Mammogram Classification and Segmentation through Deep Learning

Diogo Lourenço Cortez

Thesis to obtain the Master of Science Degree in Information Systems and Computer Engineering

Supervisors: Prof. Jacinto Carlos Marques Peixoto do Nascimento
Prof. Carlos Jorge Andrade Mariz Santiago

Examination Committee
Chairperson: Prof. Rui Filipe Fernandes Prada
Supervisor: Prof. Jacinto Carlos Marques Peixoto do Nascimento
Member of the Committee: Prof. Andreas Miroslaus Wichert

January 2021
Acknowledgements

Firstly, I would like to show gratitude towards my supervisions, Prof. Jacinto Nascimento and Prof. Carlos Santiago, for embarking on this project with me and providing insight throughout its evolution. Their share of knowledge, as well as cunning observations during development stages helped improving the quality that ultimately lies in this research.

Secondly, I would like to thank my parents and family for providing me with all the conditions needed to pursue the education I was privileged to receive. I indeed value the sustained efforts destined to increase my lifelong potential.

Finally, I would like to show appreciation towards my friends and colleagues who gladly endured my disposition and sharing my journey while studying at Técnico Lisboa. Knowing I found such lifelong charismatic people brings me joy. “Home is whenever I’m with you”.
Resumo

Dada a extrema importância na detecção antecipada de cancro de mama, uma busca preponderante de técnicas de diagnóstico auxiliado por computadores levou os investigadores de aprendizagem profunda a procurar possíveis aplicações no ramo de rastreio mamográfico. Neste trabalho são comparadas as abordagens tradicionais e as contemporâneas. Dado o sucesso destas últimas, são abordados modelos de aprendizagem profunda e os seus desafios no contexto do rastreio mamográfico. Além disso, este trabalho propõe o uso de treino com metodologias e arquitecturas de redes neurais convolucionais sofisticadas para classificação e segmentação de lesões no conjunto de dados INbreast, publicamente disponível. Usando um modelo com a arquitectura *Attention Dense U-Net*, a segmentação de lesões é feita com um coeficiente de Dice de $(0.71 \pm 0.08)$ para massas e de $(0.58 \pm 0.05)$ para micro-calculificações. Incorporando estas segmentações, a classificação segundo um modelo com arquitectura *Multi-Vista DenseNet* mostra resultados competitivos em relação ao estado da arte na classificação totalmente automatizada de exames de rastreio da mama (nas classes Normal, Benigno e Maligno), atingindo uma AUC média das três classes de $(0.79 \pm 0.06)$.

Palavras-chave: Aprendizagem Profunda, Imagiology Médica, Modelos Multi-Vista, Transferência de Conhecimento, Redes Neuronais Convolucionais, Segmentação de Imagem
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. In this work, traditional and novel approaches are compared. Given the success of the latter, DL models and challenges are thoroughly reviewed in the context of mammography screening. Moreover, this work proposes the use of sophisticated Convolutional Neural Network (CNN) model architectures and training for classification and segmentation of lesions publicly available in the INbreast dataset. Using an Attention Dense U-Net model, lesion segmentations are extracted with a Dice Coefficient of (0.71 ± 0.08) for masses and (0.58 ± 0.05) for Micro-Calcifications (MCs). As a result, the 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 ± 0.06).

Keywords: Deep learning, Medical Imaging, Multi-View Models, Transfer Learning, Convolutional Neural Networks, Image Segmentation
# Contents

Acknowledgments ................................................................. iii
Resumo .................................................................................. v
Abstract ................................................................................ vii
List of Tables .......................................................................... xi
List of Figures ......................................................................... xiii
List of Acronyms ..................................................................... xv

1 Introduction .............................................................................. 1
   1.1 Objectives ......................................................................... 2
   1.2 Thesis Outline .................................................................... 2

2 Theoretical Background ......................................................... 3
   2.1 Fundamental Concepts ....................................................... 3
      2.1.1 Computer-Aided Detection and Diagnosis ....................... 3
      2.1.2 Deep Learning ............................................................ 4
      2.1.3 Convolutional Neural Networks ..................................... 8
      2.1.4 Semantic Image Segmentation ..................................... 13
   2.2 Mammography Screening .................................................... 16
      2.2.1 Traditional Methods vs. Deep Learning ......................... 16
      2.2.2 Multi-View Mammography ........................................... 17
      2.2.3 Dataset Impact ........................................................... 17

3 Proposal .................................................................................. 19
   3.1 Task Description .............................................................. 19
   3.2 Classification ..................................................................... 19
   3.3 Segmentation ..................................................................... 21
   3.4 Validation ......................................................................... 23

4 Implementation ........................................................................ 24
   4.1 Dataset and Tools ............................................................. 24
   4.2 Classification Model .......................................................... 25
   4.3 Segmentation Model .......................................................... 27
List of Tables

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Training Hyper-Parameters of Classification Models.</td>
<td>26</td>
</tr>
<tr>
<td>4.2</td>
<td>Training Hyper-Parameters of Segmentation Models.</td>
<td>28</td>
</tr>
<tr>
<td>5.1</td>
<td>Comparative Classification Model Evaluation, regarding [14]</td>
<td>30</td>
</tr>
<tr>
<td>5.2</td>
<td>Comparative Mass Segmentation Model Evaluation, regarding [79, 127]</td>
<td>32</td>
</tr>
<tr>
<td>5.3</td>
<td>Detailed MC Segmentation Model Evaluation.</td>
<td>33</td>
</tr>
</tbody>
</table>
List of Figures

2.1 Example of a Neuron Activation. ......................................................... 4
2.2 Example of a Fully Connected Deep Neural Network. ............................ 5
2.3 Example of a Convolution Operation with Stride 1 and no Padding. .......... 8
2.4 Example of the Max-Pooling Operation with Stride 2. ............................ 9
2.5 Example of CNN Architecture for Classification. ................................. 9
2.6 AlexNet, as in [71]. ........................................................................... 10
2.7 Residual Learning, as in [42]. ............................................................... 10
2.8 Inception Module Blocks, as in [129]. .................................................. 11
2.9 DenseNet, as in [50]. .......................................................................... 12
2.10 Example of a Deconvolution Operation with Stride 1. ......................... 13
2.11 Fully Convolutional Network (FCN) architecture with skips, as in [82]. 14
2.12 SegNet architecture, as in [4]. ............................................................. 14
2.13 U-Net architecture, as in [108]. ........................................................... 14
2.14 Attention Dense U-Net, as in [79]. ....................................................... 15
2.15 AUNet, as in [127]. .......................................................................... 15
2.16 Mammography Processing with Traditional Machine Learning (ML) Methods, as in [30]. 16

3.1 Training Workflow of Classification Models. ........................................... 20
3.2 Training Workflow of Segmentation Models. ........................................... 22

4.1 Dataset Class Distribution. .................................................................... 24
4.2 Classification Model Architectures. .......................................................... 25
4.3 Segmentation Model Architecture. .......................................................... 27
4.4 Decoder Block Variations. ...................................................................... 27

5.1 Evaluation Between Proposed Classification Models. ............................ 29
5.2 Model FA-7* Correct (a-c) and Incorrect (d) Classifications on Fully Automated Test Cases. 30
5.3 Evaluation Between Proposed Mass Segmentation Models. ................... 31
5.4 Evaluation Between Proposed MC Segmentation Models. ...................... 32
5.5 Lesion Segmentation Comparison by Model. .......................................... 33
List of Acronyms

**AUC**  Area Under ROC Curve

**BIRADS**  Breast Imaging Report and Data System Score

**BN**  Batch Normalization

**CAD**  Computer-Aided Detection

**CC**  Craniocaudal

**CNN**  Convolutional Neural Network

**DL**  Deep Learning

**FCN**  Fully Convolutional Network

**LR**  Learning Rate

**MC**  Micro-Calcification

**ML**  Machine Learning

**MLO**  Mediolateral Oblique

**NN**  Neural Network

**ReLU**  Rectified Linear Unit

**SGD**  Stochastic Gradient Descent

**SIS**  Semantic Image Segmentation

**TL**  Transfer Learning
Chapter 1

Introduction

Recent studies still acknowledge breast cancer as the leading cancer among women in every European country [28] and in the United States [116]. However, declining breast cancer mortality trends have been reported in most European countries [13], as well as in the United States [116]. This decline is mainly due to the combined effects of earlier detection and treatment [13, 28, 113, 116, 120, 121].

