Mammogram Classification and Segmentation
Through Deep Learning

Diogo Lourenço Cortez, Jacinto Nascimento and Carlos Santiago

Instituto Superior Técnico
Department of Computer Science and Engineering
Av. Rovisco Pais 1, 1049-001 Lisbon, Portugal

Abstract—Given the extreme importance in early detection of breast cancer, a compelling search for Computer-Aided Detection (CAD) techniques drove Deep Learning (DL) researchers to investigate potential mammography screening applications. This work proposes the use of sophisticated Convolutional Neural Networks (CNNs) for classification and segmentation of mass and Micro-Calcification (MC) lesions on the publicly available INbreast dataset. Using an Attention Dense U-Net model, lesion segmentations are extracted with a Dice Coefficient of ($0.71 \pm 0.08$) for masses and ($0.58 \pm 0.05$) for MCs. As a result, classification using a Dense Multi-View model incorporating these predicted lesion segmentations shows competitive results regarding the State of the Art in fully-automated classification of breast screening exams (Normal, Benign, Malignant), achieving a 3-Class Mean AUC of ($0.79 \pm 0.06$).

Index Terms—Deep learning, Medical Imaging, Multi-View Models, Transfer Learning, Convolutional Neural Networks, Image Segmentation.

1 INTRODUCTION

Recent studies still acknowledge breast cancer as the leading cancer among women in every European country [8] and in the United States [31]. However, declining breast cancer mortality trends have been reported in most European countries [3], as well as in the United States [31]. This decline is mainly due to the combined effects of earlier detection and treatment [33].

Mammography is the only imaging test that has reduced breast cancer mortality [18]. Although Computer-Aided Detection (CAD) in mammography is a part of routine radiologist assistance in exam interpretation, there is controversy around lack of evidence in traditional CAD benefits to diagnostic performance [22]. Lack of performance can lead to unnecessary anxiety, invasive biopsy and potential unidentified lesions, so improvement in screening is needed to lessen harm [9].

Traditional CAD programs typically use handcrafted features to detect regions, but are unable to determine their clinical significance and actionability [39]. Moreover, traditional methods require expert definition of thresholds and are often not ideal [27], making this approach unappealing. In recent years, there has been a shift from these rule-based, problem-specific approaches into more generic ones, like DL. This new generation of CAD tools are arising in the wake of recent developments in DL, in particular, CNNs [13, 16, 19, 32, 35].

In recent studies, breast cancer detection indicates that the performance of radiologists when using this new CAD technologies show promise in the aid of practitioners [39]. Despite the fact that these models show success levels close to expert practitioners, hybrid performance of both expert radiologists and CAD tools exceed individual performance.

There are two distinct tasks in which these CAD systems are helpful: classification and detection. In breast cancer, classification tasks determine the severity of the screening, whereas detection tasks identify suspicious regions.

Regarding classification, the medical practise of analysing both Cranio-caudal (CC) and Mediod-lateral Oblique (MLO) views [20] has been successfully adapted for DL, formulating a Multi-View approach that leads to better performance [4, 39]. Moreover, these studies further boost their model performance by adding previously segmented lesion channels into the input.

Regarding segmentation, recent models are based on encoder-decoder architectures, of which the U-Net is example [30]. However, following recent breakthroughs in language translation [36], attention mechanisms established a dominant position in the encoder-decoder mappings for image applications [37, 38]. This methodology has been successfully applied in the field of mass segmentation [23, 34], where different styles of attention mechanisms are proposed to aid the U-Net style architectures. On one side, the grid-like attention mechanism of the Attention Dense U-Net [23] is inspired on Attention Gate proposed by [26], that suppresses irrelevant spatial feature activations. On the other, the channel-wise attention mechanism of the AUNet [34] resembles the Squeeze and Excitation attention present in [15], which scale important feature channels. The latter also employs Dense Upsampling [17] to enhance the decoder part of
the network.

