Multi-modal HyperDense-Net Classification For Breast Diagnosis

Madalena Pedreira*
madalena.pedreira@tecnico.ulisboa.pt
Instituto Superior Técnico
Lisboa, Portugal

Supervisor: Jacinto Nascimento
Supervisor: Carlos Santiago

ABSTRACT

In this paper, following the importance of early detection in breast cancer to decrease today's alarming mortality rates, traditional and innovative approaches on breast cancer diagnosis will be assessed. In the wake of the emerging research on Computer-Aided Detection (CAD) practices with Deep Learning (DL) approaches in the mammography screening field and given DL model's current successes in this context, an approach on the modern practices to overcome the particular challenges of data insufficiency, information loss and lack of computational power will also be featured. Finally, this work proposes an advanced DL solution through a Multimodal architecture for classification of mammography screenings utilizing an Hyperdense Network, which will process inputs of an unestablished private dataset.

CCS CONCEPTS

• **Deep Learning** → **Convolutional Neural Network**; *Hyperdense Network*; *Breast Cancer*; Classification; Weak-label; • **Dataset** → Curation.

KEYWORDS

Unbalanced dataset, Medical Imaging, Multimoda

1 INTRODUCTION

Current studies showed that Breast Cancer (BC) was the main cause of death in the EU-27 in 2016 [15], registering a 1.9% death rate; among women, BC accounted for 3.7% of all deaths and is the second leading cause of worldwide cancer related deaths with a percentage of 11.6% [5]. Though hereditary and genetic factors, accompanied by the personal and family history weight on the count of BC cases, demographic and social-economic development are the main halters, earlier detection and treatment being the leading reasons these numbers are tendentially decreasing, as this cancer is more responsive to less aggressive treatment in earlier stages [30]. Mammography screening was found to produce substantial reductions in the mortality rate of BC in women participating in these exams [13, 14, 25, 30]. Furthermore, the mammography imaging test was proved to be the only imaging test that reduces BC's mortality [12], bringing light to the importance of the correct interpretation of these exams. There are, nevertheless, properties that influence the effectiveness of the diagnosis, like the density structure of a breast. This factor impacts sensitivity in mammography screening, particularly in image-based exams, being important to cross-factor these with different image modalities. More particularly, current evidence indicates that Ultrasound (US) of the breast is an important adjunct to mammography and clinical examination in the further assessment of breast abnormalities, as it surpasses the impairment of breast density (sensitivity problems) [19, 35].

Nowadays, Computer-Aided Diagnosis (CAD) is part of the professional's routines, assisting with image interpretation, acting as an automated second reader by signaling potentially suspicious sites for radiologists to review. Convolutional Neural Networks (CNNs) arise as a new generation of CAD devices in the wake of recent developments in Deep Learning.

Once comparing breast cancer detection performance of radiologists utilizing these new CAD technological approaches, recent studies [26, 33] suggest that the CAD-DL imbued systems show great potential for the assistance of medical professionals.

This work is going to focus on a single CAD system task: classification. Classification tasks applied to BC determine the severity of a screening. Since it is standard for mammography screenings to take multiple screenings for each breast of a patient (Multi-View) as opposed to the traditional image classification tasks that take a single image (Single-View), some adjustments are to be made. In particular, a Medio-Lateral Oblique (MLO) and Cranio-Caudal (CC) mammography views are to be taken into consideration in the classification task. In addition to these, the proposed model will also make use of the ultrasound image modality, consequently assigning a Multimodal system architecture.

Concurrently, the curation process of a set of breast images of the previously enumerated three modalities will also be made. These endeavors will ultimately produce a usable dataset to feed the Hyperdense Network. Following current efforts to develop weak-label model arquitectures, this work developed a usable dataset with such labels. Taking into consideration the lack of publicly available datasets and, moreover, the lack of weak-label datasets in the breast cancer screening field, this attempt was especially intriguing.

The system will be further explained in the proceeding sections as well as for its evaluation parameters. As for the proposed dataset creation, a description of its pre-arrangement efforts, preprocessing operations and validation will also be clarified.

2 OBJECTIVES

This work has two fundamental objectives: the development of a Deep Learnining classification model to be used in a CAD system to ultimately assist medical practitioners in the diagnosis of BC exams; and the curation of a series of breast examinations to feed the model.

The first objective was tackled with an automated model, under a Multimodal Hyperdense-Net architecture [10], that takes 3 different breast exams as input - two related to the CC and MLO mammography images and one related to the ultrasound modality. The foreseen output classes of the system will be of Normal Tissue, Benign and Malignant, associated with no lesion whatsoever and lower or higher probability of a cancerous tumor respectively.

Concurrently, efforts to curate a set of breast images of the previously enumerated three modalities will also be made. These endeavors will ultimately produce a usable dataset to feed the Hyperdense Network following a weak-label paradigm.

3 THEORETICAL BACKGROUND

3.1 Computer-Aided Diagnosis

Computer-Aided Diagnosis (CAD) systems are an interdisciplinary technology that combines elements of artificial intelligence and computer visualization models to assist doctors in the interpretation of medical images [8, 29]. More specifically, they process radiological and pathological images to act as a medical professional's second read. This could deeply benefit the radiologist's work, as it relies on a short period of time to make a diagnosis decision which usually depends on the analysis of a minimum of two medical opinions[21].

With the introduction of CAD systems in the health scene, recent studies have been trying to make sense if their assistance is relevant [20, 24, 26]. DL approaches have been making significant improvements in CAD systems in radiology areas [28]. Current literature focused on these new approaches, found that DL CAD systems may enhance the sensitivity of mammography. More particularly, these could detect masses, architectural distortions, and asymmetries on the screenings regardless of the densities of the tissues as well as for identifying malignant lesions that are obscured by benign-suspicious lesions [4, 23, 31]. This technology is now correctly marking most (96.1%) asymptomatic breast cancers detected with digital mammographic screening, with acceptable false-positive marks (1.80 per patient) [37]. Another clear advantage was the decrease in reading time, which exhibited an improvement by 29.2% with CAD [4]. This factor is especially important as cancer's diagnosis at an early stage is more likely to be treated successfully [36]. Moreover, as BC diagnosis requires more than a single medical opinion in the screening context, studies believe CAD systems are potentially equal to another medical 'read', that is, CAD may hypothetically take the place of a second doctor's opinion and, with further advances in the area, even as first read system.

