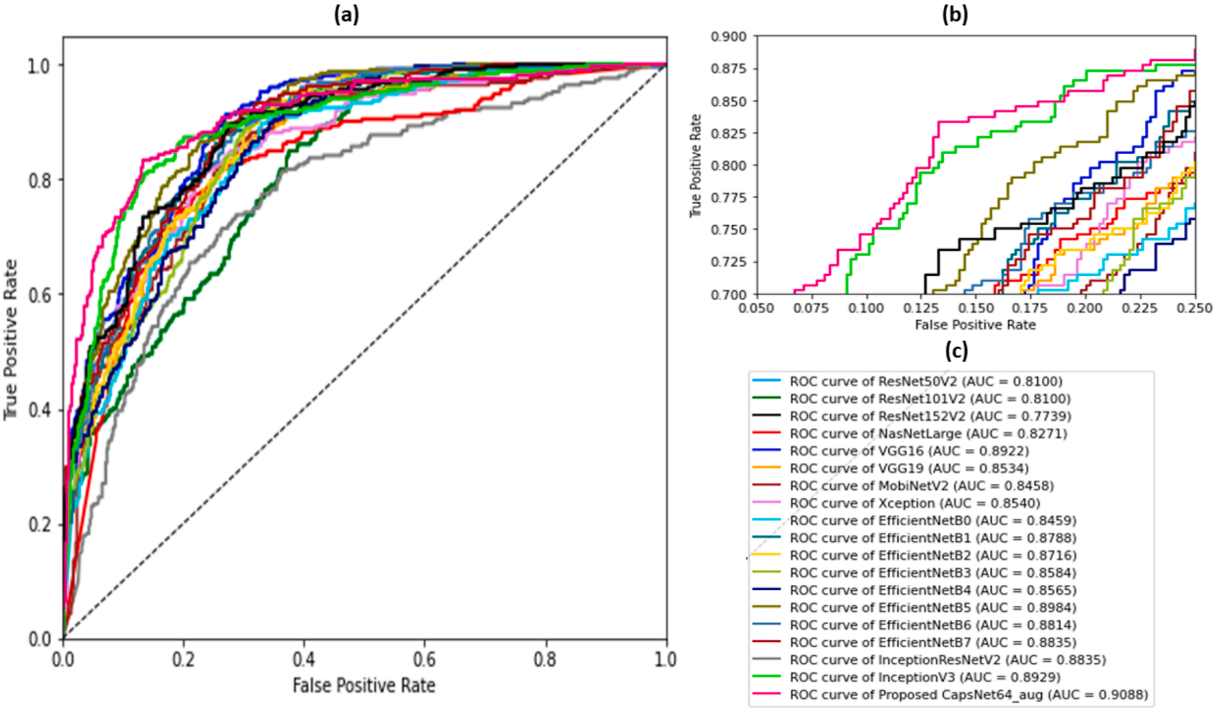


Fig. 3. (a) Multi-classes test micro average receiver operating characteristic curves of the implemented deep learning models, (b) Zoomed version of the multi-classes test micro average receiver operating characteristic curves of the implemented deep learning models, (c) legend for both curves in (a) and (b) and their corresponding micro average AUC values.

Table 5
Test results of breast masses' multi-classification using different deep learning models for the DDSM and INbreast databases.

	Datasets	Test mean accuracy% (+/- std)	Test mean f1 score % (+/- std)	Test mean precision% (+/- std)	Test mean Recall % (+/- std)	Test mean specificity% (+/- std)	Test macro- mean AUC (+/- std)
InceptionResNetV2	DDSM	75.85% (+/- 0.18)	77.08% (+/- 0.18)	77.08% (+/- 0.18)	77.08% (+/- 0.18)	88.54% (+/- 0.09)	0.875 (+/- 0.01)
	INbreast	37.50% (+/- 0.48)	37.50% (+/- 0.48)	37.50% (+/- 0.48)	37.50% (+/- 0.48)	68.75% (+/- 0.24)	0.835 (+/- 0.00)
InceptionV3	DDSM	77.12% (+/- 0.50)	79.07% (+/- 0.50)	79.23% (+/- 0.50)	78.91% (+/- 0.50)	89.65% (+/- 0.25)	0.892 (+/- 0.00)
	INbreast	62.50% (+/- 0.48)	58.06% (+/- 0.50)	60.00% (+/- 0.50)	56.25% (+/- 0.50)	81.25% (+/- 0.24)	0.855 (+/- 0.00)
Proposed CapsNet64_aug	DDSM	78.81% (+/- 0.41)	80.12% (+/- 0.35)	74.95% (+/- 0.37)	86.33% (+/- 0.37)	85.16% (+/- 0.25)	0.906 (+/- 0.00)
	INbreast	62.50% (+/- 0.48)	64.86% (+/- 0.41)	57.14% (+/- 0.41)	75.00% (+/- 0.40)	71.88% (+/- 0.30)	0.876 (+/- 0.00)