2 PROPOSAL

The main contribution of this work lies in the development of the semi and fully-automated tasks in [4], with the use of more sophisticated models. More concretely, the applicability of Dense [16] core architectures was explored in the public INbreast dataset [25] for both classification and segmentation of potential lesions, providing a fair assessment of performance variation.

2.1 Dataset and Tools

The INbreast dataset consists in 410 annotated scans with precise segmentation maps of masses and Micro-Calcification (MC) lesions. Each scan corresponds to a single breast CC or MLO view of a mammography exam. These scans have an imbalanced distribution of 6 classes, corresponding to the Breast Imaging Report and Data System Score (BIRADS) [1]. Following the work in [4], these classes were reformulated into 3 classes: Normal, when BIRADS = 1; Benign, when BIRADS ∈ {2,3}; and Malignant, when BIRADS ∈ {4,5,6}. Both distributions follow in Figure 1.

![Figure 1. Original (Left) and Reformulated (Right) Dataset Class Distributions.](image)

2.2 Classification

In order to track the reaction of dense architectures to the typical breast imaging methodologies, a staged progression of the models was proposed as follows:

- (Baseline): Firstly, the baseline models 1-A and 1-B made use of solely CC and MLO scans, respectively, to predict the output classes.
- (Disjoint Scans): Model 2 tested the training of a single model on disjoint CC and MLO inputs.
- (Joint Scans): Inspired by the view topology in [39], model 3 investigated the impact of a Multi-View format via a dual input model (two branches for joint CC and MLO scans, respectively).
- (Auxiliary Data): Inspired by the incorporation of information channels in [4], models 4-A, 4-B, 5 and 7 explored the improvement of adding auxiliary binary masks of lesions (masses and MCs).
- (Channel Stacking): In order to validate the dual branch topology, model 6 tested the impact of junction by channel stacking on a single branch model, instead of the dual branch approach.

- (Pre-Training): Inspired by the Transfer Learning (TL) improvements in [4], the models 4-A*, 4-B*, 5*, 7* verified the boost of pre-training on the Imagenet dataset [6].
- (Full Automation): Contrary to previously mentioned items, that were trained using the dataset ground truth lesion masks, the most promising classification architectures, were later evaluated on fully-automated trials (as in [4]), by training and testing using segmented mass and MC lesions from the most promising segmentation models developed in this work.

Figure 2 depicts the training workflow of each classification model, for clarity.

![Figure 2. Training Workflow of Classification Models.](image)

Being a multi-class classification problem, training used cross-entropy as the loss function. After training, the proposed classification models were evaluated between each other according to the metrics of Area Under ROC Curve (AUC), Dice Coefficient (Equation (1)), Precision (Equation (2)) and Sensitivity (Equation (3)). It is to note that, by default, every 2-Class metric was extended to this multi-class problem via averaging in a one-vs-rest classification (where, sequentially, each class is denoted as positive and the remainder negative).

\[
\text{DiceCoefficient} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{1}
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{2}
\]

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}
\]

In addition to this internal evaluation, the best performing classification models will be subject to comparative evaluation regarding the work in [4], according to the metrics of Lesion Classification AUC (2-Class AUC of “Benign vs Malignant”, where it is assumed that all cases contain at least one mass or MC), Breast Screening AUC (2-Class AUC of “Normal/Benign” vs “Malignant”), Mean AUC (the same average of 2-Class AUCs as mentioned in the internal evaluation), Sensitivity, and finally, Specificity (the formula follows in Equation (4)).

\[
\text{Specificity} = \frac{TN}{TN + FP} \tag{4}
\]
2.3 Segmentation

Inspired by the work in [23, 34], this work proposed the following mass segmentation development:

- **(Mass Detection):** Firstly, models I-A and II-A verified the results of the Attention Dense-U-Net and Dense AU-Net on the INbreast dataset. The contribution is two-fold due to the fact that the Attention Dense-U-Net architecture [23] has no published results on the INbreast Dataset, while the work in [34] did not study the results of a dense backbone for the AU-Net and only accounted for cases containing mass lesions of the INbreast dataset.