3.2 Multi-View and Multimodal Mammography Models

Early detection and treatment of breast cancer is a fundamental task to decrease its mortality. Contemporary literature on the field states that no single modality has high enough sensitivity for a reliable diagnosis [6]. As a result, the current workflow for radiologists involves the analysis from at least two radiologists of several imaging modalities, mammography (with CC and MLO views), Ultrasound (US) and Magnetic Resonance Imaging (MRI). The combination of these modalities can significantly increase diagnostic accuracy, resulting in better patient care and reducing the number of unnecessary biopsies. Concurrently, as Deep Learning architectures reached the medical scene, the same methodologies were sought to develop models capable of detecting and classifying mammography exams and breasts' mass segmentation. However, in traditional image classification tasks only a single image (Single-View) is considered, which is translated into CC and MLO mammography views being separately analysed. These systems produced unsatisfactory results, which led to their depreciation until a more performant architecture was developed. It is nowadays customary to utilize both CC and MLO views of mammography to detect tissue anomalies, therefore relying on a Multi-View approach to produce an output, contrary to the previous Single-View methodology. This Multi-View network configuration presented increased performance results [32]. Unimodal models, however, tend to ignore the holistic context of medical images that can only be provided by multiple outlooks. These needs can be fulfilled with the use of multi-modality. Since the current SoA models provide the representation capabilities to conjugate supplementary modality information, the US or MRI modality are used together with the mammography modality to clarify certain system results.

In order to potentiate the best model performance, different network configurations are possible to merge the various modalities' information - early fusion, late fusion are amongst them. Current methods involve multi-layer fusion [27] or concatenation [22]. In this work a complex fusion strategy is used, in which each imaging modality has a path, and dense connections occur not only between the pairs of layers within the same path, but also between those across different paths, not requiring neither early nor late fusion practices [10]. As this method produced good results in the BC multimodal-diagnosis models, the same approach will be used in this work.

3.3 Hyperdense Network

Research over the last years has brought to light the performant behavior of densely-connected networks in the medical image scene. Inspired by the trend, hyper-dense architectures started to arise in the hopes of grasping multimodal image problems, extending the dense connectivity philosophy to the already proved to be competent tactic of considering various modalities. This architecture extends the establishment of deep connections not only between layers within the same modality path, but also between layers throughout different modality paths. Considering the formerly in-

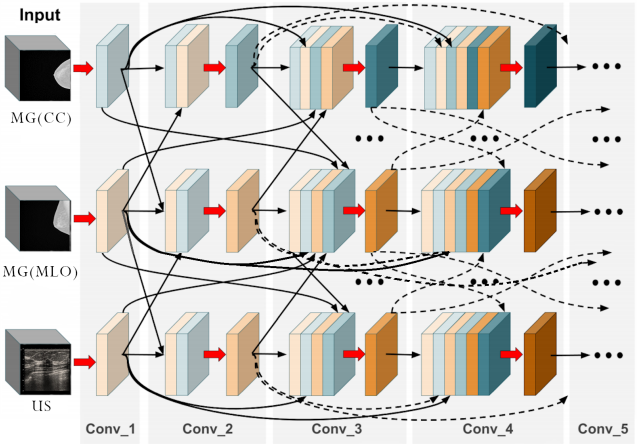

Figure 1: A section of the proposed HyperDenseNet in this work's case of three image modalities.

troduced function described in Equation (??) of the DenseNet, the HyperDense Net introduces a more effective feature representation, that translates the complex relations within and in-between all abstraction levels. Considering this work's existence of 3 image pathways, let x_l^1, x_l^2 and x_l^3 signify the outputs of the l^{th} layer in streams 1, 2 and 3 respectively, the output of the l^{th} layer in stream s is defined in Equation (1):

$$x_l^s = H_l^s([x_{l-1}^1, x_{l-1}^2, x_{l-1}^3, x_{l-2}^1, x_{l-2}^2, x_{l-2}^3, ..., x_0^1, x_0^2, x_0^3]) \quad (1)$$

Furthermore, to decrease information lost during the intermediate CNN deterministic operations, it was found beneficial shuffling and interleaving feature map elements in intermediate layers, to stimulate more information exchange while safeguarding the previous needed operations. For this reason, Equation (1) is adapted to Equation (2):

$$x_l^s = H_l^s(\pi_l^s([x_{l-1}^1, x_{l-1}^2, x_{l-1}^3, x_{l-2}^1, x_{l-2}^2, x_{l-2}^3, ..., x_0^1, x_0^2, x_0^3])) \quad (2)$$

Ultimately, π is a function that randomly permutes the feature maps given an input. Consequently, the utilized equation concatenates feature maps in a different order for each branch and layer. Figure (1) illustrates the proposed mechanisms of this architecture. Each gray region represents a convolutional block, whereas red arrows indicate convolution operations. Additionally, black arrows represent the network's direct connections between feature maps from different layers, within and in-between the different streams. As a result, the input of each convolutional block (the maps before the red arrows) are the permuted concatenation operation (Equation (2)) of the outputs (maps after the red arrow) of all the preceding layers from both paths. Addedly, this architecture prioritizes one of the modalities' extracted features. In other words, one of the paths is considered to be the main modality in the network. The modality with the highest correlation to the pertinent features is usually nominated to be the main modality. The secondary modalities' features are fused into the main modality. More specifically, the proposed model fuses the output of the last layer of each Denselayer of the secondary modalities to the corresponding denser layer of the main modality. This can be represented by Equation 3, where $x_{US,l}$ and $x_{MLO,l}$, secondary modalities, are the output of the dense layer from the US and MLO modality respectively of the dense layer l.

$$\bar{y} = \bigoplus_{x_{US,l}, x_{MLO,l}}^{x_{CC,l}} \quad \forall l \in 1, 2, 3... \quad (3)$$

By doing so, this network ensures that the relevant features are reinforced whilst additional features that may increase the description of the model are included. In order to assess the relevance of each modality, unimodal experiments are required, resulting in the selection of the best performing modality as the main one. The output of the main modality is associated with a max pooling layer and a ReLU layer, which will finally be associated with the classifier. In this work, the classifier will output between one of these three classes: *Normal*, *Benign* or *Malignant*.

3.4 Dataset Impact

The DL system's performance is reliant on the gathering and analysis of measurement data (*i.e* of datasets). The quality of the training data is the differentiation factor of the functioning system, as for the same exact DL model, different datasets produce different accuracy results. Finding a high-quality dataset, to assure the maximization of the performance of a DL system is, therefore, an imperative task [16]. Authors specially advocate the impact of dataset size and image resolution on the learning competencies of DenseNets applied to mammographies, concluding that performance is directly proportional to both [9, 17]:

- (Impact of Amount of Images): Considering deep-learning architecture's paradigm and, more particularly, convolutional neural networks that solely rely on the visual interpretation of images, providing a sufficient number of examples for the model is essential for robust predictive results [3]. Moreover, as breast lesions have different sizes and shapes, it is essential to classify diverse tumor morphologies.
- (Impact of Image Resolution): The quality (resolution) of the input images is considered as an influencing factor on the performance of deep neural networks. Various elements can have an impact on the image quality, including the presence Understanding the optimal balance between the resolution of input images and the computational and model efforts is essential to achieve maximum performance.