Mammography is the only imaging test that has reduced breast cancer mortality [25, 69]. Although Computer-Aided Detection (CAD) in mammography is a part of routine radiologist assistance with image interpretation, a controversial study found that there was no evidence that traditional CAD improved diagnostic performance of mammography in any metric, and showed decrease in sensitivity when compared to diagnosis without CAD [76]. Lack of performance can lead to unnecessary anxiety, invasive biopsy and potential unidentified lesions, so improvement in screening is needed to lessen unnecessary harm [32].

Traditional CAD programs typically use handcrafted features to detect regions, but are unable to determine their clinical significance and actionability [5, 145]. Moreover traditional methods require expert definition of thresholds and are often not ideal [97], making this approach unappealing. In recent years, there has been a shift from these rule-based, problem-specific approaches into more generic ones, like Deep Learning (DL). This new generation of CAD tools are arising in the wake of recent developments in DL, in particular, Convolutional Neural Networks (CNNs) [42, 50, 71, 117, 130].

In recent studies, breast cancer detection compared the performance of radiologists when using this new CAD technologies and results indicate that these DL models show promise in the aid of practitioners [107, 145]. Despite the fact that these models show success levels close to expert practitioners, hybrid performance of 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.

Typical image classification tasks evaluate a single input image (Single-View). This is not true in mammography screening, where it is standard to evaluate two screenings for each breast of a patient (Multi-View). This analysis is comprised of two mammographic views of each breast: the Mediolateral
Oblique (MLO) and Craniocaudal (CC) views [73].

Multi-View approaches have been the de facto standard for competing DL models in recent mammography studies [14, 32, 125, 145]. Between these Multi-View approaches, different trends will be further discussed in this work, giving special focus to holistic approaches, following the work in [14].

1.1 Objectives

The ultimate goal of this work is to develop deep learning models to be used in a CAD system in order to assist practitioners in the evaluation of mammography exams. There are two objectives in this work, which will be directly evaluated with regard to the work in [14]:

1. (Classification) - The development of the semi-automated model, which comprises DL architectures that take as input breast cancer images along with the segmentations of lesions (Micro-Calcifications (MCs) and Masses) and output classes (Normal Tissue, Benign and Malignant);

2. (Segmentation) - The development of the fully-automated model, which comprises DL architectures that map the breast cancer screenings into masks of potential lesions (MCs and Masses).

The proposed solutions to these objectives will be explained in Chapter 3, while their implementation and evaluation will be further disclosed in Chapters 4 and 5, accordingly.

1.2 Thesis Outline

In order to provide the reader with the organization of this document, the following itemization lists a short summary of the discussed contents in each subsequent chapter:

• Chapter 2 familiarizes the reader with the fundamental knowledge to understand this work. It also addresses the current State of the Art and related work, firstly in the domain of DL, and afterwards with more concrete information in the domain of mammography screening.

• Chapter 3 features the proposed methodology used to accomplish the objectives mentioned in the previous section. In detail, it depicts the workflow, evaluation and validation methodology followed in the experiments.

• Chapter 4 describes how the proposed methodology was implemented. More specifically, it orderly gives insights on the dataset, tools and proposed models. Moreover, it overviews the training and architectural details used to materialize the proposed solutions.

• Chapter 5 presents the experimental results, with respect to the order of objectives. Following their presentation, remarks are made to justify expected and/or unexpected outcomes.

• Chapter 6 summarizes the conclusions of this work. Firstly, it overviews the experiment as a whole, following with contributions and downfalls, finally ending with future work.
Chapter 2

Theoretical Background

This Chapter provides the reader with the essential knowledge required to fully understand this work. Firstly, Section 2.1 extensively explores the fundamental building blocks and motivation of DL in this work, followed by Section 2.2, which further discusses the involvement of such applications in the field of mammography screening.

2.1 Fundamental Concepts

The covered concepts in this Section are disclosed in the following order: Section 2.1.1 explores the application of DL in the field of mammography screening, giving motivation to this field of study. Section 2.1.2 reviews the evolution of DL, showing the refinement of techniques throughout the years. Section 2.1.3 provides an overview of novel architectures that emerged in DL for computer vision tasks. Finally, Section 2.1.4 deepens the focus to image segmentation, in an effort to provide additional comprehensive tools to aid breast screening practitioners.

2.1.1 Computer-Aided Detection and Diagnosis

Although CAD is not limited to a single type of cancer, this work will focus on CAD systems designed to help radiologists in the routine detection and characterization of breast cancer in screening mammography.

In the United States, a recent study found that 20% of newly diagnosed breast cancers and almost one-third of interval cancers (those diagnosed within 1 or 2 years of being reported as normal, depending on whether the screening program is annual or biennial) are evident on previous mammograms [46]. CAD technologies emerged to aid radiologists in this missed exams, and even though they can be helpful, the body of evidence in support of CAD is equivocal [60].

A recent massive study on CAD helpfulness found that screening performance was not improved with CAD on any metric [76]. This raised some controversial questions considering the budget spent on traditional CAD systems, which the same study estimated to be over 400 million dollars a year, with no added value and in some cases decreased performance.
The main downfall of traditional CAD systems is their inability to access clinical significance and actionability of the found lesions [5, 145]. Thus, the need to improve these systems is pertinent.

Artificial Intelligent approaches (in particular, DL approaches) have shown promise for new CAD tools in many areas of radiology [17, 70, 103, 126, 140, 148]. In more recent studies, the authors of [107, 145] claim that DL algorithms are reaching comparable performance to that of radiologists. The advances to deep learning (disclosed in further sections), the adaptive learning nature of neural networks, and the positive results on screening mammography studies combined create a great motivation for new, more accurate and robust CAD decision support tools for mammographic screening as first or second readers.

2.1.2 Deep Learning

Deep learning is a subfield that derived from Machine Learning (ML) and Neural Networks (NNs), which date back to the 1940s. Although many new DL techniques emerged to this day, the building blocks of NNs remained the same.

NNs are a brain-inspired type of supervised learning, which is a technique that uses previously labelled data. The model tries to predict ground truth labels $y_t$, of inputs $x_t$, that follow a known distribution $(x_t, y_t) \sim \rho$, using a fixed function $F(x_t, \theta) = \hat{y}_t$. Optimizing the network parameters, $\theta$, corresponds to minimizing a loss function, $L(y_t, \hat{y}_t)$, that quantifies the disparity of ground truth and predicted labels.

In mathematical terms, the equation for the neuron excitation can be found in Equation (2.1). The input of sample $t$, $x_t$, is calculated by multiplying the input with the weights, $\theta$, and then passing through a non-linear activation function, $h$. Figure 2.1 illustrates this simple one layer NN model, the perceptron.

$$\hat{y}_t = F(x_t, \theta) = h(x_t, \theta) = h(\sum_{i=1}^{N} (x_{t,i} \times \theta_i))$$  \hspace{1cm} (2.1)

![Figure 2.1. Example of a Neuron Activation.](image)

To train the model parameters, $\theta$, Gradient Descent is applied. This is an iterative process that uses the gradient of the loss function, $L(y_t, \hat{y}_t)$, (defined in Equations (2.2) and (2.3)) to update the network parameters. The learning rate, $\eta$, is a hyperparameter that controls the amplitude of the updates.

$$\nabla L(\theta) = \left(\frac{dL}{d\theta_1}, \frac{dL}{d\theta_2}, ..., \frac{dL}{d\theta_N}\right)$$  \hspace{1cm} (2.2)

$$\theta^{\text{new}} = \theta^{\text{old}} - \eta \nabla L(\theta)$$  \hspace{1cm} (2.3)

If we consider the simple one layer perceptron, it becomes apparent that the model can only correctly classify linear separable problems. The XOR problem raised issues considering the limited mapping
ability of this methods, killing research in neural networks for a decade [89]. This problem was later solved by stacking multiple layers, and became the motivation for deeper architectures (deep reeers to the number of layers).