recall values of 37.50%.

Table 6 presents the test results of the multi-classification of breast masses in mammograms according to the four breast tissue density categories. The CapsNet64_aug model scores its best metric values for the D3 and D4 breast densities with an accuracy value of 75.41% and 91.11%, an F1-score of 72.04% and 93.12%, a precision value of 64.14% and 91.74%, a recall score of 82.17% and 94.59%, a specificity of 77.07% and 95.73%, and a macro average of the three classes AUC values of 0.87 and 0.968, respectively. The InceptionV3 shows the best multi-classification performance for D2 breast density mammograms with an accuracy, F1-score, precision, recall, specificity, and macro average AUC of 78.13%, 78.12%, 78.13%, 78.13%, 89.06%, and 0.89, respectively. The InceptionResNetV2 models give the best results in multi-classifying D1 breast density mammograms with accuracy, F1-score, precision, recall, specificity, and macro average AUC values of 76.00%, 76.39%, 76.39, 76.39%, 88.19%, and 0.902, respectively.

Breast masses in digital mammograms are often classified into two classes (normal/abnormal) or into three classes but in two stages in which masses are classified into normal and abnormal then into benign and malignant. Hence, one-stage multi-classification of breast masses still challenging and for comparison purposes we tested the performance of our model not only for multi-classification but also for binary classification of breast masses into normal and abnormal. In this paper, we present and investigate the novel CapsNet model's performance in classifying of suspicious masses in the breast into normal, benign, and malignant. However, for comparison purposes, we also conduct a binary classification of breast masses into normal and abnormal based on the CapsNet64_aug model that gives the best multi-classification results among the six CapsNet models investigated in this work. As Table 7 shows, the binary classification of breast masses using the CapsNet64_aug model reaches an accuracy value of 96.03%, F1-score of 96.27%, a precision value of 94.28%, a recall value of 98.38%, a specificity score of 93.97%, and a macro average AUC value 1. The model also shows better results in classifying DDSM mammograms with

Table 6
Test results of breast masses' multi-classification using different deep learning models regarding the breast density categories.

	Breast density category	Test mean accuracy% (+/- std)	Test mean f1 score% (+/- std)	Test mean precision% (+/- std)	Test mean recall% (+/- std)	Test mean specificity% (+/- std)	Test macro- mean AUC (+/- std)
InceptionResNetV2	D1	76.00% (+/- 0.43)	76.39% (+/- 0.43)	76.39% (+/- 0.43)	76.39% (+/- 0.43)	88.19% (+/- 0.22)	0.902 (+/- 0.00)
	D2	70.83% (+/- 0.45)	70.83% (+/- 0.45)	70.83% (+/- 0.45)	70.83% (+/- 0.45)	85.42% (+/- 0.23)	0.845 (+/- 0.00)
	D3	65.57% (+/- 0.47)	65.57% (+/- 0.47)	65.57% (+/- 0.47)	65.57% (+/- 0.47)	82.79% (+/- 0.24)	0,836 (+/- 0.00)
	D4	86.67% (+/- 0.34)	86.06% (+/- 0.34)	86.06% (+/- 0.34)	86.06% (+/- 0.34)	93.03% (+/- 0.17)	0.964 (+/- 0.00)
InceptionV3	D1	74.00% (+/- 0.44)	74.83% (+/- 0.44)	74.83% (+/- 0.44)	74.83% (+/- 0.44)	87.41% (+/- 0.22)	0.891 (+/- 0.00)
	D2	78.13% (+/- 0.41)	78.12% (+/- 0.41)	78.13% (+/- 0.41)	78.13% (+/- 0.41)	89.06% (+/- 0.21)	0.888 (+/- 0.00)
	D3	68.85% (+/- 0.46)	67.92% (+/- 0.47)	69.07% (+/- 0.47)	66.81% (+/- 0.47)	85.05% (+/- 0.23)	0.874 (+/- 0.00)
	D4	84.44% (+/- 0.36)	82.21% (+/- 0.36)	82.21% (+/- 0.36)	82.21% (+/- 0.36)	91.11% (+/- 0.18)	0.934 (+/- 0.00)
Proposed CapsNet64_aug	D1	74.00% (+/- 0.44)	74.79% (+/- 0.38)	70.34% (+/- 0.39)	79.86% (+/- 0.38)	83.25% (+/- 0.23)	0.924 (+/- 0.00)
	D2	75.00% (+/- 0.43)	74.53% (+/- 0.37)	68.11% (+/- 0.39)	82.29% (+/- 0.38)	80.73% (+/- 0.26)	0.881 (+/- 0.00)
	D3	75.41% (+/- 0.43)	72.04% (+/- 0.37)	64.14% (+/- 0.39)	82.17% (+/- 0.38)	77.07% (+/- 0.28)	0.872 (+/- 0.00)
	D4	91.11% (+/- 0.28)	93.12% (+/- 0.22)	91.74% (+/- 0.23)	94.59% (+/- 0.21)	95.73% (+/- 0.14)	0.968 (+/- 0.00)