Despite these needs, finding a large high resolution dataset of breast exams is a difficult chore [9], as optimal datasets are very difficult and expensive to produce: dataset compilation relies on the very time-consuming tasks of gathering data, labeling it, and verifying the made correspondences. With this said, in this work, having a correctly labeled, large-enough dataset of both the mammography (CC and MLO views) and ultrasound modalities is a crucial precondition for its development.
With the interest of studying the necessary procedures to produce a usable dataset and, more specifically, to fully understand the pipeline of a dataset curation,this work focused its exertions on curating images of the mammography and ultrasound modalities.

4 PROPOSED SOLUTION

Following the state-of-the-art concepts gathered in the last section, this section describes the two distinct propositions this work aims to achieve.

4.1 Dataset Curation

This work proposes the curation of a private dataset to be used in the classification system's implementation stages. The data curation work ultimately produced two artifacts:

- (Repository): A github repository with two folders that contain sets of labeled and validated images with two distinct preprocessing treatments ('stretched' or 'preserved').These images have a 224×224 pixel size.
- (Database) :A .csv file that serves as a database, where every DICOM exam's information is stored. More particularly, for each patient's entry, the following fields are featured: the

Patient ID, the Acquisition Date, the Image Modality {MG, US}, the Modality Subtype {CC, MLO, NA}, the Laterality {Left, Right} ,the Associated BI-RADS {1-5} and links to the 224×224 image and DICOM repositories.

Both of these artifacts were created with the help of several data-processing scripts:

- (Data-Labeling Script): A python program is responsible to traverse through the different DICOM-filled folders, inspect their fields, extract the relevant database information and manually associate them to the medical diagnosis.
- (Validation Script): Cross-factors the developed database with the original exam's data, outputting for each element of the dataset if there are any inconsistencies between the two.
- (Preprocessing Scripts):A preprocessing script responsible for standardizing the data's appearance before it is fed into the network was created.

4.2 Classification Model

The development of classification model with a Hyperdense Network. This network processes the mammography image modality, with both CC and MLO views and the US image modality. More particularly, in this architecture, dense connections not only within each of the enumerated views but also across different modality streams will be featured, consequently promoting a multi-path architecture of a complete Hyperdense Network.
The network will take as input both mammography and ultrasound images and will output between either Normal, Benign, and Malignant classes. The classes correspond to BIRAD intervals that will be discussed in more detail in Section 4 6.1.5. This network's organization pursues the intuition that connected branches, as opposed to sequential branches, provide better results in multimodal systems, as it was implied in [7]. Some empirical results are to be made in order to identify the main branch to weigh heavier on the output .
By observing the training workflow of the classification model, one can understand that it is a multi-class classification problem, as the model's output has three different classes. Moreover, the proposed model will learn in a weak-label end-to-end fashion. For that reason, as the loss function, training will use cross-entropy, whose expression is in Equation (4).

$$L_{cross-entropy}(y_t, \hat{y}_t) = -\sum_{c=1}^{C} y_{t,c} \times log(\hat{y}_{t,c}) \tag{4}$$

where C denotes the number of classes , while $y_{t,c}$ and $\hat{y}_{t,c}$ denote the true and predicted probabilities of sample t belonging to class c, respectively.

For the purpose of studying the proposed Hyperdense net architecture, in addition to the unimodal testings on a densenet architecture, the progression of the hyper network's testings will follow:

- (Baseline): model 1 has a Hyperdense Net architecture with a densenet-121 branch architecture. The model will use the 'stretched'
- (DenseNet Custom) : model 2 will have a Hyperdense Net architecture with a custom densenet-101 (model 1-A) and densenet-37 (model 1-B) branch architecture. This model will use the dataset that had the best results in the Baseline model 1.
- (Class Weight): This model (B-I) will use a custom Hyperdense network with Cross Entropy Loss to mitigate the unbalanced classification threat.

Finally, once the model reaches its optimal point, a comparison between the unimodal densenet trials and the hyperdense net results will be made, in order to assess if the multimodal approach with a hyper connectivity pattern had better classification performance in a weak-label paradigm.

5 EVALUATION METRIC

This work will utilize the most descriptive metrics in the breast cancer classification task. More particularly, after training, Precision (Equation (5)), Recall (Equation (6)) and Specificity (Equation (7)) will be used as metrics to assess the model quantitatively. The used evaluation metrics had to take into account the fact that this work utilizes an unbalanced dataset - since the dataset has more examples of a specific class, the probability of the model leaning towards that diagnosis becomes higher, which jeopardizes the model's prediction capabilities whilst, concurrently, might mislead the interpretation of the model's results [**?**]ref:m141, ref:m142).

For that reason, the equations for each of these metrics were adapted. Instead of calculating the global metric, across all samples and classes, it calculates the metric for each class separately, and averages the metrics across classes, weighing each class by its support. Furthermore, as these are binary class metrics, an adaptation for them to fit this work's multiclass problem was made by averaging the final score in a one-vs-rest approach. Each of these systems of measurement follow below. The used notation signals TP as *True Positive* cases, TN as *True Negative* cases, FP as *False Positive* cases and FN as *False Negative* cases.

$$Precision_{macro} = \frac{1}{|C|} \sum_{c=1}^{|C|} \frac{TP_c}{TP_c + FP_c} \tag{5}$$

$$Recall_{macro} = \frac{1}{|C|} \sum_{c=1}^{|C|} \frac{TP_c}{TP_c + FN_c} \tag{6}$$

$$Specificity_{macro} = \frac{1}{|C|} \sum_{c=1}^{|C|} \frac{TN_c}{TN_c + FP_c} \tag{7}$$

The Receiver Operating Characteristic is also going to be plotted. From the ROC curve, it is also possible to extract the Area Under the Curve (AUC) score of a model, which expresses the degree or measure of separability in a scale of 0 to 1, 0.5 being an uninformative classifier. Both of these metrics were also adapted to fit the one-vs-rest approach.
Furthermore, a $N \times N$ Confusion Matrix is also generated for each model, where N is the number of classes (3 in this work).
Utilizing k-fold validation and, more particularly, Stratified K-fold Validation because of the unbalanced nature of the dataset would have several advantages in this work. The model was instead run 5 times, with random splits of the dataset which is computationally equivalent to k-fold validation.