Deep neural networks stack many layers, allowing for multiple stages of non-linear transformations. These hierarchical architectures discover the representations needed for detection or classification. With higher abstract level representations, increasingly more complex functions can be learned [74]. The training of deep models follows Equation (2.2), with the application of chain rule differentiation for hidden layers and is commonly called Back-Propagation [109]. For the sake of completeness, Equation (2.4) and Figure 2.2 shows the archetype of a fully connected deep NN with K layers and N outputs.

\[ \hat{y}_t = F(x_t, \theta) = F_{f_{c,K}}(\ldots, F_{f_{c,2}}(F_{f_{c,1}}(x_t, \theta_1), \theta_2)\ldots, \theta_K) \]  

*(Figure 2.2. Example of a Fully Connected Deep Neural Network.)*

Linear regression is used to refer a slightly more sophisticated model with an additional parameter called intercept, \( b \). This translates Equation (2.1) into Equation (2.5). This parameter \( b \), hereafter called bias, extends the mapping into affine functions that do not necessarily cross the origin. Instead of the addition of the bias term, one can look at the model with only weights, where the entry \( x_{t,0} \) carries an extra entry, \( x_{t,0} \), that is always set to 1. This trainable parameter is also learned with back-propagation using Equations (2.2) and (2.3).

\[ \hat{y}_t = F(x_t, \theta) = h(x_t, \theta + b) = h(\sum_{i=1}^{N} (x_{t,i} \times \theta_i + b)) = h(\sum_{i=0}^{N} (x_{t,i} \times \theta_i)) \]  

With this basis, deep learning has considerably evolved due to the increase of computational power, and the broader availability of larger datasets that reduce the degree of difficulty for statistical generalization. Bellow follows a list of advanced algorithmic changes that propelled this evolution:

- **Activation Functions**: Modern literature recommends the use of the Rectified Linear Unit (ReLU) [92] as an activation function. It was already used in earlier works, but only began to be favoured in 2009 and after, due to studies pointing out that ReLU increased network performance and learning, when faced to networks that used other activation functions [34, 59, 92]. The ReLU activation function can be found in Equation (2.6):

\[ h(z) = \max(0, z) = \begin{cases} 0, & z \leq 0 \\ z, & z > 0 \end{cases} \]
Prior to the use of ReLU, researchers used the sigmoid activation function. Sigmoid functions suffer from the vanishing gradient problem. The gradient of small input values to sigmoid functions tends to get smaller as the gradients are propagated into backward hidden layers. This results in slow learning, making training NNs with sigmoid activation units strongly discouraged [101]. ReLU does not suffer from vanishing gradient [34]. The gradient will be 1 for positive inputs, and 0 for negative ones. The downside is that once the gradient is 0, the corresponding unit will become inactive in the network. This problem is commonly called Dying ReLU, and Leaky ReLU was one attempt at solving it. Instead of outputting 0 for negative inputs, Leaky ReLU introduces a small negative slope. This slope reduces sparsity but increases robustness for optimization [86].

Other two notable activations are the Exponential Linear Unit (ELU) [21], and the Maxout Neurons [36]. Even though they offer advantages over ReLU or Leaky ReLu, they require increased computation and parameters, respectively. In [34, 71, 90, 101], authors discuss advantages and disadvantages of different activation functions. Modern mammography screening studies commonly use ReLU, as this activation function allows for faster learning while keeping satisfactory performance [14, 32, 77, 84, 125, 145].

- **Learning Rate**: Learning Rate (LR) is one of the most important hyperparameters in DL [9]. High values of LR can provide divergent learning behaviour, while low values can converge but rather slowly. To deal with this issue, it is common practice to start with a higher LR value and gradually lower it during training. This learning policy is called LR decay, of which step, quadratic, square root and linear decay are examples [90].

Another way to hasten training is to optimize the gradient descent. In practice, Stochastic Gradient Descent (SGD) [64] is the method used to approximate this optimization strategy. Many fancier variations of SGD such as Momentum [128], AdaGrad [24], RMSprop [135] and Adam [65] have been developed in efforts to fasten the convergence of the network parameters, by tweaking with the vanilla learning rule (Equation (2.3)). AdaGrad and RMSprop stem from the same method that computes adaptive LRs for each parameter. On the other side, Momentum accelerates SGD in relevant directions, dampening the zig-zag effect of vanilla update steps. Adam mixes the adaptive learning and momentum techniques. Empirically, Adam reportedly shows faster convergence, while Momentum with LR decay converges slower to a better minima [61, 143].

- **Regularization**: Overfitting [41] is a major concern when training complex models. Very expressive models can learn very complex relationships between inputs and outputs of training data, resulting in sampling noise. Some overfitting reducing techniques are L1 or L2 regularization [93, 95], max-norm [122], and dropout [123].

Dropout prevents overfitting by randomly selecting neurons during training to be inactive in each mini-batch (given a sampling probability). This way, the network trains an ensemble network of exponential different networks. During testing, all units are active and the shared weights are downscaled to approximate the average prediction of the different networks.
In [119], the author compares these methods with Convolutional Neural Networks (Section 2.1.3) and shows that dropout is generally the best regularization technique. The authors of recent studies in mammography screening have used dropout [20, 32, 67, 77, 110, 142].

Another noticeable regularization technique is Batch Normalization (BN) [57]. The authors say that BN layers (Equation (2.7)) help mitigating the internal covariant shift (the shifts in the distributions of inputs in each layer that change in training, when previous layer parameters change). However, in spite of empirical benefits, there is no reason consensus to support these results [10, 66, 111].

$$BN(x_k) = \gamma_k \frac{x_k - E[x_k]}{\sqrt{\text{Var}[x_k]}} + \beta_k$$ (2.7)

Because of its regularization effect, other methods such as Dropout are less (if even) required to prevent overfitting. Recent studies incentivise the combination of BN with decreased dropout value (such as 0.3 or 0.4) [2, 150]. A modification called Group Normalization [146] functions in a similar matter, although not suffering from the limitation of having a sufficiently large batch size, which can be a problem in mammography studies. This method is relatively new, although not yet widely tested in mammography research (notable exception found in [12]).

• **Data Dependence**: Deep learning in image recognition has made remarkable progress due to the public access to large scale annotated datasets, along with the revival of Convolutional Neural Networks (Section 2.1.3). In [32], researchers show the impact in performance with larger datasets with higher resolution images on mammography screening. However, obtaining such annotated dataset for medical imaging still remains a challenge [115].

To compensate the lack of data, Transfer Learning (TL) [83, 133] is commonly used for small datasets in mammography screening [14, 55, 77, 110]. TL is a technique where the initialization of network parameters, $\theta$, of model $F$, is based on parameters $\tilde{\theta}$ of a secondary model $\tilde{F}$, that learned task $\tilde{T}$ using an easy access, large dataset $\tilde{D}$. This is motivated from the fact that in lower to mid level layers, parameters $\tilde{\theta}$ can learn abstract feature extraction that can represent latent knowledge to task $\tilde{T}$ and $T$, allowing it to be reused when training model $F$ using the relatively small dataset, $D$.

In [132], authors show that neither shallow nor deep tuning is the optimal choice, concluding that the ideal tuning depth is task and dataset dependant. Hence, experimenting different depths of fine tuning is recommended. Moreover, TL can improve model generalization even after substantial fine-tuning [149], making it an even more attractive method to compensate lack of data.