- **(MC Detection):** Then, the same architectures were naively used for the detection of MCs, denoted by models I-B and II-B for the Attention Dense-U-Net and Dense AU-Net, respectively. This was motivated by the interest in holistic approaches, which is not the case for the work in [4] that made use of cascaded architectures for both mass and MC lesions [7, 24].

- **(Pre-Training):** Inspired by the TL improvements in [34], the models I-A*, II-A*, I-B*, II-B* verified the boost on previous models by pre-training in the Imagenet dataset. It is to note however that the pre-training in [34] was done on the CBIS-DDSM dataset [21].

Figure 3 depicts the training workflow of each segmentation model, for clarity.

![Figure 3. Training Workflow of Segmentation Models.](image)

Being a binary segmentation problem, training used a combined loss function (Equation (6)) that combines both Dice (Equation (5)) and Cross Entropy Losses for each pixel. This was inspired by the work in [34, 40], which states that the solely using Dice Loss leads to unstable optimization. Furthermore, the presence of a Dice Loss component forces the model to detect the positive class (Lesions), which suffers a great number imbalance in comparison to the background class. The mentioned equations are presented below, where $L_{\text{Dice}}$, $L_{\text{BCE}}$, $L_{\text{Combined}}$ denote the dice, cross entropy and combined loss, respectively; $y_{t,i,c}$ and $\hat{y}_{t,i,c}$ the true and predicted probabilities of pixel $i$ of sample $t$ being in the class $c$, accordingly; $C$ the amount of classes; and $N$ the amount of pixels.

$$L_{\text{Dice}}(y_t, \hat{y}_t) = 1 - \frac{2 \times \sum_{c=1}^{C} \sum_{i=1}^{N} (y_{t,i,c} \times \hat{y}_{t,i,c}) + 1}{\sum_{c=1}^{C} \sum_{i=1}^{N} (y_{t,i,c} + \hat{y}_{t,i,c}) + 1}$$  \hspace{1cm} (5)

$$L_{\text{Combined}}(y_t, \hat{y}_t) = L_{\text{Dice}}(y_t, \hat{y}_t) + L_{\text{BCE}}(y_t, \hat{y}_t)$$ \hspace{1cm} (6)

After training, the segmentation models were evaluated between each other using Dice Coefficient, Precision, Sensitivity, and $\Delta A$ (inspired by [34], having its formula in Equation (7), where $A_{\text{GT}}$ and $A_{\text{Pred}}$ represent the ground truth and predicted lesion area, respectively). The motivation for the chosen metrics was the focus on the severely under-represented positive class (lesions).

$$\Delta A = \frac{|A_{\text{GT}} - A_{\text{Pred}}|}{A_{\text{GT}}} = \frac{|(TP + FN) - (TP + FP)|}{TP + FN}$$ \hspace{1cm} (7)

After this internal evaluation, the mass segmentation models will also be subject to comparative evaluation in regarding the work in [23, 34], according to the metrics of Dice Coefficient, Sensitivity, Specificity, and $\Delta A$. It is to note that while the internal evaluation of mass segmentations was done on the whole dataset, the external evaluation was compared only on the INbreast samples that contained mass lesions. The motivation behind the comparison is due to the fact that the results can be comparable to the experiment in [34].

3 IMPLEMENTATION

3.1 Data Processing

The pre-processing in this work followed the same methodology as in [4], where mammograms were enhanced via local contrast normalization, followed by Otsu’s segmentation [28] to select a tight bounding box containing the breast region, and flipping, so that the pectoral muscle was always located on the right-hand side. To relieve computational and memory costs, images were then downsized via bi-cubic interpolation to 264x264 resolutions. Following the downsize, the data was normalized to have zero-mean and unit standard deviation.