6 IMPLEMENTATION

This section does an inspection on the implementation specifics of each objective of this work.

6.1 Dataset Curation

In this work, the assembling of a usable dataset with both mammographic views (Cranio-Caudal and Medio-Lateral Oblique) and ultrasound modality was made.It considers the data labeling, validation, and preprocessing tasks that compose this dataset's assembling workflow, which is schematized in Figure (2).

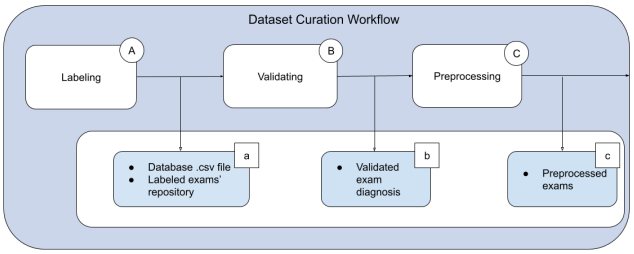

Figure 2: The dataset curation workflow.

6.1.1 Data Labeling. The data labeling task revolved around assigning a label to a sufficient amount of DICOM files. As this work's convictions revolve around a weak label approach, this work consisted solely of associating the exam with its corresponding medical diagnosis.
In order to label the data, this work utilized the Pydicom library that provides an interface for DICOM object manipulation and reading. A python script was responsible to traverse through the different DICOM-filled folders, inspect their fields, extract their relevant information and manually associate them to their medical diagnosis. The script takes as input a path to the folder containing the DICOM files and outputs the file that is going to be stored in a repository that serves as the database. It is stored in two different formats: the raw DICOM file and a 224× 224.jpg file. Due to the DICOM format intricacies and diagnosis-assignment spreadsheet impediments, the data labeling task was not automated and required manual checking of the exam's diagnosis and inputting it onto the script.
Every DICOM file's information was stored in a database - a .csv file - with the following fields: the Patient ID, the Acquisition Date (date the team acquired the exam), the Image Modality {MG}, the Modality Subtype {CC, MLO}, the Laterality {Left, Right}, the Associated BI-RADS {1-5} and links to the 224× 224 image and DICOM repositories. Other data annotations -breast density or lesion positioning, for instance - were non-existing and fell out of the scope of this work's weak label approach.

6.1.2 Validation. Limiting incorrectly labeled data is an important task in a data curation effort. In order to ratify the labeling process, a python validation script was created. This script crossfactored the initial spreadsheet with the developed database information (the csv file that was generated by the data-labeling script) and required two distinct stages. By traversing the files the program would output a file that categorized the database information's states for each of the exams. Following the tagging process of both files, a statistical report on the amount of information that falls into each category is outputted as well.

6.1.3 Preprocessing. Current literature supports preprocessing as an essential step in mammogram image analysis [mcite]. It involves image normalization, grayscale contrast enhancement, breast segmentation, and noise removal.This pipeline is shown in Figure (3).

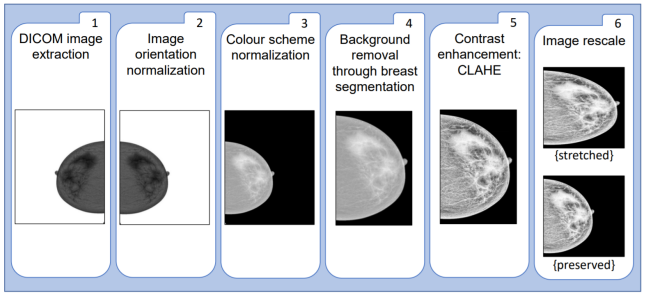

Figure 3: The dataset preprocessing pipeline.

In the first stage of the pipeline (1), the .TIFF image is extracted from the DICOM file. The .TIFF's extraction was made with the pydicom library. Then, image orientation is normalized (2) by horizontally flipping all images with 'Right' laterality in order to have all breasts facing right. Thereafter, the image's colour scheme is also normalized (3), to ensure the breast contains the brighter pixels and the background contains the darker pixels. Both step (2) and (3) are skipped in the Ultrasound modality, as they don't share the same image particulars.
Removing artifacts, labels and the pectoral muscle follows (4). These artifacts are often texts in the mammogram, indicating the modality of the exam, patient information or the hospital's name. In order to achieve breast segmentation, the contrast of the image is increased by 4 times, ensuring that the pixels are either white or black. Thereafter, the image is analyzed as a l*c matrix, where the average brightness of pixels is computed for each row and column, generating two uni-dimensional matrices. Through the analysis of these matrices, it is possible to detect the extent of bright pixels contained in each row or column, from which the breast area is detected and thus achieving segmentation. This procedure is done for both mammography vistas (CC and MLO) and for the ultrasound modality.
Then CLAHE is applied. It partitions the image into 8X8 pixel tiles and performs histogram equalization in each of them. This method depends on the presence of noise in each tile - if the tile has no noise, its histogram is in the 8×8 region; when noise is present, those pixels are uniformly scattered between the surrounding tiles before employing histogram equalization, to prevent noise amplification. Lastly, bilinear interpolation to remove the visible borders between the tiles is applied.

The final preprocessing stage involves image resizing (6) to 224×224 pixels, since after the cropping stage (4) the image has an unbalanced width-length ratio. The rescale is done in two different

manners in the mammography modality: stretching the image to fit the 224×224 size and preserving the breast's proportions by adding additional black pixels on the right to predo the rest of the mammogram. The first image type was named "stretched" and the other "preserved".

6.1.4 Statistical Report. Table 3 illustrates the statistical information of the curated dataset on the number of exams per class per modality.

Table 1: Dataset statistical information.

Modality	Modality Subtype	Class	Number of Images
US	NA	Normal	109
US	NA	Benign	92
US	NA	Malignant	87
MG	CC	Normal	227
MG	CC	Benign	208
MG	CC	Malignant	104
MG	MLO	Normal	180
MG	MLO	Benign	156
MG	MLO	Malignant	103

There are a total number of 288 Ultrasound curated images, 539 Cranio-Caudal Mammographies and 439 Medio-Lateral-Oblique Mammographies. The dataset is unbalanced not only on the amount of exams per modality (there is a gap between the number of ultrasound and MG-MLO, and finally to the MG-CC exam) but also on the amount of exams per diagnosis - across every exam modality, there is a lack of 'Malignant' diagnosed exams, whilst the *Normal* class has a larger number of examples.
The presence of an unequal distribution of images between classes has an impact on the model's performance and is a common problem in datasets. Consequently, the statistical distribution of exams has to be taken into consideration once evaluating the system, as this model's predictive capabilities can be jeopardized by the lack of a balanced number of examples between classes. Finally, since the Hyperdense Network has as input three images of ultrasound and CC and MLO mammograms from the same patient. An overview of the amount of patients that have all or some of the modalities is stated below:

- Number of patients with 3 modality types of exams: 195
- Number of patients with 2 modality types of exams: 78
- Number of patients with 1 modality type of exams: 53

6.1.5 Final Considerations. Some simplification for this work's classification task was made - the BIRAD score was replaced by a medical qualitative diagnosis. Table 2 illustrates this work's association between the BIRADs score and the classification task's class.