Other data augmentation options used in research is the artificial enlargement of training data. This technique alleviates the need for data, drastically improving performance under smaller datasets [77]. In mammography screening, this can be achieved by randomcropping [32], rotating and rescaling [84] or flipping for patch level approaches [137]. In full image approaches that do not require this enlargement, flipping is usually applied in prepossessing so that all breasts face the same direction [14, 67].
Data augmentation is specifically usefully when dealing with imbalanced datasets in medical imaging [88]. This way, datasets can overcome their bias towards over-represented classes. In modern DL libraries [19, 100], another steadily way to deal with this imbalance is class weighting [102]. In practice, class weighting reformulates the standard loss function to accentuate the learning of minority classes. An example of inverse frequency class weighting follows below (Equation (2.8)), where $Freq_j$ denotes the frequency of class $j$; $L(y_t, \hat{y}_t)$ and $L^*(y_t, \hat{y}_t)$, the standard and reformulated loss of prediction $\hat{y}_t$ with respect to $y_t$, accordingly.

\[
L^*(y_t, \hat{y}_t) = \frac{L(y_t, \hat{y}_t)}{Freq_{argmax(y_t)}}
\]  

(2.8)

In Section 2.1.2, overall DL techniques were reviewed. It is to note that the benefits discussed are applied to the general purpose, and are still subject to dataset and application variances. Since the purpose of this work is focused on mammography screening, the following Section 2.1.3 explores more pertinent DL models currently available in the field of computer vision for classification.

### 2.1.3 Convolutional Neural Networks

CNNs have been around since 1980s [29]. They were inspired by the concept of receptive fields, based on the studies of brain response to different features such as shapes, colours, or textures [53]. Nowadays, CNNs prove to be the State of the Art in many computer vision tasks [27, 31, 40, 54, 105].

Unlike fully connected layers, convolutional layers were specifically designed to preserve input spatial structure. They make use of the convolution operation to slide shared weights along an image, computing dot products to extract features, often called feature maps. The convolution operation and the output size follow in Equations (2.9) and (2.10), respectively, where $\ast$ denotes the convolution operator. Figure 2.3 shows a simple convolution with stride 1 and no padding.

\[
F_{cv}(x_t, \theta) = h(x_t \ast \theta + b)
\]  

(2.9)

\[
Size_{output} = (Size_{input} - Size_{kernel} + 2 \times Padding) / Stride + 1
\]  

(2.10)

**Figure 2.3.** Example of a Convolution Operation with Stride 1 and no Padding.

In [151], we can visualize that this layered extraction is compatible with the earlier concepts of receptive fields [53]. Although not explicitly programmed, each stacking layer performs feature extraction operations to create increasingly high levels of abstract representations, like detecting blobs and edges in lower layers, and more discriminative artefacts (like eyes or noses) as the network progresses.
Another important operation used in CNNs is the pooling operation. Pooling layers operate over each feature map individually, performing subsampling, thus reducing the feature representations to more manageable sizes and so, the number of trainable parameters in the network. Another useful result of pooling is that it also makes the input invariant to local translation [35]. The most common pooling operation is max-pooling [18, 112], shown in Figure 2.4. This is arguably because since each feature encodes the presence of spatial pattern, it could be more informative to look at the maximal presence of different features. The pooling output size follows Equation (2.10).

In Figure 2.5, a simple CNN for classification is shown for clarity (using convolutions and pooling as feature extractors and fully connected layers for classification). Since 2012, CNNs conducted a breakthrough in modern computer vision with the introduction of AlexNet [71], ZFNet [151], VGG [117], GoogLeNet [130], ResNet [42], and SeNet [48] to the ImageNet annual competitions [22].

It is very computational and memory expensive to train a Deep CNN from scratch. Nowadays, the aforementioned networks are used to pre-train models to alleviate the lack of data. The main trends introduced were:

- **Novel Learning**: AlexNet (Figure 2.6) and ZFNet introduced the use of ReLU, Dropout, and the later removed Local Response Normalization [71]. These novel features at the time greatly improved learning and generalization as discussed in Section 2.1.2. The later mentioned Local Response Normalization was a normalization layer that permitted larger activations to be more pronounced within a neighbourhood. This was achieved by making lateral inhibition of nearby neurons of a large activating neuron. Nowadays, more refined regularization techniques such as Batch Normalization replaced Local Response Normalization [104].
• **Deeper Networks:** Following AlexNet and ZFNet, the work in the field of CNNs shifted to deeper networks using smaller filters. VGG initially proposed this methodology, allowing the increase of the receptive field, while decreasing the number of parameters of the network. This trend continued to rise in further works as deeper networks allow for more non-linearity, and thus, more complex models [42, 50].

• **Residual Learning:** ResNet pushed the depth of models to the extreme (ResNet-152 is eight times deeper than VGG-19). This network reformulated convolutional layers that learn residual functions (Figure 2.7a). They make use of shortcut connections to allow easier optimization of deeper networks. ResNet also widespread the use of bottleneck convolutions (Figure 2.7b). These consist in 1x1 convolutions to reduce and increase the filter dimensions, before and after standard convolutions, effectively saving time complexity while allowing filter depth to be maintained for residual learning [141].

Figure 2.7. Residual Learning, as in [42].

However, the depth of these models can lead to vanishing gradient, and thus, slow learning. To address this, Stochastic Depth [51] was introduced. Similarly to dropout, this method inhibits entire blocks (instead of neurons) at training time, leaving identity mapping as means to bypass them. This leads to substantial less training time and improves experimental test error. Moreover, ResNets were further improved by pre-activation normalization [43], which allowed training of extremely deep ResNets (over 1000 layers).

Interesting debates on wider instead of deeper residual learning, such as ResNext [147] and Wide Residual Networks [150], tried to tackle the problem of diminishing feature reuse [124]. This is an inherent problem to identity mapping, where the gradient flow is not forced to go through residual blocks, and hence, producing lower utility blocks, that train slowly. Although these models show improvements in natural images (twice as fast learning), their application in mammography is yet to be ample (one exception can be found in [84]).
**Inception Models:** GoogLeNet (later called Inception V1) introduced the Inception Module with the main goal of achieving high performance with lower computational cost. The Inception Module incorporates split branching within its blocks of different sized convolutional filters. The branching process was inspired by the Network in Network architecture [80], while the premise behind different sized filters is that salient parts (image feature context) can largely vary in size.

The Inception V1 [130] also introduced a new training methodology of auxiliary learners to speed up convergence, where mid-level Inception blocks were connected to classification heads that contributed to the overall training loss.

In subsequent work, the Inception Module was refined as follows:

- **Inception V2** [131] incorporated bottleneck convolutions before standard convolutions, and used smart factorization of convolutions in shape 1xn and nx1 to lower computational costs;
- **Inception V3** [131], presented in the same paper, rethought the auxiliary learners as regularizers, when adding batch normalization, and also reshaped convolutions to factorized 7x1 and 1x7 filters (which reduce computation but still maintain the receptive field of one 7x7 filter);
- **Inception V4** [129] standardized Inception Modules, restructured the stem and introduced Reduction blocks. While Inception blocks preserved side, Reduction blocks used valid convolutions, thus reducing feature size;
- **Inception-ResNet V1** and V2 [129] also presented in the same paper as Inception V4, explored residual learning capabilities of Inception Modules. This hybrid module also comprised Inception and Reduction blocks, where Inception blocks used identity mapping, whereas Reduction blocks used max-pooling. In order to resist instabilities in residual training, residual activations where scaled by a value of 0.1 to 0.3.

In order to preserve illustration space, Figure 2.8 depicts only Inception-C from Inception V4 and Inception-B from Inception-ResNet V1.

Through the experiments in [129], Inception V4 and Inception-ResNet models outperformed all earlier Inception versions. Residual versions differed in depth (and hence computation), at the cost of performance (less depth, worse performance). Moreover, they proved that residual connections
have equivalent performance (assuming equivalent depth), though accelerating convergence, confirming the trend of faster convergence through residual learning.

- **Densely Connected Convolutions**: Inspired by the trend of shortcut connections from previous to subsequent layers, found in FractalNet [72], Highway Networks [124], and ResNet [43], the DenseNet [50] took skip connection to the extreme. Authors investigated the propagation of features through channel-wise concatenation (Figure 2.9), instead of the usual element-wise addition.

![Figure 2.9. DenseNet, as in [50].](image)

Instead of exploiting representational power from deeper or wider architectures, DenseNet focused on feature reuse, thus building highly dense connected models that are easier to train and highly parameter efficient. In each Dense block, the feature maps of each layer are concatenated to subsequent layers inside the block (the number of channels added for each layer is given by a growth rate, \(k\)), which also introduce a regularizing effect. The paper also polished the baseline work with two distinct features:

- Bottleneck convolutions like the ones Inception Modules before convolutions (1x1 filters to reduce feature map depth). Networks that use this bottleneck are referred to as DenseNet-B.