Table 7
Test results of breast masses' multi-classification using different deep learning models regarding the breast density categories.

(a) Test results of the proposed CapsNet64_aug for binary masses' classification (Abnormal/normal)

Datasets	Test mean accuracy%(+/- std) %	Test mean f1 score%(+/- std) %	Test mean precision%(+/- std) %	Test mean recall%(+/- std) %	Test mean specificity%(+/- std) %	Test macro- mean AUC(+/- std) %
DDSM& INbreast	96.03% (+/-0.19)	96.27% (+/-0.14)	94.28% (+/-0.16)	98.38% (+/-0.12)	93.97% (+/-0.24)	0.997 (+/-0.00)

(b) Test results of the proposed CapsNet64_aug for binary masses' classification (Abnormal/normal) regarding both DDSM and INbreast databases

Datasets	Test mean accuracy%(+/- std) %	Test mean f1 score%(+/- std) %	Test mean precision%(+/- std) %	Test mean recall%(+/- std) %	Test mean specificity(+/- std) %	Test macro- mean AUC(+/- std) %
DDSM	96.61% (+/-0.17)	97.13% (+/-0.13)	95.53% (+/-0.14)	98.83% (+/-0.11)	95.31% (+/-0.21)	0.998 (+/-0.00)
INbreast	87.5% (+/-0.33)	88.24% (+/-0.26)	83.33% (+/-0.28)	93.75% (+/-0.24)	81.25% (+/-0.39)	0.986 (+/-0.00)

(c) Test results of the proposed CapsNet64_aug for binary masses' classification (Abnormal/normal) regarding all breast density categories

Breast density	Test mean accuracy%(+/- std) %	Test mean f1 score%(+/- std) %	Test mean precision%(+/- std) %	Test mean recall%(+/- std) %	Test mean specificity%(+/- std) %	Test macro- mean AUC(+/- std) %
D1	98.00% (+/-0.14)	97.69% (+/-0.15)	96.97% (+/-0.15)	98.44% (+/-0.14)	96.87% (+/-0.20)	0.999 (+/- 0.00)
D2	96.87% (+/-0.17)	97.96% (+/-0.07)	96.02% (+/-0.10)	100% (+/-0.00)	95.83% (+/-0.20)	0.998 (+/- 0.00)
D3	93.44% (+/-0.25)	94.06% (+/-0.22)	93.31% (+/-0.22)	94.83% (+/-0.22)	93.26% (+/-0.25)	0.996 (+/- 0.00)
D4	95.55% (+/-0.21)	95.21% (+/-0.10)	90.87% (+/-0.16)	100% (+/-0.00)	89.90% (+/-0.31)	0.997 (+/- 0.00)

accuracy, F1-score, precision, recall, specificity, and AUC values of 96.61%, 97.13%, 95.53%, 98.83%, 95.31%, and 1, respectively, against those of the INbreast mammograms with scores of 87.5%, 88.24%, 83.33%, 93.75%, 81.25%, and 0.99. The binary classification performance of the CapsNet64_aug model reaches good results for the D1 breast category mammograms followed by those of the D2, D4, and D3.