Given the small size of the INbreast dataset, validation through three-way splitting would not be appropriate. Instead, validation through stratified k-fold cross-validation [11] (with k = 5) was used, and metrics were averaged between splits to provide a more meaningful model evaluation. It is also to note that the normalization step in pre-processing was applied to each training and testing splits separately, to stay truthful to the cross-validation methodology.

All models described in this Chapter were implemented in Python, assisted with publicly available libraries such as Keras [5]. Pre-training was also implemented using Keras, which provides public DL models with pre-trained weights on the Imagenet dataset [6].
In order to meet the demands of computational power, training was executed under the GPUs publicly available using the Google Colab platform [12].

3.2 Classification Model

Figure 4 depicts the architectures of both single and dual branch topologies. The model core is faithful to the original DenseNet-BC publication in [16], keeping the initial aggressive downsize of a 7x7 strided convolution followed by 3x3 max-pooling, bottleneck layers inside dense blocks and compression on transition layers. Besides that, the overall architecture drew inspiration from [4].

During pre-training trials, weight initialization was based on weights from a DenseNet-121 trained on the ImageNet dataset. Otherwise, weights were initialized with a normal distribution, following Xavier initialization [10]. During trials without pre-training, a DenseNet-37 was used instead. The structural difference lies in the number of core convolutional layers: DenseNet-121 and DenseNet-37 use a combination of (6,12,24,16) and (4,4,4,4) convolutions in dense blocks 1 through 4, respectively. The reasoning why architectures without pre-training use a DenseNet-37 instead of DenseNet-121, is due to overfitting. In preliminary trials, results shown lesser ability to generalize when using a DenseNet-121 without pre-training, even when applying regularization techniques such as dropout and weight loss, hence the downsacle.

Models leveraged from weight regularization and dropout to improve their generalization ability. L1 regularization was tested but had lesser effect than L2 regularization, which was used. Still facing sub-optimal generalization, dropout proved to match closer training and test performance, without the deterioration of overall performance.

When dealing with class imbalance, data augmentation through random rotation and flipping were tested. Furthermore, class weighting [29] was also experimented on. The latter delivered better results, even when using data augmentation generating balanced batches. Exponential weight decay also proved a better Learning Rate (LR) scheduler than a constant LR, and therefore was used. For the sake of completeness the formula for Exponential LR Decay (Equation (8)) is presented below, where \( \eta_p \) denotes the LR at training epoch \( p \) and \( \gamma \) the decay factor, respectively:

\[
\eta_p = \eta_0 \times \gamma^p
\]  

Training parameters were chosen based on training stability throughout models, so as to not give an unfair advantage to any in particular. In summary: the compression factor and growth rate of DenseNets were kept at recommended values 0.5 and 32, respectively [16]; starting LR set to \( 10^{-3} \); LR decay set to 0.95; batch size set to 32; dropout set to 0.2; training was optimized with Adam using \( (\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-7}) \); L2 factor set to \( 10^{-5} \) and \( 10^{-4} \), while training for 100 and 150 epochs in trials with and without pre-training, respectively.

The reason why models with and without pre-training had different weight regularization factors and training epochs is due to fairness regarding their weight nature. On one hand, the larger size of pre-trained models meant that they had much more learnable parameters, matched by a lower regularization factor. On the other, pre-trained models required less training epochs to converge.

3.3 Segmentation Model

The segmentation models made use of a U-Net like network employing attention and dense core to segment mammography image inputs. Figure 5 gives an overview of the architecture used for mass and MC detection. The dense core is the same as in Section 3.2, using a DenseNet-BC core (DenseNet-121 and DenseNet-37 for models with and without pre-training, respectively).