- Feature map compression within transition layers (layers between dense blocks). Authors introduced a scalar hyperparameter of compression, \(\psi\), which scales down the feature map depth in transitional layers \((0 < \psi < 1)\). If a transitional layer produced \(m\) feature maps, then the corresponding output would be \(\psi \times m\) deep. In the original paper, \(\psi = 0.5\). Networks using this compression are referred to as DenseNet-C, while networks using both bottleneck convolutions and compression are called DenseNet-BC.

Through experiments in [50], authors claimed the stronger gradient flow because of the collective knowledge propagated by these new skip connections through implicit deep supervision (individual layers receive additional supervision from the loss function through their skip connections). They also experimented better performance than ResNets with substantial less parameters having more diverse features that maintain all complexity of levels through the network representations.

It is also worth mentioning that other novel model techniques such as Feature Map Exploitation (eg. Se Net [48] and CMPE-SE [49]) and Channel Boosting (eg. Channel Boosted CNN [62]) are starting to appear in natural imaging classification. While these models have good performances in their domain, their lack of presence in mammography screening makes them less appealing to this work, when compared to Residual, Inception, and Dense models, which have stronger prevalence in classification mammography studies.
In Section 2.1.3, CNN building blocks were discussed on proper use of DL for image contexts. As of now, the reader has insight to understand how CNNs are used to automatically extract features in classification problems. The following Section 2.1.4 will discuss how this feature extraction can be leveraged to identify possible breast lesions in the context of mammography screening.

### 2.1.4 Semantic Image Segmentation

In the context of mammography screening, failure in performance can be critical [38, 136]. Even though some insights have uncovered the “black box” properties of deep NNs [151], it would be useful in CAD systems to visualize the regions of mammographies that make these models infer their decisions [20, 107]. In addition, the use of these detected lesions can be further applied to aid the training of such networks [14, 145]. This work is one such case, where the detection of breast lesion masks can be leveraged as context information relevant for exam classification. This problem can be seen as pixel-wise classification, and is formally defined as Semantic Image Segmentation (SIS) [16].

In typical CNNs, feature maps are downsized through the network. In SIS, the image size must be preserved to obtain pixel-level classifications for the input image. In practice, network design of SIS models comprises both downsampling and upsampling operations to reduce computation costs [26]. Deconvolution is the most common learnable upsampling operation [94]. In literature, it is often called transposed or fractionally strided convolution. Below follow the equations for the deconvolution operation and output size (Equations (2.11) and (2.12)), as well as illustration (Figure 2.10, where / \(* \) denotes the deconvolution operator).

$$F_{dcv}(x_t, \theta) = h(x_t / \theta \star + b) \quad (2.11)$$

$$Size_{output} = Stride \times (Size_{input} - 1) + Filter \quad (2.12)$$

![Figure 2.10. Example of a Deconvolution Operation with Stride 1.](image)

As of now, SIS is a hot topic of discovery in DL. Below, the main trends of the subject are reviewed:

#### • Encoder Decoder Architectures:

Kick-starting deep learning SIS in 2015, Fully Convolutional Networks (FCNs) [82] replaced the final fully connected layers in traditional CNN classification by convolutional ones. With this switch, the input dimensions could be restored in deeper layers, permitting segmentation. The paper also introduced the use of shortcut connections from previous layers to the final deconvolution layer. This way, the network combines coarse, high layer information with fine, low layer information resulting in slightly sharper segmentations (Figure 2.11).
Figure 2.11. FCN architecture with skips, as in [82].

- **Up-Sampling Refinement:** Motivated by the coarse results of FCN segmentations (even when post-processing using Conditional Random Fields [15]), SegNet [4] aimed for more accurate boundary location mappings. This was mainly credited to the novel use of "tied unpooling". During the encoder part of the network, pooling switch variables memorize the location of max-activating units that later restore the location of activations in max-unpooling, during decoding (Figure 2.12). Although not usually present in medical imaging, an adaptation of SegNet, NablaNet [87], is a notable exception.

Figure 2.12. SegNet architecture, as in [4].

Also FCN based, U-Net [108] is a CNN specifically designed for Biomedical Image Segmentation. Motivated by the lack of training data, it introduced a network that did not required an abundant annotated dataset to train. Instead, the work relied on a network that could efficiently use data augmentation. The architecture still consists of symmetric contracting and expanding paths, to fuse high-level feature maps that are rich in semantic information with low-level feature maps that are rich in location information, leading to better segmentation results [47]. The shortcut connections are made with simple feature map concatenations (Figure 2.13).

Figure 2.13. U-Net architecture, as in [108].

It is to note, however, that as the backbone networks described in Section 2.1.3 were refined, novel implementations of U-Net architectures, such as Residual [78] or Dense [39] U-Nets started to emerge in medical imaging.
Attention Models: Current dominant methods for sequence transduction (e.g., language translation) are based on encoder-decoder architectures. The best performing ones use attention mechanisms to establish the encoder-decoder mappings [138]. This attention mechanisms also propelled CNNs in computer vision for natural images [139, 144]. In mammography screening, two notable novel implementations of attention for lesion segmentation are noted:

- **Attention Dense-U-Net [79]:** This network is inspired by the use of Attention Gates [96] and Dense Blocks [50] to enrich the decoding and encoding stage of U-Nets, respectively. The attention mechanism in CNNs forces the networks to give more focus to the selected prospective salient parts of the inputs or features [3]. This is achieved by obtaining gating coefficients to use element-wise multiplication that suppresses irrelevant feature activations (Figure 2.14). This element-wise multiplication is pixel-wise, preserving spatial structure.

![Overall Architecture](image1)
![Attention Gate](image2)

**Figure 2.14.** Attention Dense U-Net, as in [79].

- **AUNet [127]:** In this network, attention is used to carefully select rich-informative channels using channel-wise attention. The main differences is that AUNet is residual based and uses dense upsampling in the decoding process [52]. The rest of the network is equivalent where Upsampling Blocks represent the Attention Gates, although more closely resembling the Squeeze-and-Excitation attention that derives channel-wise attention [48] (Figure 2.15). Using Global Average Pooling [80], the channel-wise descriptor is obtained (comprising fully connected layers), which are then used as scalars in channel-wise multiplication for the attention mechanism.

![Overall Architecture](image3)
![AU Block](image4)

**Figure 2.15.** AUNet, as in [127].

SIS is still a hot topic in computer vision research. In Section 2.1.4, some current competing models for image segmentation were disclosed, giving emphasis to models used in medical imaging. Given the
specific context of mammography imaging in this work, further techniques explored in literature will be discussed in Section 2.2, to give insight on current methodologies used in the field.

2.2 Mammography Screening

As mentioned in Section 2.1, modern neural network research is providing major breakthroughs in its applications. However, it is to note that the field is guided by mathematical and engineering disciplines, functioning as an approximation that is designed to achieve statistical generalization [35]. In this sense, some domain contextualization can be enforced to enrich model application, which is the focus of the following Sections: Section 2.2.1 proceeds a brief contextualization on the historic progress of CAD tools in medical imaging. Section 2.2.2 motivates the need to reformulate the topography of DL architectures to match the standard data structure of mammography exams. Section 2.2.3 explores the DL downfall of scarce annotated data in medical imaging and modern workarounds.

2.2.1 Traditional Methods vs. Deep Learning

Prior to DL, the traditional methods applied in CAD followed the conventional ML workflow illustrated in Figure 2.16. The main difference between DL and traditional approaches is the feature extraction process. In the later, feature selection can become a tedious task as the discriminative power of features used in CAD system varies. Moreover traditional methods deal with problems such as varying size of masses, architectural distortion, detection of bilateral asymmetry and threshold definition, making this alternative unappealing. While it is not the focus of this work to discuss traditional methods and shortcomings in detail, the work in [30] provides an excellent overview of the current State of the Art on the subject.