5. Discussions

Breast mass classification in digital mammograms is crucial for the early detection of breast cancer. Many works opted for machine learning methods where feature extraction techniques are used prior to the classification process to extract shape and texture descriptors from masses [5,6,37,39]. The efficiency of the classification process in machine learning depend on the feature extraction step, which needs to be selected carefully and used with optimal parameters. However, deep learning models learn and extract features accordingly during the training process without using any external technique. Hence the parameters that require optimization, in this case, are those of the model and compiler. In this context and by the emergence of the GPU era, deep learning models have been widely used in medical imaging applications, particularly for breast masses classification in the breast. Most of the used deep learning classification works conduct a binary classification of the suspicious breast masses into normal and abnormal or malignant and benign [36,41,43,44,46,47]. However, the classification of the normal cases should be included in the classification process so the model can also learn the characteristics of the fibro-glandular tissue that often gets misdiagnosed in the dense and extremely dense breast categories. Thus, some authors opted for a 2-level classification process where the suspicious masses are first classified into normal and abnormal; then, the abnormal cases are classified into benign and malignant [40,45]. Other authors used a one-stage classification process of suspicious breast masses into normal, benign, and malignant [39,48].

In this paper, we propose a one-stage classification of breast masses using a new CapsNet architecture. As mentioned previously, some methods conduct a multi-stage classification. For example, Khan et al [40] suggest a multi-level classification process based on a convolutional neural network. Their model classifies first the breast masses into normal and abnormal, then into benign and malignant. The features learned from both stages are then combined for a final prediction, and the best results of the proposed method are achieved using data augmentation. For the abnormal/normal classification stage, the model achieves an accuracy of 93.73%, a recall of 96.31%, a specificity of 90.47%, and an AUC of 0.934. Zhang et al. [45] also proposed a multi-level classification of breast masses using a multi-scale attention DenseNet model. In the first stage, masses are classified into normal and abnormal, then in the second stage into benign and malignant. The binary classification of masses into normal and abnormal achieves an accuracy of 94.92%, a recall of 96.52%, and AUC of 0.947. Our results show that our proposed CapsNet64_aug model outperforms existing deep learning models in classifying suspicious breast masses into normal and abnormal and attains an accuracy of 96.03%, an F1-score of 96.27%, a precision of 94.28%, a recall of 98.38%, a specificity of 93.97%, and an AUC of 0.997 (cf. Table 8).

Few works opted for a one-stage multi-classification of suspicious masses in the breast. These models learn the different breast tissues in one-stage simultaneously. Aboutalib et al. [48] proposed a multi-stage classification of breast masses based on a modified version of the AlexNet, which achieves an AUC value of 0.68. This score is low compared to that of the binary classification using the same model. Our proposed CapsNet model used for the multi-classification of breast masses shows promising performance and outperforms that of the existing methods with an accuracy of 77.78%, an F1-score of 77.45%, a precision of 71.54%, a recall of 84.54%, a specificity of 83.15%, and an AUC of 0.9 (cf. Table 8).

Our proposed CapsNet architecture takes advantage of the novel capsule concept that solves the limitation of the conventional CNN models and significantly reduces the training time, which makes it suitable also for large size ROIs like breast masses. Our proposed model shows good performance for binary classification and multi-classification of suspicious breast masses, particularly for extremely dense mammograms. In addition, our model is trained from scratch and outperforms the existing deep learning model classification. However, we believe that our new architecture can be expanded, and its corresponding parameters can be fine-tuned using a larger input size if trained using more powerful GPUs.