![Segmentation Model Architecture](image)

The overall shape is inspired by the work in [23, 34], while staying true to the dense core of [16]. This is controversial due to the fact that the architectures in [23, 34] use skip connections solely from the encoder blocks, which is impossible for the last decoder block of this work due to the aggressive downsize of the first convolution and pooling proposed by [16]. In this sense, the network was adapted to fit the U-Net style of skip connections via an additional convolutional path from the input. Another notable modification is the reduction...
in filter size after each decoder block by a factor of 2, inspired by the reduction of filter size seen in DenseNet-BC. This helped reduce the network to a more tangible size, while not significantly deteriorating performance.

Figure 6 presents the decoder blocks of both segmentation architectures, where the mentioned Attention Gate and Channel-Wise Attention are faithful to the corresponding original works [23, 34].

Figure 6. Attention Dense-U-Net (Left) and Dense AUNet (Right) Decoder Blocks.

As in Section 2.2, the dense core weight initialization of pre-training trials was based on weights from a DenseNet-121 trained on ImageNet, and otherwise according to Xavier initialization [10].

To improve the generalization ability of models, weight loss and data augmentation through random rotation was used. L2 weight regularization lead to stabler generalization improvements, while data augmentation helped twofold by increasing training size and improving detection of different lesion orientations that could occur between training and testing splits. The rotation ranged from angles $\theta \in [-20^\circ, 20^\circ]$ so as not to excessively distort the input.

Parameter choosing was again based on unbiased training stability. To summarize the differences: starting LR set to $10^{-4}$ and decay set to 0.96, to follow the otherwise step decay proposed in [34]; AUNet reduction factor in the channel wise attention mechanism was set to the recommended value of 16; batch size set to 8, due to memory constraints; L2 factor and epochs set to $10^{-5}$ and 100 for in trials with and without pre-training, respectively.

4 RESULTS

4.1 Classification Results

In order to visualize the impact on performance by each breast imaging methodology, the quantitative results for each model are presented in Figure 7 (hereinafter, FA-5* and FA-7* denote the fully-automated trials of models 5* and 7*, respectively). Furthermore, the comparison between the best performing models and the work in [4] is presented in Table 1 (where the column 'Auto' denotes if the trial is fully-automated or not, while 'LC AUC' and 'BS AUC' denote the Lesion Classification AUC and Breast Screening AUC, respectively).

In addition, some classifications of model FA-7* are depicted in Figure 8. This model leveraged auxiliary information channels containing lesion masks provided by models I-A* and I-B* for masses and MCs, respectively. The first three columns present correct classifications of each class, while the last column presents incorrect classifications. The colours red and orange denote ground truth masses and MCs, while blue and cyan regard the input segmentation masses and MCs, accordingly.

Figure 7. Evaluation Between Proposed Classification Models.

Figure 8. Model FA-7* Correct (a-c) and Incorrect (d) Classifications on Fully-Automated Test Cases.

To finish the experimental classification results, a summary of preliminary conclusions follow:

- (Mass and MC Masks): The inclusion of lesion masks substantially increased classification perfor-
Table 1. Comparative Classification Model Evaluation, regarding [4]

<table>
<thead>
<tr>
<th>Method</th>
<th>Auto</th>
<th>LC AUC</th>
<th>BS AUC</th>
<th>Mean AUC</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>[4]</td>
<td>No</td>
<td>0.94±0.05</td>
<td>0.91±0.08</td>
<td>0.87±0.08</td>
<td>0.69±0.28</td>
<td>0.92±0.08</td>
</tr>
<tr>
<td>5*</td>
<td>No</td>
<td>0.90±0.02</td>
<td>0.86±0.04</td>
<td>0.88±0.03</td>
<td>0.73±0.06</td>
<td>0.86±0.03</td>
</tr>
<tr>
<td>7*</td>
<td>No</td>
<td>0.88±0.04</td>
<td>0.86±0.05</td>
<td>0.86±0.06</td>
<td>0.70±0.08</td>
<td>0.84±0.05</td>
</tr>
<tr>
<td>[4]</td>
<td>Yes</td>
<td>0.78±0.09</td>
<td>0.86±0.09</td>
<td>0.72±0.10</td>
<td>0.66±0.14</td>
<td>0.69±0.23</td>
</tr>
<tr>
<td>FA-5*</td>
<td>Yes</td>
<td>0.77±0.08</td>
<td>0.74±0.10</td>
<td>0.75±0.06</td>
<td>0.58±0.08</td>
<td>0.76±0.05</td>
</tr>
<tr>
<td>FA-7*</td>
<td>Yes</td>
<td>0.86±0.07</td>
<td>0.85±0.03</td>
<td>0.79±0.06</td>
<td>0.61±0.04</td>
<td>0.80±0.03</td>
</tr>
</tbody>
</table>