Table 2: The association between the radiologist's diagnosis (BIRAD) and the system's designed output class.

BIRAD	Class
{1}	Normal
{2,3}	Benign Finding
{4,5}	Malignant Finding

6.2 Classification Model

This section contemplates the several implementation efforts that had to be made in order to create a Hyperdense model and is subdivided into the three fundamental tasks that ultimately compose the stages of this work's model execution - unimodal testing, data loading customization and hyperparameter assessment.

6.2.1 Unimodal Testing. Since the proposed Hyperdense Network architecture has three distinct interconnected branches that contemplate a Densenet network topology, unimodal testing with a Densenet model, as reported in literature [11, 34], is extremely important and had to be performed. The opted densenet architecture was the densenet-121. Designating which of mammograms had the most valuable information was tested in this architecture.
The results indicated the CC mammography vista as the best modality, this was due to the lack of Ultrasound exams that is usually better in the unimodal trials .

6.2.2 Target Specification. Following the network's pipeline, the three mammography examinations have to be fed into three branches of the Hyperdense network. Despite the exams being of the same patient, these can have different diagnostic results, as each of the mammograms show different visible lesions in their modality domain. Dealing with inputs with multiple targeted outputs adds a different layer of complexity for this model's learning process.
In this work's context, it is more desirable to reduce the rates of False Negatives (a less severe wrong diagnosis) than it is to reduce the rates of False Positives (a more severe wrong diagnosis) in the medical field. Considering a more conservative approach to the overall diagnosis in this problem domain could potentially save lives.
For that reason, the designated target of each three-image "batch" is the one with the medical exam with the highest (more severe) diagnosis.

6.2.3 Hyperparameter Assessment. The Hyperparameter assessment of the Hyperdense network is reliant on its previous densenet-121 architecture. Consequently, unimodal testing was also intended to inspect the several hyperparameters, and Table 3 illustrates their settlement.

In order to prevent overfitting several changes were made on the branche's architecture. More particularly, on the number of convolutional layers: a combination of (16,16,16) convolutions in dense blocks was adopted as opposed to the previous (6, 12,24,16) dense block configuration of the densenet-121. These adjustments didn't allow for the usage of the preloaded Image Net weights but

Table 3: Hyperparameter Specification

Parameters/Model	Densenet-121	Hyperdense Net
Starting LR	$\eta_0 = 10^{-5}$	$\eta_0 = 10^{-5}$
LR Decay	Exponential ($\gamma = 0.98$)	MultiStepLR ($\gamma = 0.1$)
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.99$, $\epsilon = 1 \times 10^{-8}$)	
Number of Epochs	100	100
Batch Size	32	32
L_2 Regularization	1×10^{-4}	1×10^{-4}
Number of Folds	$k = 5$	5
Dropout	$p = 0.4$	$p = 0.4$
Densenet Compression	$\theta = 0.5$	$\theta = 0.5$
Densenet Growth Rate	$k = 32$	$k = 32$

were necessary as model trials with a pretrained densenet-121 network were worse than with a densenet-101 without pretraining. The dense block reduction also accounted for fewer parameters for the model to train.

Weight regularization techniques together with dropout were also applied to improve the model's generalization capabilities. L_2 regularization was opted for as network trials produced slightly better results than with L_1 regularization. Dropout was also added - in the densenet trials, a bigger probability proved to achieve better performance, whilst in the HyperdenseNet architecture it was preferable to use a lower value ($p = 0.25$).

Data Augmentation techniques were applied to overcome data shortage. More particularly, conducted experiments proved random rotation and random flipping to be more useful. Multi step LR decay proved to be better for the Hyperdense Net model than the Exponential LR decay that was used in the densenet trials. Multistep LR decays the learning rate of each parameter group by gamma ($\gamma = 0.1$ in this work's model) once the number of epoch reaches two milestones - half of the number of epochs and in the last quarter of the model's training stages.

sectionResults Firstly, the densenet unimodal testing is illustrated in Figure (4). These trials used a densenet-121 model with a (6, 12, 24, 16) convolutional dense blocks' configuration. As stated in Section 6.2.1, this model found the Cranio-Caudal view of a mammography exam as the modality with the best information. Despite literary studies presenting the ultrasound modality as the one with the biggest information gain, as this modality had less images than the CC mammographies, the results were expected. By curating more ultrasound exams, the main modality is expected to shift towards the original one.

The dataset that achieved better results in the unimodal trials was the preprocessed 'stretched' dataset, and this finding was extrapolated to the Hyperdense network's results. This conclusion follows the intuition that an increased number of pixels of the breast is preferable than unadulterated pixels. Furthermore, in future work, with the help of clinical physicians, following publicly available dataset's trends like the InBreast dataset or the CBIS-DDSM, by

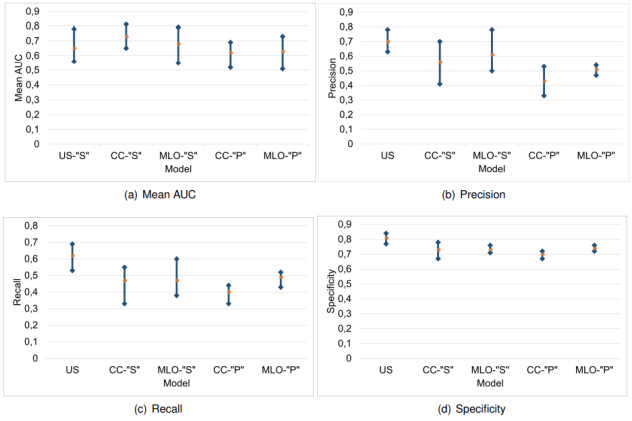

Figure 4: Evaluation results for the unimodal tests with a densenet-121 model. [Compared Models- US: Ultrasound exams resuls; CC-"S": Cranio-Caudal mammography with 'Stretched' preprocessing; MLO"-S": Medio-Lateral Oblique mammography with 'Stretched' preprocessing; CC-"P": Cranio-Caudal mammography with 'Preserved' preprocessing; MLO"-P": Medio-Lateral Oblique mammography with 'Preserved' preprocessing]

cropping the image in the area where the tumor is present, the model is expected to achieve better performance. This assumption is stated throughout SoA literature as the model would be able to restrict its feature extraction on the lesion's area and wouldn't be misled by uninformative pixels that can illustrate ambiguous shapes, consequently leading to a wrong classification. Utilizing images with better resolution would also have a positive impact on the network's performance.