6. Conclusion

In this paper, we suggest an automatic method for breast masses classification in mammograms based on capsule-based deep learning model. Our proposed CapsNet architecture has seven times less computational times compared to the original CapsNet model. This makes the classification of suspicious masses ROIs with larger input size possible on the existing affordable GPUs. The proposed CapsNet model is trained from scratch using data augmentation, and its corresponding performance is compared to that of the top 18 best conventional deep learning models, including ResNet50V2, ResNet101V2, ResNet152V2, NasNetLarge, VGG16, VGG19, MobiNetV2, Xception, EfficientNetB0, EfficientNetB1, EfficientNetB2, EfficientNetB3, EfficientNetB4, EfficientNetB5, EfficientNetB6, EfficientNetB7, InceptionResNetV2, and InceptionV3. All these models were trained using the same training parameters as the proposed CapsNet model. The training and testing processes of our CapsNet model used mammograms from two public databases and took into consideration the four categories of breast density. The proposed model's binary classification (normal/abnormal) performance achieves an accuracy, an F1-score, a precision, a recall, a specificity, and AUC values of 96.03%, 96.27%, 94.28%, 98.38%,

Table 8
Comparison of proposed CapsNet model test results with related work deep learning results.

	Datasets	Predicted classes	Validation/ test	Accuracy	F1-score	Precision	Recall	Specificity	AUC
Milosevic et al. [53] 2015	MIAS	Normal/Abnormal	Test set	62%			20.4%	87.2%	
	Local database		Test set	83.7%			80.7%	86.7%	
Dong et al. [54] 2017	MIAS	Normal/Abnormal	Test set	94.14%					0.96
Yu et al. [55] 2020	MIAS	Normal/Abnormal	Test set	89.06%					
Ramadan et al. [56] 2020	MIAS	Normal/Abnormal	Test set	92.10%			91.40%	96.80%	0.95
Khan et al. [40] 2020	CBIS – DDSMAnd MIAS	Normal/Abnormal	Test set	93.73%			96.31%	90.47%	0.93
Zhang et al. [45] 2020	DDSM	Normal/Abnormal	Test set	94.92%			96.52%		0.95
The proposed CapsNet64_aug	DDSMand INbreast	Normal/Abnormal	Test set	96.03%	96.27%	94.28%	98.38%	93.97%	0.99
Jadoon et al. [39] **2017**	DDSM MIAS LLNL and RWTH	Normal, Benign, malignant	Validation set	83.74%	-	88.1%	88.8%	80.1%	0.85
Aboutalib et al. [48] **2018**	DDSM And FFDM	Normal, benign, malignant	**Test set**						0.68
The proposed CapsNet64_aug	**DDSM and INbreast**	**Normal, benign, malignant**	**Validation set**	**87.23% (+/- 0.13)**	**88.93% (+/- 0.11)**	**83.33% (+/- 0.15)**	**95.34% (+/- 0.04)**	**88.50% (+/- 0.12)**	
			Test set	**77.78%**	**77.45%**	**71.54%**	**84.54%**	**83.15%**	**0.90**

93.97%, and 0.997, respectively. In addition, the one-stage multi-classification results (normal, benign, and malignant) show an accuracy, an F1-score, precision, recall, specificity, and AUC of 77.78%, 77.45%, 71.54%, 84.54%, 83.15%, and 0.9, respectively. Our proposed CapsNet architecture achieves promising results and can be expanded and trained using more powerful GPUs.

CRediT authorship contribution statement

Khaoula Belhaj Soulami: Conceptualization, Methodology, Data curation, Software, Investigation, Writing – original draft. **Naima Kaabouch:** Validation, Visualization, Supervision, Project administration, Writing – review & editing. **Mohamed Nabil Saidi:** Conceptualization, Validation, Supervision.

Declaration of Competing Interest

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

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J Clin 68 (6) (Nov. 2018) 394–424.

[2] Eurostat, “Health Statistics: Atlas on Mortality in the European Union; Office for Official Publications of the European Union: Luxembourg.” 2009.

[3] B. Lauby-Secretan, et al., Breast-cancer screening–viewpoint of the IARC Working Group, N Engl J Med 372 (24) (Jun. 2015) 2353–2358.

[4] B.O. Anderson Breast Cancer—Thinking Globally Science 343 6178 Mar. 2014 1403 1403.

[5] K.B. Soulami, M.N. Saidi, A. Tamtaoui, A CAD System for the Detection of Abnormalities in the Mammograms Using the Metaheuristic Algorithm Particle Swarm Optimization (PSO), in: R. El-Azouzi, D.S. Menasche, E. Sabir, F. De Pellegrini, M. Benjillali (Eds.), Advances in Ubiquitous Networking 2, vol. 397, Springer Singapore, Singapore, 2017, pp. 505–517.