![Figure 2.16. Mammography Processing with Traditional ML Methods, as in [30].](image)

One of the attractive aspects of DL methods, in comparison to the traditional ones, is that they bypass the feature extraction and selection challenges due to their inherent automatic feature extraction nature,
which produces the optimal high-level features for the task. This is specially useful in the context of mammography screening, where data is structured in a non conventional topology, further evidenced in the following Section 2.2.2.

### 2.2.2 Multi-View Mammography

In medical practice, it is standard to use both CC and MLO views of a mammography to detect abnormalities. This holds true for recent mammography studies, which reformulate previous Single-View by Multi-View approaches with more robust representations that increase model performance [63, 125]. Although troubling for traditional ML models, the inherent high representation powers of DL models facilitate the junction of both views in upper layers [14]. In [145], authors argue that view-wise joining provided best validation results, also observing that the CC view is more predictive than the MLO view. Moreover, they further enrich the model input by adding previously segmented lesion heat-maps (similar to the work in [14]), confirming that this extra data inclusion increased performance.

Other studies also explore the extent of auxiliary meta-data (such as patient age or breast density), and reformulate exams in a multi-modal fashion [68]. However, due to privacy reasons, such datasets are challenging to find. Additionally, developing DL models with limited annotated data is still an active area of research. This is specially true in medical imaging, where getting a large annotated dataset is costly [132]. In the following Section 2.2.3, dataset impact on learning and possible workarounds are further discussed.

### 2.2.3 Dataset Impact

The success of DL models is highly data driven. As mentioned in Section 2.1.3, the introduction of largely annotated datasets such as ImageNet were a milestone that propelled DL research in computer vision. In [32], authors advocate the impact of dataset size and image resolution on the learning capabilities of CNNs applied to mammographies, concluding that performance is directly proportional to both. Moreover, they successfully show how to train high resolution models (2600x2000 input images) in an end-to-end fashion, using aggressive convolution/pooling, and clever feature map concatenation [80].

Despite the success of this high resolution trend [145], the use of such high resolution images is very data and computationally expensive. When such dimensionality is unfeasible, it is common practice to downscale the images to \(\sim 250 \times 250\) pixels, the standard in natural image datasets [42].

To overcome data driven issues, an overview on current techniques that mitigate these problems in mammography research follows bellow:

- **Patch Level Approaches:** To avoid training with small datasets, earlier works developed models that processed minimal overlapping patches of the initial image (for either pre-training full image models [114], or multi-scale approaches [84]). This way, relatively small datasets can be extended (each image provides multiple samples). However, careful extraction of both negative and positive patches is recommended to avoid divergent learning [152].
• **Scarce Annotation Workarounds:** Mammography imaging datasets usually have image-level labels such as Breast Imaging Report and Data System Score (BIRADS), while datasets with lesion annotations are scarce. In order to tackle this issue, Weakly Supervised Learning [99] approaches have been explored [6, 56]. They make use of global annotations to train a typical classification networks based on image-level labels and then inspect top layer activations in order to refine segmentations. In [8], a weakly supervised technique is further enhanced by using an ordered loss function, to enforce importance of certain errors over others.

• **Small Data Size Approaches:** In recent studies, undisputed models like Conditional Generative Adversarial Networks [58] have proved successful in mammography research [118]. The network makes use of an FCN based generator to produce segmentations to be discriminated in the usual min-max problem of generative adversarial networks. The key advantage of these models is that they can work well even with a relatively small dataset.

• **Domain Transfer:** Due to the low size of common public datasets in mammography [45, 91], direct comparison between methodologies may not be correct. Some models can be more sensitive to certain dataset specific variances, yielding untrue performance variations. One interesting methodology that tries to provide a common ground is the adaptive deep segmentation system in [134]. Using Coupled Generative Adversarial Networks [81], the network can transfer the samples from different mammography datasets, which then can be examined by a general network.

• **Transfer Learning:** As discussed in Section 2.1.2, Transfer Learning has been successfully applied in DL mammography research to facilitate training under smaller datasets [14, 106]. This work will take advantage of this technique, motivated by the aforementioned success in the applications of related work in the field.

In this chapter, the State of the Art was reviewed in the field of DL. More specifically, its applications to novel CAD tools were explored. The reader is now familiarized with the building blocks of this work, and can now understand the following Chapter 3, where the tools discussed in Chapter 2 will be used to propose models that can accomplish the objectives described in Section 1.1.
Chapter 3

Proposal

Leveraging the knowledge gathered in Chapter 2, this Chapter details the approach that this work proposes to accomplish the goals defined in Section 1.1. In sequence, Section 3.1 formalizes each task with respect to the objectives, giving a brief introduction to both. Sections 3.2 and 3.3 detail information about the training workflow and model evaluation conducted in the classification and segmentation tasks, respectively. Finally, Section 3.4 presents the validation methods used to ensure robust results.

3.1 Task Description

Given the objectives of this work in Section 1.1, the proposed Deep Neural Networks for each task are described as follows:

1. (Classification) - A DenseNet architecture is proposed to solve the classification objective. The task is to develop the classification of mammography image inputs along with lesion mass and MC masks into the correct class (either Normal, Benign, or Malignant). These masks are based on the segmentation results and dataset ground truth lesions for the fully and semi-automated trials, respectively;

2. (Segmentation) - An Attention Based Dense U-Net architecture is proposed to solve the segmentation objective. The task is to develop the segmentation of mammography image inputs into binary lesion masks (masses and MCs), to be later used on the automated classification trials.

In the two subsequent Sections 3.2 and 3.3, further details are provided in regards to the methodology used in the corresponding experiments.

3.2 Classification

This section describes the proposed solution relative to the Classification task mentioned in Section 3.1. Evidenced by success of DenseNets in the natural image domain [50], this work proposes the exploration of dense core architectures for classification of mammography images in the INbreast dataset [91].
In order to track the reaction of dense architectures to the typical breast imaging methodologies, a staged progression of the models was tested 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 [145], 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 [14], 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 TL improvements in [14], the models 4-A*, 4-B*, 5*, 7* verified the boost of pre-training on the Imagenet dataset [22].
- **(Full Automation):** In addition to evaluating the classification results using the dataset ground truth lesion masks, the most promising classification architectures, were later evaluated on fully-automated trials (as in [14]), by training on automatic segmented masses and MCs lesions from the most promising segmentation models developed in this work.

Figure 3.1 depicts the training workflow of each classification model, for clarity:

![Figure 3.1. Training Workflow of Classification Models.](image)

Being a multi-class classification problem, training used cross-entropy [74] as the loss function. For completeness, the formula follows in Equation (3.1), where \( y_{t,c} \) and \( \hat{y}_{t,c} \) denote true and predicted probability of sample \( t \) belonging to class \( c \), respectively, while \( C \) denotes the number of classes:

\[
L_{CE}(y_t, \hat{y}_t) = - \sum_{c=1}^{C} y_{t,c} \times \log(\hat{y}_{t,c})
\]  

(3.1)

After training, the proposed classification models were quantitatively evaluated between each other according to the metrics of Area Under ROC Curve (AUC), Dice Coefficient (Equation (3.2)), Precision
(Equation (3.3)) and Sensitivity (Equation (3.4)), of which formulas follow below. 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).

\[
\text{DiceCoefficient} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{3.2}
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{3.3}
\]

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3.4}
\]

In addition to this internal evaluation, the best performing classification models will be subject to comparative evaluation regarding the work in [14], according to comparable 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 below in Equation (3.5)).

\[
\text{Specificity} = \frac{TN}{TN + FP} \tag{3.5}
\]

The main motivation behind the direct comparison is due to the relatively small size of the dataset used, which heavily impacts the learning of DL models. In the next Section 3.3, the workflow of the segmentation task is discussed.

### 3.3 Segmentation

This section describes the proposed solution relative to the Segmentation task mentioned in Section 3.1. Resuming the exploration of dense models and its applications on the INbreast dataset, this work proposes to investigate the results of using a dense core to segment the dataset masses and MCs.

Inspired by the breast mass lesion segmentation work in [79, 127], this work proposed the following model 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 [79] has no published results on the INbreast Dataset, while the work in [127] 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 segmented lesions in [14] that made use of cascaded architectures for both mass and MC lesions [23, 85].
• (Pre-Training): Inspired by the TL improvements in [127], 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 [127] was done on the CBIS-DDSM dataset [75].