The baseline Hyperdense Net model features a Densenet-121 architecture in each branch. Further testing with the block's configurations was made, as well as with the proper optimization techniques. The results are presented in Figure (5) The model's block configura-

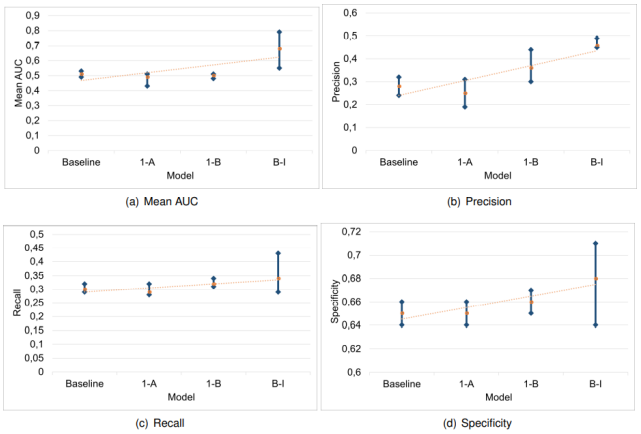

Figure 5: Evaluation results from the proposed classification models.

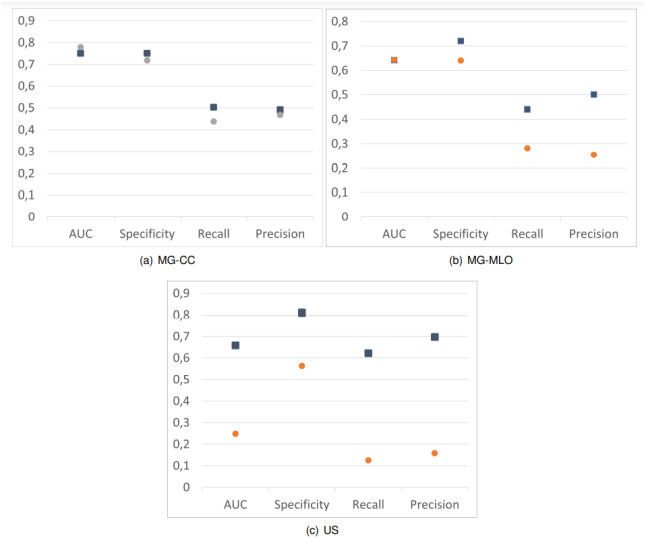

Figure 6: : Comparison between each branch of the Hyperdense network (orange) and each unimodal test (blue) in each evaluation metric.

tions pursued simpler arquitectures at first, to leverage the complex nature of the Hyperdense's branches that were harming its abstraction strength. Through a variety of testings, the Hyperdensenet's results were unresponsive to the decrease of the dense block's complexity. Only by decreasing the number of dense blocks from 4 to 3 and keeping a high number of trainable parameters, did the network start to regain performance. More concretely, the branches' dense block architecture was restricted to (16, 16, 16).

Once admitting the unbalanced nature of the dataset was being too harmful to the network's results, a Weighted Random Sampler was added to balance out the dataset's class count. This technique increased the model's predictive capabilities, and could testify that such a method is crucial for models whose input batches have different class rates in the dataset.

Additionally, a comparison between a Hyperdense Architecture with a multiclass approach and a single view approach can also be analysed. By isolating each hyperdense branch prediction and comparing its results with the unimodal densenet architecture's results for each of the modalities, conclusions on the utility of the interchanged information between modalities can be extracted in this work's settings. The comparison is illustrated in Figure (6), where the blue squares are a Hyperdense net's branch result and the orange dots are a Densenet's unimodal result for each modality for each evaluation metric. The illustrated comparison between the hyperdense branches and the unimodal test for each of the modalities' results indicates that the hyperdense net's information exchange between the branches is not enough to outperform the increased architectural complexity of the Dense net. That is, the lower complexity of each Densenet in the Hyperdensenet's branches, once comparing it to the unimodal Densenet's complexity is possibly compromising the model's predictive capabilities.

Finally the Hyperdense Network classification results did not reach the same capacity as models with the same architecture with annotations. Limited work is published with weakly supervised deep learning models in the Breast Cancer field, specially with this model's architecture, which limits the capability to compare this work's results.

However, the way the results stagnated in these values can be explained by a number of factors. Firstly, as three distinct mammographic exams are featured as input that have both different semblance and different tumour appearances, each HyperdenseNet branch is being fed with too few and dissimilar images. Despite having constant feature exchange, the fact that these features can be so divergent and with such little image representation, can ultimately jeopardize the output result. This constraint could be moved past with more data or with the usage of data annotations. Furthermore, as previously stated, a considerable amount of incomplete input batches are passed through the network, which threatens the model's training stage. Considering that each input is reliant on three images, the input quantity is dropped to a third of the dataset's amount, which results in an insufficient amount of images for the model to learn each class.

Such a deep architecture did not have the proper amount of images to build a robust predictive capability comparable to SoA Hyperdense models that make use of annotated datasets (Supervised Classification models). Nevertheless, these can be seen as promising results in the field of Weak Supervision, with possibility of improvement. The next Chapter will culminate this work's findings with a conclusion on the overall judgments and suggestions on the possible advances of these findings.

7 CONCLUSION

Considering the high BC mortality rates, Computer-Aided Detection (CAD) systems have been developed in the hopes of assisting radiologists in the detection of breast lesions. Traditional CAD systems utilize fairly simple mechanisms to identify such lesions, which resulted in various studies that questioned the relevance of these practices. Recent work promoting the usage of Deep Learning (DL) models have made major breakthroughs on the performance

evaluation of CAD approaches. In this work, a refined DL model implementation following today's SoA practices was discussed and thoroughly evaluated, alongside the description of a standardized pipeline for data curation. Subsequently, this work supports the discussion of different innovative DL architectures for screening considering a multimodality overview. More particularly, a Hyper-Dense Network is proposed to be implemented following a multimodal and weak-label approach. Furthermore, due to the large annotated datasets shortage, some techniques to overcome input scarcity were also addressed, culminating in the establishment of a usable dataset with ultrasound, cranio-caudal and medio-lateral oblique mammograms.