[6] K. B. Soulami, “A CAD system for the detection and classification of abnormalities in dense mammograms using electromagnetism-like optimization algorithm.,” pp. 1–8, 2017.

[7] K. B. Soulami, “Breast Cancer: Segmentation of Mammograms using Invasive Weed optimization and SUSAN algorithms.,” pp. 85–91, 2019.

[8] K.B. Soulami, Detection of breast abnormalities in digital mammograms using the electromagnetism-like algorithm, Multim. Tools Appl. 78 (10) (2019) 12835–12863.

[9] K.B. Soulami, An evaluation and ranking of evolutionary algorithms in segmenting abnormal masses in digital mammograms, Multim. Tools Appl. 79 (27–28) (2020) 18941–18979.

[10] K.B. Soulami, N. Kaabouch, M.N. Saidi, A. Tamtaoui, Breast cancer: One-stage automated detection, segmentation, and classification of digital mammograms using UNet model based-semantic segmentation, Biomedical Signal Processing and Control 66 (Apr. 2021), 102481.

[11] F. Mohanty, S. Rup, B. Dash, B. Majhi, M.N.S. Swamy, Mammogram classification using contourlet features with forest optimization-based feature selection approach, Multimed Tools Appl 78 (10) (May 2019) 12805–12834.

[12] T. Sadad, A. Munir, T. Saba, A. Hussain, Fuzzy C-means and region growing based classification of tumor from mammograms using hybrid texture feature, Journal of Computational Science 29 (Nov. 2018) 34–45.

[13] A.A. Shastri, D. Tamrakar, K. Ahuja, Density-wise two stage mammogram classification using texture exploiting descriptors, Expert Systems with Applications 99 (Jun. 2018) 71–82.

[14] A. Ghasemzadeh, S. Sarbazi Azad, E. Esmaeili, Breast cancer detection based on Gabor-wavelet transform and machine learning methods, Int. J. Mach. Learn. & Cyber., Jul. 10 (7) (2019) 1603–1612.

[15] G. E. Hinton, A. Krizhevsky, and S. D. Wang, “Transforming Auto-Encoders,” in Artificial Neural Networks and Machine Learning – ICANN 2011, Berlin, Heidelberg, 2011, pp. 44–51.

[16] S. Sabour, N. Frosst, G.E. Hinton, “Dynamic routing between capsules”, in Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 2017, pp. 3859–3869.

[17] W. Huang, F. Zhou, DA-CapsNet: dual attention mechanism capsule network, Sci Rep 10 (1) (2020).

[18] R. Mukhometzianov, J. Carrillo, CapsNet comparative performance evaluation for image classification, ArXiv (2018).

[19] X. Jiang, Y. Wang, W. Liu, S. Li, J. Liu, CapsNet, CNN, FCN: Comparative Performance Evaluation for Image ClassificationCapsNet, CNN, FCN: Comparative Performance Evaluation for Image Classification, Comparative Performance Evaluation for Image Classification” 9 (6) (2019) 840–848.

[20] X. Zhang, S.-G. Zhao, Cervical image classification based on image segmentation preprocessing and a CapsNet network model, International Journal of Imaging Systems and Technology 29 (1) (2019) 19–28.

[21] S. Bonheur, D. Štern, C. Payer, M. Pienn, H. Olschewski, M. Urschler, “Matwo-CapsNet, A Multi-label Semantic Segmentation Capsules Network” (2019).

[22] R. LaLonde and U. Bagci, “Capsules for Object Segmentation,” arXiv:1804.04241 [cs, stat], Apr. 2018, Accessed: Jan. 09, 2021. [Online]. Available: http://arxiv.org/abs/1804.04241.

[23] R. Karthik, R. Menaka, G.S. Kathiresan, M. Anirudh, M. Nagharjun, Gaussian Dropout Based Stacked Ensemble CNN for Classification of Breast Tumor in Ultrasound Images, IRBM (2021).

[24] W.-C. Shia, L.-S. Lin, D.-R. Chen, Classification of malignant tumours in breast ultrasound using unsupervised machine learning approaches, Sci Rep 11 (1) (2021).