Performance across all metrics, validating the information gained when using this technique.

- **(View Modality):** Architectures capable of understanding both CC and MLO views showed overall reduced metric variance. Moreover, the dual-branch junction topology proved notably better performance, when in comparison with the channel stacking approach. On the other hand, the Single-View models using ground truth lesion masks shown slightly better performance. This minor shift is justifiable by the size of the dataset, as Single-View models inherently have double the training data (each exam constitutes two samples).

- **(Transfer Learning):** TL actively helped model initialization, speeding up model convergence (fewer training epochs), and allowing increased model depth and, therefore, complexity.

- **(View Preference):** MLO-only models surprisingly outperformed CC-only models, in contrast to results in [39]. This is a mere remark, as the dataset size used in this work is substantially lower. Thus, this variation in performance should be a product of dataset variances. In addition, this feat is not representative of a real life scenario, where both views should be correctly assessed.

- **(Full Automation):** In fully-automated trials, the Multi-View topology showed substantial increase in performance, compared to the Single-View approach. This could be explained by the fact that the presence or absence of lesions detected in one view produce informative features relevant in the Multi-View classification (seen in the upper and lower exams of Figures 8a and 8c, respectively). In contrast, Figure 8d presents two over-sensitive classifications of this phenomenon.

### 4.2 Segmentation Results

In order visualize the impact on performance by each mass segmentation architecture, the quantitative results of each model are depicted in Figure 9. Moreover, the comparison of mass segmentation results regarding the State of the Art is presented in Section 4.2, where the column ‘Training Dataset’ and ‘Pre-Train’ denote the training and pre-training datasets of each trial, respectively. In trials without pre-training ‘-’ is used instead.

![Figure 9](image-url)  
**Figure 9.** Evaluation Between Proposed Mass Segmentation Models.

In order to visualize the impact on performance by each MC segmentation architecture, the quantitative results of the corresponding models are presented in Figure 10.

![Figure 10](image-url)  
**Figure 10.** Evaluation Between Proposed MC Segmentation Models.

For detail, the results of each MC segmentation model are tabulated in Table 3 (using the same column description as Section 4.2).

In order to visualize the segmentation results, some samples of predicted lesions by architecture are depicted in Figure 11. It is to note that, while the representation overlaps mass and MC lesions and joins both views for
Table 2. Comparative Mass Segmentation Model Evaluation, regarding [23, 34]