Figure 3.2 depicts the training workflow of each segmentation model, for clarity:

![Figure 3.2. Training Workflow of Segmentation Models.](image)

Being a binary segmentation problem, training used a combined loss function (Equation (3.8)) that aggregated both Dice (Equation (3.6)) and Cross Entropy Losses (Equation (3.7)) for each pixel. This was inspired by the work in [127, 153], which states that solely applying the Dice Loss results in unstable optimization. Additionally, 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 output of pixels:

$$L_{\text{Dice}}(y_t, \hat{y}_t) = 1 - \frac{2 \sum_{c=1}^{C} \sum_{i=1}^{N} (y_{t,i,c} \times \hat{y}_{t,i,c}) + 1}{\sum_{c=1}^{C} \sum_{i=1}^{N} (y_{t,i,c} + \hat{y}_{t,i,c}) + 1}$$  \hspace{1cm} (3.6)

$$L_{\text{BCE}}(y_t, \hat{y}_t) = -\frac{1}{N} \sum_{c=1}^{C} \sum_{i=1}^{N} (y_{t,i,c} \log(\hat{y}_{t,i,c}))$$ \hspace{1cm} (3.7)

$$L_{\text{Combined}}(y_t, \hat{y}_t) = L_{\text{Dice}}(y_t, \hat{y}_t) + L_{\text{BCE}}(y_t, \hat{y}_t)$$  \hspace{1cm} (3.8)

After training, the proposed segmentation models were quantitatively evaluated between each other using the metrics of Dice Coefficient, Precision, Sensitivity, and $\Delta A$ (inspired by [127] and having its formula below in Equation (3.9), where $A_{\text{GT}}$ and $A_{\text{Pred}}$ represent the ground truth and predicted lesion area, respectively). The main motivation for the chosen metrics was the focus on the severely under-represented positive class (Mass and MC lesions).

$$\Delta A = \frac{|A_{\text{GT}} - A_{\text{Pred}}|}{A_{\text{GT}}} = \frac{|(TP + FN) - (TP + FP)|}{TP + FN}$$  \hspace{1cm} (3.9)

In addition to this internal evaluation, the mass segmentation models will also be subject to quan-
titative evaluation in relation to the work in [79, 127], according to the comparable metrics of Dice Co-efficient, Sensitivity, Specificity, and finally, $\Delta A$. 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 IN-breast samples that contained mass lesions. The motivation behind the comparison is again due to the fact that the results can be comparable to the experiment in [127]. The next Section 3.4 provides the methodologies used to ensure valid results.

### 3.4 Validation

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 [35] (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.

In this Chapter, the methodology of this work was presented. More specifically, the proposed solutions for the objectives of this work were detailed. The reader now understands which are the tools being used, and how they are put forth to ensure valid results. The next Chapter 4 shows the implementation details regarding the specifics of model architectures and training hyper-parameters used in the experiments.
Chapter 4

Implementation

According to the methodology proposed in Chapter 3, this Chapter explores the implementation details regarding each discussed task. In order, Section 4.1 features the tools and dataset used. Sections 4.2 and 4.3 each highlight the implementation process and training regarding the Classification and Segmentation tasks, respectively. The models were refined using available techniques mentioned in Chapter 2, with respect to learning trial results, while not giving focus to any model in particular.

4.1 Dataset and Tools

This work makes use of the publicly available dataset INbreast [91], having 410 annotated scans along with precise segmentation maps of mass and MC lesions. Each scan corresponds to a single breast scan of either a CC or MLO view. This dataset has an imbalanced distribution of 6 classes, corresponding to BIRADS [7]. Following the work in [14], these classes were reformulated into 3 classes: Normal, when BIRADS = 1; Benign, when BIRADS \( \in \{2,3\} \); and finally, Malignant, when BIRADS \( \in \{4,5,6\} \).

Both original (Figure 4.1a) and reformulated (Figure 4.1b) distributions follow in Figure 4.1:

![Figure 4.1. Dataset Class Distribution.](image)

The pre-processing in this work followed the same methodology as in [14], where mammograms were enhanced via local contrast normalization, followed by Otsu’s segmentation [98] 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.

All models described in this Chapter were implemented in Python, assisted with publicly available libraries such as Keras [19] (using Tensorflow [1] as Back-End). Pre-training was also implemented using Keras, which provides public DL models with pre-trained weights on the Imagenet dataset [22].

In order to meet the demands of computational power, training was executed under the GPUs publicly available using the Google Colab platform [37].

In the two subsequent Sections 4.2 and 4.3, further details are provided in regards to the implementation and training details used in the classification and segmentation experiments, accordingly.

4.2 Classification Model

In this Section, the implementation details regarding the classification task are discussed. As mentioned in Section 3.2, these models made use of a DenseNet to classify mammography image inputs. Represented below, Figure 4.2 gives an overview of the dense architectures of both single and dual branch topologies. The model core is faithful to the original DenseNet-BC publication in [50], keeping the initial aggressive downsize of a 7x7 strided convolution followed by 3x3 max-pooling, as well as bottleneck layers inside dense blocks and compression on transition layers. On the other hand, the overall model architecture was inspired by the work in [14]:

![Figure 4.2. Classification Model Architectures.](image)

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 [33]. 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 downscale.
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 techniques such as random rotation and flipping were tested. Furthermore, class weighting was also experimented on. The latter delivered better results, even when using data augmentation generating balanced batches. Exponential weight decay also proved a better LR scheduler than a constant LR, and therefore was used during classification trials. For the sake of completeness the formula for Exponential LR Decay (Equation (4.1)) is presented below, where $\eta_p$ denotes the LR at training epoch $p$ and $\gamma$ the decay factor, respectively:

$$
\eta_p = \eta_0 \times \gamma^p \tag{4.1}
$$

Parameters and its scalars were chosen based on training stability throughout models, so as to not give an unfair advantage to any in particular. Additionally, the compression factor and growth rate of DenseNets were kept at recommended values from the original work [50]. In the following Table 4.1, a general overview of hyper-parameters is aggregated for summarization:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Models With Pre-Training</th>
<th>Models Without Pre-Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starting LR</td>
<td>$\eta_0 = 10^{-3}$</td>
<td></td>
</tr>
<tr>
<td>LR Decay</td>
<td>Exponential ($\gamma = 0.95$)</td>
<td></td>
</tr>
<tr>
<td>Optimization Strategy</td>
<td>Adam ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-7}$)</td>
<td></td>
</tr>
<tr>
<td>Epochs</td>
<td>100</td>
<td>150</td>
</tr>
<tr>
<td>Batch Size</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>Dropout Probability</td>
<td>$p = 0.2$</td>
<td></td>
</tr>
<tr>
<td>L2 Regularization</td>
<td>$\lambda = 10^{-5}$</td>
<td>$\lambda = 10^{-4}$</td>
</tr>
<tr>
<td>DenseNet Compression</td>
<td>$\psi = 0.5$</td>
<td></td>
</tr>
<tr>
<td>DenseNet Growth Rate</td>
<td>$k = 32$</td>
<td></td>
</tr>
</tbody>
</table>

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

In the next Section 4.3, implementation and training details of the segmentation task are discussed instead.
4.3 Segmentation Model

In this Section, the implementation details regarding the segmentation task are discussed. As mentioned in Section 3.3, these models made use of a U-Net like network employing attention and dense core to segment mammography image inputs. Represented below, Figure 4.3 gives a general overview of the dense architecture used for both mass and MC detection. The dense core is essentially the same as the previous Section 4.2, using a DenseNet-BC core (DenseNet-121 and DenseNet-37 for models with and without pre-training, respectively):

![Figure 4.3. Segmentation Model Architecture.](image)

The overall shape is inspired by the work in [79, 127], while staying true to the dense core of [50]. This is controversial due to the fact that the architectures in [79, 127] 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 [50]. 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.

In Figure 4.4, the decoder block of the Attention Dense-U-Net (Figure 4.4a) and Dense AUNet (Figure 4.4b) follow below for completeness, where the mentioned Attention Gate, Channel-Wise Attention and Dense Upsampling are faithful to the corresponding related works:

![Figure 4.4. Decoder Block Variations.](image)

Similarly to Section 3.2, 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 [33].