[25] E.Y. Kalafi, A. Jodeiri, S.K. Setarehdan, N.W. Lin, K. Rahmat, N.A. Taib, M. D. Ganggayah, S.K. Dhillon, Classification of Breast Cancer Lesions in Ultrasound Images by Using Attention Layer and Loss Ensemble in Deep Convolutional Neural Networks, Diagnostics (Basel) 11 (10) (Oct. 2021) 1859.

[26] R. Woitek, et al., A simple classification system (the Tree flowchart) for breast MRI can reduce the number of unnecessary biopsies in MRI-only lesions, Eur Radiol 27 (9) (2017) 3799–3809.

[27] R. Fusco, M. Di Marzo, C. Sansone, M. Sansone, A. Petrillo, Breast DCE-MRI: lesion classification using dynamic and morphological features by means of a multiple classifier system, European Radiology Experimental 1 (1) (Jun. 2017) 10.

[28] H. Liu, H. Zhan, D. Sun, Y. Zhang, Comparison of BSGI, MRI, mammography, and ultrasound for the diagnosis of breast lesions and their correlations with specific molecular subtypes in Chinese women, BMC Medical Imaging 20 (1) (Aug. 2020) 98.

[29] K.J.W. Taylor, C. Merritt, C. Piccoli, R. Schmidt, G. Rouse, B. Fornage, E. Rubin, D. Georgian-Smith, F. Winsberg, B. Goldberg, E. Mendelson, Ultrasound as a complement to mammography and breast examination to characterize breast masses, Ultrasound Med Biol 28 (1) (2002) 19–26.

[30] A. Evans, R.M. Trimboli, A. Athanasiou, C. Balleyguier, P.A. Baltzer, U. Bick, J. Camps Herrero, P. Clauser, C. Colin, E. Cornford, E.M. Fallenberg, M. H. Fuchsjaeger, F.J. Gilbert, T.H. Helbich, K. Kinkel, S.H. Heywang-Köbrunner, C. K. Kuhl, R.M. Mann, L. Martincich, P. Panizza, F. Pediconi, R.M. Pijnappel, K. Pinker, S. Zackrisson, G. Forrai, F. Sardanelli, Breast ultrasound: recommendations for information to women and referring physicians by the European Society of Breast Imaging, Insights Imaging 9 (4) (2018) 449–461.

[31] R. Fadil, A. Jackson, B. Abou El Majd, H. el ghazi, and N. Kaabouch, “Classification of Microcalcifications in Mammograms using 2D Discrete Wavelet Transform and Random Forest,” Jul. 2020, pp. 353–359.

[32] S. Radhakrishna, S. Agarwal, P.M. Parikh, K. Kaur, S. Panwar, S. Sharma, A. Dey, K. K. Saxena, M. Chandra, S. Sud, Role of magnetic resonance imaging in breast cancer management, South Asian J Cancer 07 (02) (2018) 069–071.

[33] H. Ojeda-Fournier, C.E. Comstock, MRI for breast cancer: Current indications, Indian J Radiol Imaging 19 (2) (Jun. 2009) 161–169.

[34] H. Shahid The University of Texas Health Science Center San Antonio J.F. Wiedenhoefer The University of Texas Health Science Center San Antonio C. Dornbluth The University of Texas Health Science Center San Antonio P. Otto The University of Texas Health Science Center San Antonio K.A. Kist The University of Texas Health Science Center San Antonio An overview of breast MRI 7 13.

[35] A. Rodríguez-Ruiz, E. Krupinski, J.-J. Mordang, K. Schilling, S.H. Heywang-Köbrunner, I. Sechopoulos, R.M. Mann, Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System, Radiology 290 (2) (2019) 305–314.

[36] U. Bhosle, J. Deshmukh, Mammogram classification using AdaBoost with RBFSVM and Hybrid KNN–RBFSVM as base estimator by adaptively adjusting γ and C value, Int. j. inf. tecnol. 11 (4) (Dec. 2019) 719–726.

[37] R. Vijayarajeswari, P. Parthasarathy, S. Vivekanandan, A.A. Basha, Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform, Measurement 146 (Nov. 2019) 800–805.