<table>
<thead>
<tr>
<th>Method</th>
<th>Training Dataset</th>
<th>Pre-Train</th>
<th>Dice Coefficient</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>∆A</th>
</tr>
</thead>
<tbody>
<tr>
<td>[34]</td>
<td>INbreast</td>
<td>-</td>
<td>0.64±0.08</td>
<td>0.66±0.07</td>
<td>-</td>
<td>0.52±0.21</td>
</tr>
<tr>
<td>I-A</td>
<td>INbreast</td>
<td>-</td>
<td>0.49±0.16</td>
<td>0.36±0.16</td>
<td>1.00±0.00</td>
<td>0.59±0.18</td>
</tr>
<tr>
<td>II-A</td>
<td>INbreast</td>
<td>-</td>
<td>0.40±0.10</td>
<td>0.35±0.08</td>
<td>0.98±0.02</td>
<td>0.65±0.31</td>
</tr>
<tr>
<td>[34]</td>
<td>INbreast</td>
<td>CBIS-DDSM</td>
<td>0.79±0.06</td>
<td>0.81±0.07</td>
<td>-</td>
<td>0.38±0.15</td>
</tr>
<tr>
<td>I-A*</td>
<td>INbreast</td>
<td>ImageNet</td>
<td>0.71±0.08</td>
<td>0.59±0.10</td>
<td>1.00±0.00</td>
<td>0.35±0.11</td>
</tr>
<tr>
<td>II-A*</td>
<td>INbreast</td>
<td>ImageNet</td>
<td>0.58±0.14</td>
<td>0.50±0.09</td>
<td>0.99±0.01</td>
<td>0.31±0.14</td>
</tr>
<tr>
<td>[34]</td>
<td>CBIS-DDSM</td>
<td>-</td>
<td>0.82±0.00</td>
<td>0.85±0.03</td>
<td>-</td>
<td>0.27±0.30</td>
</tr>
<tr>
<td>[23]</td>
<td>DDSDM [14]</td>
<td>-</td>
<td>0.82±0.01</td>
<td>0.78±0.08</td>
<td>0.85±0.09</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3. Detailed MC Segmentation Model Evaluation.

<table>
<thead>
<tr>
<th>Method</th>
<th>Training Dataset</th>
<th>Pre-Train</th>
<th>Dice Coefficient</th>
<th>Sensitivity</th>
<th>Precision</th>
<th>∆A</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-B</td>
<td>INbreast</td>
<td>-</td>
<td>0.52±0.06</td>
<td>0.44±0.07</td>
<td>0.66±0.04</td>
<td>0.35±0.08</td>
</tr>
<tr>
<td>II-B</td>
<td>INbreast</td>
<td>-</td>
<td>0.43±0.03</td>
<td>0.39±0.07</td>
<td>0.51±0.10</td>
<td>0.22±0.06</td>
</tr>
<tr>
<td>I-B*</td>
<td>INbreast</td>
<td>ImageNet</td>
<td>0.58±0.05</td>
<td>0.51±0.05</td>
<td>0.66±0.07</td>
<td>0.22±0.06</td>
</tr>
<tr>
<td>II-B*</td>
<td>INbreast</td>
<td>ImageNet</td>
<td>0.49±0.09</td>
<td>0.52±0.05</td>
<td>0.49±0.15</td>
<td>0.38±0.44</td>
</tr>
</tbody>
</table>

Figure 11. Lesion Segmentation Comparison by Model.

Concluding the experimental segmentation results of this work, the following itemization summarizes the preliminary conclusions:

- **(Transfer Learning):** Again, TL helped model initialization, allowing increased model depth, resulting in better overall performance. However, while the TL improvements using ImageNet [6] were substantial, they showed inferior performance than [34], which did pre-training using CBIS-DDSM [21]. This goes according to the fact that not only the INbreast and CBIS-DDSM datasets share similar domains (mammography images), but also due to the experiment in [34] sharing the tasks of both pre-trained and fine-tuned models (mass segmentation).

- **(Attention Module):** Regarding the architectures, the grid attention mechanism of the Attention Dense-U-Net showed greater performance than the channel-wise mechanism of Dense-AUNet. In particular, the precision for both mass and MC segmentation was substantially better and more robust. Interestingly, the sensitivity levels of both architectures was similar. This indicates that the Dense-AUNet was oversensitive by comparison. Additionally, this seems to explain the performance deviation of the Dense-AUNet for the experiment in this work and the one in [34], whose training was done exclusively on the exams containing masses.

- **(Mass vs MC Detection):** The smaller size and higher class imbalance, in principle, should make the detection of MC harder than the detection of masses. In fact, the difference in the results reflect these statements. However, while the metric performance is not overwhelming, the results seen in Section 4.1 seem to indicate that both mass and MC segmenta-
tions proved sufficient to improve fully-automated model in [4].