To improve the generalization ability of models, weight loss and data augmentation through random
rotation was used in the segmentation trials. Again, L2 weight regularization lead to stable 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 $\in [-20^\circ, 20^\circ]$ so as not to excessively distort the input.

Parameter choosing was again based on unbiased training stability. In particular, the LR was lower, which was matched by a higher LR decay that smoothly followed the otherwise step decay proposed in [127]. Moreover, the AUNet reduction factor in the channel wise attention mechanism was set to the recommended value. Furthermore, batch size was substantially lower, in comparison to classification trials, due to memory constraints. The summarized overview of hyper-parameters follows in Table 4.2:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Dense Attention-U-Net</th>
<th>Dense AUNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starting LR</td>
<td>$\eta_0 = 10^{-4}$</td>
<td></td>
</tr>
<tr>
<td>LR Decay</td>
<td>Exponential ($\gamma = 0.96$)</td>
<td></td>
</tr>
<tr>
<td>Optimization Strategy</td>
<td>Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-7}$)</td>
<td></td>
</tr>
<tr>
<td>Epochs</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Batch Size</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>L2 Regularization</td>
<td>$\lambda = 10^{-5}$</td>
<td></td>
</tr>
<tr>
<td>DenseNet Compression</td>
<td>$\psi = 0.5$</td>
<td></td>
</tr>
<tr>
<td>DenseNet Growth Rate</td>
<td>$k = 32$</td>
<td></td>
</tr>
<tr>
<td>Dense AUNet Reduction</td>
<td>$r = 16$</td>
<td></td>
</tr>
</tbody>
</table>

In this Chapter, the implementation details of this work was presented. More specifically, the motivation and structure of architectures for the proposed solutions of this work are detailed. The reader now understands how training was put forth to reach the results of this work. The next Chapter 5 shows the experimental evaluation of each task, following with discussion to give reason to possible expected and/or unexpected outcomes.
Chapter 5

Results

In this Chapter, the evaluation of the proposed solutions detailed throughout Chapters 3 and 4 is presented. Quantitative results are aggregated visually, having comparative evaluation relative to the State of the Art tabulated for detail. After this quantitative evaluation follows the display of a few sample classifications and segmentations, finally concluding with a brief summary of remarks about the results. In order, Sections 5.1 and 5.2 present the results for the classification and segmentation tasks, respectively.

5.1 Classification Results

This Section presents the classification results of the experiments described in Sections 3.2 and 4.2. In order to visualize the impact on performance by each breast imaging methodology, the quantitative results for each model follow in Figure 5.1 (hereinafter, FA-5* and FA-7* denote the fully-automated trials of models 5* and 7*, respectively):

![Figure 5.1. Evaluation Between Proposed Classification Models.](image)

Proceeding this evaluation, the comparative results of the most competitive models are presented in Table 5.1 (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):

<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>[14]</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><strong>0.92±0.08</strong></td>
</tr>
<tr>
<td>5*</td>
<td>No</td>
<td>0.90±0.02</td>
<td>0.86±0.04</td>
<td><strong>0.88±0.03</strong></td>
<td><strong>0.73±0.06</strong></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>[14]</td>
<td>Yes</td>
<td>0.78±0.09</td>
<td>0.86±0.09</td>
<td>0.72±0.10</td>
<td><strong>0.66±0.14</strong></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>0.78±0.05</td>
</tr>
<tr>
<td>FA-7*</td>
<td>Yes</td>
<td><strong>0.86±0.07</strong></td>
<td>0.85±0.03</td>
<td><strong>0.79±0.06</strong></td>
<td>0.61±0.04</td>
<td><strong>0.80±0.03</strong></td>
</tr>
</tbody>
</table>

Following this evaluation, some classifications of model FA-7* are depicted in Figure 5.2. 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:

To finish the experimental classification results in this work, an itemized summary of preliminary conclusions follows (with respect to the highlighted method):

• **(Mass and MC Masks):** The inclusion of lesion masks substantially increased classification 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 [145]. This statement is a mere remark, as the dataset size used in this work is substantially lower by comparison. 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 exams of rows and columns (2, a) and (3, c) of Figure 5.2). In contrast, the first two rows in column d present over-sensitive classifications of this phenomenon.

In this Section, the results of the classification models were displayed. In the next Section 5.2, the mass and MC segmentation results are depicted and discussed.

5.2 Segmentation Results

This Section presents the lesion detection results of the experiments described in Sections 3.3 and 4.3. In order visualize the impact on performance by each mass segmentation architecture, the quantitative results of each model follow in Figure 5.3:
Proceeding this evaluation, the comparison of mass segmentation results regarding the State of the Art is presented below in Table 5.2, 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:

<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>ΔA</th>
</tr>
</thead>
<tbody>
<tr>
<td>[127]</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>[127]</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>[127]</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>[79]</td>
<td>DDSM [44]</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 5.2: Comparative Mass Segmentation Model Evaluation, regarding [79, 127]

In order to visualize the impact on performance by each MC segmentation architecture, the quantitative results of the corresponding models follow in Figure 5.4:

Figure 5.4. Evaluation Between Proposed MC Segmentation Models.

Proceeding this evaluation, the results of each MC segmentation model are tabulated below in Table 5.3 for consistency (using the same column description as Table 5.2):
Table 5.3: 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>$\Delta A$</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>

In addition to the quantitative evaluation, a sample of lesion segmentations by architecture follows. It is to note that, while the representation overlaps mass and MC lesions and joins both views for each exam, the segmentations were separately produced by the respective Single-View model for each lesion type. Depicted in Figure 5.5, each row represents the same exam being segmented by the respective column architecture. As in Figure 5.2, the outlines in red and orange depict ground truth masses and MCs, while blue and cyan outline the segmentation prediction of masses and MCs:

(a) Models I-A and I-B  (b) Models I-A* and I-B*  (c) Models II-A and II-B  (d) Models II-A* and II-B*

Figure 5.5. Lesion Segmentation Comparison by Model.

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

- (Transfer Learning): Again, TL actively helped model initialization, allowing increased model depth, resulting in better overall performance. However, while the TL improvements using ImageNet [22] were substantial, they showed inferior performance than [127], which did pre-training using CBIS-DDSM [75]. 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 [127] sharing the tasks of both pre-trained and fine-tuned models (mass segmentation).
• (Attention Module): Regarding the architectures, the grid attention mechanism found in 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 outstandingly better and more robust (higher mean and lower variance). Interestingly, the sensitivity levels of both architectures was similar. This indicates that the Dense-AUNet was over-sensitive by comparison. Additionally, this seems to explain the performance deviation of the Dense-AUNet for the experiment in this work and the one in [127], whose training was done exclusively on the exams containing mass lesions.

• (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 5.1 seem to indicate that both mass and MC segmentations proved sufficient to improve the classification of mammograms in a real life scenario, when compared to the work in [14].

In this Section, the results of the segmentation models were displayed. The next Chapter 6 concludes this document with final remarks about the experiments.
Chapter 6

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 document such tools were developed for the task of both classification and detection of lesions in mammography exams. 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 fully-automated classification on the INbreast dataset [91]. In the following Sections 6.1 and 6.2, the main contributions and proposed future work are presented.

6.1 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 [14] or require abundant private datasets to train [145]. 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 [91] to provide a fair assessment of architectural performance.

A downside of the INbreast dataset is its inferior size, which, inspired by the work in [14], was compensated by the use of TL on the Imagenet dataset [22]. The data dependence downfall 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 [127] and [79] 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 [50], employing an initial aggressive convolution and pooling to reduce the feature size.

Regarding the proposed development of the semi-automated model in [14], 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 that the approach in [14] which made use of dedicated cascaded pipelines to automatically extract masses [23] and MCs [85].

6.2 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 practice 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 [79, 145]. 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 [79].

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 [75], while also testing the performance according to the methodology in [127], 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 [11], which feature the inclusion of additional exam types such as Ultra Sound or Magnetic Resonance Imaging for enhanced assessment of breast cancer.
Bibliography