Finally, contemplating the presented results, the possible drawbacks - data insufficiency, unavailability of high computational power, unbalanced dataset and model architectural constraints - were also discussed, highlighting the necessary requirements to better the parametrized evaluation criteria.

7.1 Achievements

Study in the area of DL hints to a future with improved supporting tools for mammography screening practitioners. More particularly, deep architectures with a weak label approach can move past the constraint of data unavailability in the breast cancer field that is due to the lack of time and funding resources. Although achieving performant CAD models through weak label classification could have a positive impact on their future employment in medical facilities and DL research, classification models with a weak supervision paradigm is still an open area of investigation [1, 38].

On the other hand, creating a standartized curation process with preprocessing and validating scripts stands was one of this works biggest accomplishments. Despite the number of curated images and the initial labeling difficulties, with more curated data, data augmentation techniques and class weighting techniques, this threat could be diminished.

Considering the proposed classification model, there were no substantial improvement on the results to state. However, this work's implementation following a weak label approach with an unbalanced set of images has sufficient facilities to indicate that with a more complex pipeline or a bigger dataset size, one could extract better results. These alternatives are described in the next section.

7.2 Future Work

This work's classification results added to the collective knowledge on the importance of a balanced dataset in a complex three-input architecture. Various limitations were stated throughout the sections, computational power and dataset size being the most relevant ones. Considering different model architectures can also be seen as possible future improvement.

Firstly, as the computational resources were scarce, downsampling the images to (224×224) was necessary. Future work could try to soothe the downsampling to higher resolutions either by increasing the computational abilities or by establishing a network pipeline capable of decomposing a mammogram into smaller manageable items and passing each section through the network as input whilst still contributing to the same output result, like in the work of [2]. Additionally, architectural changes that are also being reported in literature to accommodate higher resolution images can also be employed. In [18] the authors present a network capable of having (2600×2000) pixels image resolution by reducing the size of the feature maps in the early stages of the network - by managing the strides size in the first convolutional and pooling layers - and in this way accepting, with less information loss, bigger inputs. As to the dataset's size, the team that aided in the development of this work is still acquiring more mammograms. The scalability of the dataset is possible and is expected to occur in the next few months. Considering the weak-label approach this work tried to explore, it is difficult to find publicly available datasets to test on different mammograms in such a manner. However, pretraining the network with annotated datasets and moving on to a weak supervision paradigm by preloading the model could possibly move past the various constraints of purely image level annotations.

Finally, as the medical and scientific research branches are still independent, medical exams are not yet seen as potential input to researchers. Establishing a gateway between the two would deeply benefit both communities and is stated in this work as a necessity to develop a better collective understanding on the Breast Cancer field.

REFERENCES

[1] D. Abdelhafiz, C. Yang, R. Ammar, and S. Nabavi. 2019. Deep convolutional neural networks for mammography: advances, challenges and applications. 20, S11 (June 2019). https://doi.org/10.1186/s12859-019-2823-4

[2] D. Abdelhafiz, C. Yang, R. Ammar, and S. Nabavi. 2019. Deep convolutional neural networks for mammography: advances, challenges and applications. 20, S11 (June 2019). https://doi.org/10.1186/s12859-019-2823-4

[3] A. Althnian, D. AlSaeed, H. Al-Baity, A. Samha, A. Bin Dris, N. Alzakari, A. Abou Elwafa, and H. Kurdi. 2021. Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. *Applied Sciences* 11, 2 (Jan. 2021), 796. https://doi.org/10.3390/app11020796

[4] R. A. Benedikt, J. E. Boatsman, C. A. Swann, A. D. Kirkpatrick, and A. Y. Toledano. 2018. Concurrent Computer-Aided Detection Improves Reading Time of Digital Breast Tomosynthesis and Maintains Interpretation Performance in a Multireader Multicase Study. *American Journal of Roentgenology* 210, 3 (2018), 685–694. https://doi.org/10.2214/ajr.17.18185

[5] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal. 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* 68, 6 (2018), 394–424. https://doi.org/10.3322/caac.21492

[6] F.M Calisto. 2017. Medical Imaging Multimodality Breast Cancer Diagnosis User Interface. (2017). https://doi.org/10.13140/RG.2.2.15187.02084

[7] G. Carneiro, J. Nascimento, and A. P. Bradley. 2017. Automated Analysis of Unregistered Multi-View Mammograms With Deep Learning. *IEEE Transactions on Medical Imaging* 36, 11 (2017), 2355–2365. https://doi.org/10.1109/TMI.2017.2751523

[8] R. A. Castellino. 2005. Computer aided detection (CAD): an overview. *Cancer Imaging* 5, 1 (2005), 17–19. https://doi.org/10.1102/1470-7330.2005.0018

[9] K. Dembrower, P. Lindholm, and F. Strand. 2019. A Multi-million Mammography Image Dataset and Population-Based Screening Cohort for the Training and Evaluation of Deep Neural Networks—the Cohort of Screen-Aged Women (CSAW). *Journal of Digital Imaging* 33, 2 (2019), 408–413. https://doi.org/10.1007/s10278-019-00278-0

[10] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. Ben Ayed. 2019. HyperDense-Net: A Hyper-Densely Connected CNN for Multi-Modal Image Segmentation. *IEEE Transactions on Medical Imaging* 38, 5 (2019), 1116–1126. https://doi.org/10.1109/TMI.2018.2878669

[11] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. Ben Ayed. 2019. HyperDense-Net: A Hyper-Densely Connected CNN for Multi-Modal Image Segmentation. *IEEE Transactions on Medical Imaging* 38, 5 (2019), 1116–1126. https://doi.org/10.1109/TMI.2018.2878669

[12] C. Dromain, B. Boyer, R. Ferré, S. Canale, S. Delaloge, and C. Balleyguier. 2013. Computed-Aided Diagnosis (CAD) in the Detection of Breast Cancer. *European Journal of Radiology* 82, 3 (2013), 417–423. https://doi.org/10.1016/j.ejrad.2012.03.0059

[13] S. W. Duffy, L. Tabár, A. M. Yen, P. B. Dean, R. A. Smith, H. Jonsson, S. Törnberg, S. L. Chen, S. Y. Chiu, J. C. Fann, M. M. Ku, W. Y. Wu, C. Y. Hsu, Y. C. Chen,

G. Svane, E. Azavedo, H. Grundström, P. Sundén, K. Leifland, E. Frodis, and T. H. Chen. 2020. Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549,091 women. *Cancer* 126, 13 (2020), 2971–2979. https://doi.org/10.1002/cncr.32859

[14] M. V. Euler-Chelpin, M. Lillholm, I. Vejborg, M. Nielsen, and E. Lynge. 2019. Sensitivity of screening mammography by density and texture: a cohort study from a population-based screening program in Denmark. *Breast Cancer Research* 21 (2019). https://doi.org/10.1186/s13058-019-1203-3

[15] Eurostat. 2020. Cancer statistics - specific cancers. https://ec.europa.eu/eurostat/statistics-explained/pdfscache/39738.pdf. [Last accessed 12-Nov-2020].