5 CONCLUSIONS

In an attempt to assist practitioners with CAD in the field of mammography screening, radiologist performance has been shown to improve using novel DL model architectures. In this work such tools were developed for the task of both classification and detection of lesions in mammography exams. More concretely, this work proposed sophisticated models that were separately capable of labelling unseen exams according to their lesion magnitude and detecting possible suspicious lesions. Furthermore, this work proposed the use of an holistic lesion detection DL model to provide competitive classification performance relative to the State of the Art, in accordance to the fully-automated classification of the INbreast dataset [25].

5.1 Contributions

Research in the field of DL is promising of a future with better assisting tools for practitioners in mammography screening. However, recent prominent work in the field makes use of relatively simple models [4] or require abundant private datasets to train [39]. In this sense, this work reviewed the evolution of DL models and algorithmic improvements to develop sophisticated models capable of the complex task of lesion classification and segmentation. Moreover, it proposed the development of such sophisticated models using the publicly available INbreast dataset [25] to provide a fair assessment of architectural performance.

A downside of the INbreast dataset is its inferior size, which, inspired by the work in [4], was compensated by the use of TL on the Imagenet dataset [6]. The data dependence of DL models was further relieved by the use of data augmentation and class weighting in lesion segmentation and classification, respectively. The latter of which surpassed the popular alternative of data augmentation, even when generating balanced batches.

Regarding the segmentation of mammograms, this work contributes by extending the work in [34] and [23] by evaluating the performance of adapted versions of the Attention Dense-U-Net and Dense-AUNet to detect mass and MC lesions in the INbreast dataset. The adaptation was faithful to the DenseNet core proposed in the original work [16], employing an initial aggressive convolution and pooling to reduce the feature size.

Regarding the proposed development of the semi-automated model in [4], no significant improvements can be stated due to the fact that the proposed models achieved similar performance. However, this is not the case for the development of the fully-automated model, which achieved significant performance improvements in both metric mean and variance. Regarding this upgrade, the main contributions lie in the proposal of an holistic DL approach capable of producing informative lesion segmentations that can be used for fully-automated exam classification. This is more appealing than the approach in [4], which made use of dedicated cascaded pipelines to automatically extract masses and MCs [7, 24].

5.2 Future Work

While the results seen in this work are appealing, some limitations were applied, which are left as suitable future work proposals. Namely, due to lack of computational ability, the inputs did not maintain the original high mammography resolutions. In fact, it is standard practise in DL to reduce the dimensionality to lower resolutions appropriate for training. This work reduced the input resolutions to 264x264 pixels. However, competitive works reduce the dimensionality to 512x512 representations [23, 39]. Future work could explore the use of such higher yet manageable resolutions, potentially validating the learnable downscaling proposed by the mentioned initial aggressive convolution and pooling present in the DenseNet architecture. On that note, this work did not fully explore the DenseNet core outside this faithful adaptation. Future work should also validate the performance of DenseNets which only downscale the inputs in transition layers, allowing for cleaner encoder-decoder skip connections, like in [23].

On the other hand, the INbreast dataset is not the only publicly available dataset used in research. For instance, future work could also explore the results of the proposed models in the CBIS-DDSM dataset [21], while also testing the performance according to the methodology in [34], which used pre-training on CBIS-DDSM and fine-tuning on the INbreast to detect masses.

Additionally, considering the performance improvement seen in this work when using feature informative channels, the unexplored inclusion of patient meta-data such as breast density is a suitable future work proposal. Moreover, this experiment explored Single and Multi-View topology networks, while not investigating the potential of Multi-Modality [2], which feature the inclusion of additional exam types such as Ultrasound or Magnetic Resonance Imaging for enhanced assessment of breast cancer.

REFERENCES