[38] A.M. Alqudah, H.M.S. Algharib, A.M.S. Algharib, H.M.S. Algharib, Computer aided diagnosis system for automatic two stages classification of breast mass in digital mammogram images, Biomed. Eng. Appl. Basis Commun. 31 (01) (Feb. 2019) 1950007.

[39] M.M. Jadoon, Q. Zhang, I.U. Haq, S. Butt, A. Jadoon, Three-Class Mammogram Classification Based on Descriptive CNN Features, BioMed Research International 2017 (2017) 1–11.

[40] H.N. Khan, A.R. Shahid, B. Raza, A.H. Dar, H. Alquhayz, Multi-View Feature Fusion Based Four Views Model for Mammogram Classification Using Convolutional Neural Network, IEEE Access 7 (2019) 165724–165733.

[41] H. Li, S. Zhuang, D. Li, J. Zhao, Y. Ma, Benign and malignant classification of mammogram images based on deep learning, Biomedical Signal Processing and Control 51 (May 2019) 347–354.

[42] S.A. Agnes, J. Anitha, S.I.A. Pandian, J.D. Peter, Classification of Mammogram Images Using Multiscale all Convolutional Neural Network (MA-CNN), J Med Syst 44 (1) (Dec. 2019) 30.

[43] L.G. Falconi, M. Perez, W.G. Aguila, A. Conci, Transfer Learning and Fine Tuning in Breast Mammogram Abnormalities Classification on CBIS-DDSM Database, Advances in Science, Technology and Engineering Systems Journal 5 (2) (2020) 154–165.

[44] A.C. Perre, L.A. Alexandre, L.C. Freire, Lesion classification in mammograms using convolutional neural networks and transfer learning, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 7 (5–6) (Nov. 2019) 550–556.

[45] C. Zhang, J. Zhao, J. Niu, D. Li, R. Stoean, New convolutional neural network model for screening and diagnosis of mammograms, PLoS One 15 (8) (Aug. 2020) e0237674.

[46] R. Arora, P.K. Rai, B. Raman, Deep feature–based automatic classification of mammograms, Med Biol Eng Comput 58 (6) (Jun. 2020) 1199–1211.

[47] L. Tsochatzidis L. Costaridou I. Pratikakis Deep Learning for Breast Cancer Diagnosis from Mammograms—A Comparative Study J. Imaging 5 3 37.

[48] S.S. Aboutalib, A.A. Mohamed, W.A. Berg, M.L. Zuley, J.H. Sumkin, S. Wu, Deep Learning to Distinguish Recalled but Benign Mammography Images in Breast Cancer Screening, Clin Cancer Res 24 (23) (Dec. 2018) 5902–5909.

[49] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer, "The Digital Database for Screening Mammography," in Proceedings of the Fifth International Workshop on Digital Mammography, M.J. Yaffe, pp. 212–218.

[50] M. Heath, et al., Current status of the Digital Database for Screening Mammography, in: in Proceedings of the Fourth International Workshop on Digital Mammography, 1998, pp. 457–460.

[51] R.S. Lee, F. Gimenez, A. Hoogi, K.K. Miyake, M. Gorovoy, D.L. Rubin, A curated mammography data set for use in computer-aided detection and diagnosis research, Sci Data 4 (1) (2017).

[52] I.C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M.J. Cardoso, J.S. Cardoso, Inbreast: toward a full-field digital mammographic database, Academic radiology 19 (2) (2012) 236–248.

[53] M. Milosevic, D. Jankovic, A. Peulic, Comparative analysis of breast cancer detection in mammograms and thermograms, Biomed Tech (Berl) 60 (1) (Feb. 2015) 49–56, https://doi.org/10.1515/bmt-2014-0047.

[54] M. Dong, Z. Wang, C. Dong, X. Mu, Y. Ma, Classification of Region of Interest in Mammograms Using Dual Contourlet Transform and Improved KNN, Journal of Sensors 2017 (Nov. 2017), e3213680.

[55] X. Yu, W. Pang, Q. Xu, M. Liang, Mammographic image classification with deep fusion learning, Sci Rep 10 (1) (Sep. 2020) 14361.

[56] S.Z. Ramadan, Using Convolutional Neural Network with Cheat Sheet and Data Augmentation to Detect Breast Cancer in Mammograms, Computational and Mathematical Methods in Medicine 2020 (Oct. 2020), e9523404.