[16] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé, and K. Crawford. 2020. Datasheets for Datasets. arXiv:1803.09010 [cs.DB]

[17] K. J. Geras, S. Wolfson, Y. Shen, N. Wu, S. G. Kim, E. Kim, L. Heacock, U. Parikh, L. Moy, and K. Cho. 2017. High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks. *CoRR* abs/1703.07047 (2017).

[18] K. J. Geras, S. Wolfson, Y. Shen, N. Wu, S. G. Kim, E. Kim, L. Heacock, U. Parikh, L. Moy, and K. Cho. 2018. High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks. arXiv:1703.07047 [cs.CV]

[19] P Gordon. 2002. Ultrasound for breast cancer screening and staging. *Radiologic clinics of North America* 40 (06 2002), 431–41. https://doi.org/10.1016/S0033-8389(01)00014-8

[20] D. Gur, J. H. Sumkin, H. E. Rockette, M. Ganott, C. Hakim, L. Hardesty, W. R. Poller, R. Shah, and L. Wallace. 2004. Changes in Breast Cancer Detection and Mammography Recall Rates After the Introduction of a Computer-Aided Detection System. *JNCI Journal of the National Cancer Institute* 96, 3 (2004), 185–190. https://doi.org/10.1093/jnci/djh067

[21] E. Heeg, Y. A. Civil, M. A. Hillen, C. H. Smorenburg, L. A. E. Woerdeman, E. J. Groen, H. A. O. Winter-Warnars, and M. T. F. D. Vrancken Peeters. 2019. Impact of Second Opinions in Breast Cancer Diagnostics and Treatment: A Retrospective Analysis. *Annals of Surgical Oncology* 26, 13 (Oct. 2019), 4355–4363. https://doi.org/10.1245/s10434-019-07907-6

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2261–2269. https://doi.org/10.1109/CVPR.2017.243

[23] L. Irwig, N. Houssami, and C. van Vliet. 2004. New technologies in screening for breast cancer: a systematic review of their accuracy. *British Journal of Cancer* 90, 11 (2004), 2118–2122. https://doi.org/10.1038/sj.bjc.6601836

[24] Y. Jiménez-Gaona, M. J. Rodríguez-Álvarez, and V. Lakshminarayanan. 2020. Deep-Learning-Based Computer-Aided Systems for Breast Cancer Imaging: A Critical Review. *Applied Sciences* 10, 22 (Nov 2020), 8298. https://doi.org/10.3390/app10228298

[25] B. Lauby-Secretan, C. Scoccianti, D. Loomis, L. Benbrahim-Tallaa, V. Bouvard, F. Bianchini, and K. Straif. 2015. Breast-Cancer Screening — Viewpoint of the IARC Working Group. *New England Journal of Medicine* 372, 24 (2015), 2353–2358. https://doi.org/10.1056/NEJMsr1504363

[26] C. D. Lehman, R. D. Wellman, D. S. M. Buist, K. Kerlikowske, A. N. A. Tosteson, and D. L. Miglioretti. 2015. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. *JAMA Internal Medicine* 175, 11 (2015), 1828–1837. https://doi.org/10.1001/jamainternmed.2015.5231

[27] F. Mahmood, Z. Yang, T. Ashley, and N. J. Durr. 2018. Multimodal Densenet. arXiv:1811.07407 [cs.CV]

[28] A. Malich, D. R. Fischer, and J. Bottcher. 2006. CAD for mammography: the technique, results, current role and further developments. *European Radiology* 16, 7 (Jan. 2006), 1449–1460. https://doi.org/10.1007/s00330-005-0089-x

[29] S. Ohsuga. 1989. Toward intelligent CAD systems. *Computer-Aided Design* 21, 5 (June 1989), 315–337. https://doi.org/10.1016/0010-4485(89)90039-0

[30] N. Perry, M. Broeders, C. de Wolf, S. Tornberg, R. Holland, and L. von Karsa. 2008. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition—summary document. *Annals of Oncology* 19, 4 (April 2008), 614–622. https://doi.org/10.1093/annonc/mdm481

[31] N. Petrick, B. Sahiner, S. G. Armato, A. Bert, L. Correale, S. Delsanto, M. T. Freedman, D. Fryd, D. Gur, L. Hadjiiski, Z. Huo, Y. Jiang, L. Morra, S. Paquerault, V. Raykar, F. Samuelson, R. M. Summers, G. Tourassi, H. Yoshida, B. Zheng, C. Zhou, and H.P. Chan. 2013. Evaluation of computer-aided detection and diagnosis systems). *Medical Physics* 40, 8 (2013). https://doi.org/10.1118/1.4816310

[32] J. M Pérez-Rúa, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie. 2019. MFAS: Multimodal Fusion Architecture Search. arXiv:1903.06496 [cs.LG]

[33] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh. 2019. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Scientific Reports* 9, 1 (2019). https://doi.org/10.1038/s41598-019-48995-4

[34] J. Song, Y. Zheng, M. Z. Ullah, J. Wang, Y. Jiang, C. Xu, Z. Zou, and G. Ding. 2021. Multiview multimodal network for breast cancer diagnosis in contrast-enhanced spectral mammography images. *International Journal of Computer Assisted Radiology and Surgery* 16, 6 (May 2021), 979–988. https://doi.org/10.1007/s11548-021-02391-4

[35] W. Teh and A. R. Wilson. 1998. The role of ultrasound in breast cancer screening. A consensus statement by the European Group for Breast Cancer Screening. *European journal of cancer* 34, 4 (1998), 449–450. https://doi.org/10.1016/s0959-8049(97)10066-1

[36] Cancer Research UK. 2018. The importance of early diagnosis. https://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important. [Last accessed 4-Nov-2020].

[37] S. K. Yang, W. K. Moon, N. Cho, J. S. Park, J. H. Cha, S. M. Kim, S. J. Kim, and J. G. Im. 2007. Screening Mammography–detected Cancers: Sensitivity of a Computer-aided Detection System Applied to Full-Field Digital Mammograms. *Radiology* 244, 1 (2007), 104–111. https://doi.org/10.1148/radiol.2441060756

[38] Z. Zhou. 2017. A brief introduction to weakly supervised learning. 5, 1 (Aug. 2017), 44–53. https://doi.org/10.1093/nsr/nwx106