Medical & Biological Engineering & Computing manuscript No.
(will be inserted by the editor)

A Real Use Case of Semi-Supervised Learning for Mammogram Classification in a Local Clinic of Costa Rica

Saúl Calderón-Ramírez · Diego Murillo-Hernández · Kevin Rojas-Salazar · David Elizondo · Shengxiang Yang · Armaghan Moemeni · Miguel A. Molina-Cabello

Received: date / Accepted: date

Total tables: 6
Total Figures: 11

Abstract The implementation of deep learning based computer aided diagnosis systems for the classification of mammogram images can help in improving the accuracy, reliability, and cost of diagnosing patients. However, training a deep learning model requires a considerable amount of labelled images, which can be expensive to obtain as time and effort from clinical practitioners is required. To address this, a number of publicly available datasets have been built with data from different hospitals and clinics, which can be used to pre-train the model. However, using models trained on these datasets for later transfer learning and model fine-tuning with images sampled from a different hospital or clinic might result in lower performance. This is due to the distribution mismatch of the datasets, which include different patient populations and image acquisition protocols.

In this work, a real world scenario is evaluated where a novel target dataset sampled from a private Costa Rican clinic is used, with few labels and heavily imbalanced data. The use of two popular and publicly available datasets (INbreast and CBIS-DDSM) as source data, to train and test the models on the novel target dataset, is evaluated. A common approach to further improve the model's performance under such small labelled target dataset setting is data augmentation. However, often cheaper unlabelled data is available from the target clinic. Therefore, semi-supervised deep learning, which leverages both labelled and unlabelled data, can be used in such conditions. In this work, we evaluate the semi-supervised deep learning approach known as MixMatch, to take advantage of unlabelled data from the target dataset, for whole mammogram image classification. We compare the usage of semi-supervised learning on its own, and combined with transfer learning (from a source mammogram dataset) with data augmentation, as also against regular supervised learning with transfer learning and data augmentation from source datasets. It is shown that the use of a semi-supervised deep learning combined with transfer learning and data augmentation can provide a meaningful advantage when using scarce labelled observations. Also, we found a strong influence of the source dataset, which suggests a more data-centric approach needed to tackle the challenge of scarcely labelled data. We used several different metrics to assess the performance gain of using semi-supervised learning, when dealing with very imbalanced test datasets (such as the G-mean and the F2-score), as mammogram datasets are often very imbalanced.

Saúl Calderón-Ramírez · David Elizondo · Shengxiang Yang
Centre for Computational Intelligence (CCI), De Montfort University, United Kingdom

Saúl Calderón-Ramírez · Diego Murillo-Hernández · Kevin Rojas-Salazar
PARMA Research Group, Instituto Tecnológico de Costa Rica, Costa Rica

Diego Murillo-Hernández · Kevin Rojas-Salazar
School of Computing Engineering, Instituto Tecnológico de Costa Rica, Costa Rica

Armaghan Moemeni
School of Computer Science, University of Nottingham, United Kingdom

Miguel A. Molina-Cabello
Department of Computer Languages and Computer Science. University of Málaga, Spain
Instituto de Investigación Biomédica de Málaga - IBIMA, Spain

E-mail: sacalderon@itcr.ac.cr, diemurillo@ic-itcr.ac.cr, kevin.rojas7@estudiantec.cr, elizondo@dmu.ac.uk, syang@dmu.ac.uk, armaghan.moemeni@nottingham.ac.uk, miguelangel@lcc.uma.es

Keywords Transfer Learning · Breast Cancer · Mammogram · Semi-Supervised Deep Learning · Data Imbalance

1 Introduction

Breast cancer is one of the leading causes of death in women around the world [57]. Nonetheless, it is widely known that diagnosing a malign breast tumor in its early stages can increase treatment effectiveness [5]. In many situations, an early diagnostic can increase survival probability significantly.

Deep learning has extensively been explored and implemented as an approach to develop Computer Aided Diagnosis (CAD) systems using medical imaging [3,9,12, 17,18]. In 2012, a neural network architecture known as AlexNet won the ImageNet 2012 challenge. It featured a large neural network architecture, which implemented a set of novel techniques, which became a core part what was later referred to as deep learning. Later it became a popular approach for image analysis tasks. Deep learning can be defined as the set of architectures, training algorithms aimed to build very large neural networks, with millions of parameters. [28]

Deep learning based systems have the potential of highly improving the diagnosis and further treatment of patients. For mammogram analysis, different deep learning architectures have been proposed, for either binary classification, BI-RADS based multi-class classification, or segmentation of regions of interest [1,29]. Frequently, previously proposed architectures for mammogram classification (binary or multi-class), use large open datasets that have been gathered in a specific group of hospitals in one or few countries. These results might not be representative for a system deployed in a small hospital/clinic from a specific country (target hospital or clinic). When implementing and deploying a deep learning solution in such target hospital/clinic, usually a very small labelled dataset is available. Using small labelled datasets frequently hampers the model's generalisation and performance. Nevertheless, cheaper unlabelled data might available in the target hospital/clinic.

In this work, we explore the following setting: take a specific target clinic or hospital to deploy a deep learning model. Such data sampled from the target hospital/-clinic must be used for evaluation purposes. A a small number of labelled observations sampled from the target hospital/clinic might be available. Additionally, a larget unlabelled dataset is available in the target hospital/clinic. Furthermore, different datasets sampled from other hospitals or clinics might also be available. The notation of such experimental settings can be formalised as follows:

– Target labelled dataset D_t^l: A small number of labelled observations n_t^l might be available which can be used for training/fine-tuning the model.
– Source labelled dataset D_s^l: Different data sources of data sampled in different hospitals/clinics might be used. Usually these datasets have a large number of labelled observations, thus $n_t^l < n_s^l$.
– Target unlabelled dataset D_t^u: A larger number of unlabelled observations n_t^u might be available and can also be used for training/fine-tuning the model. As unlabelled data is cheaper to obtain, it can often be found that $n_t^l < n_t^u$.
– Source unlabelled dataset D_s^u: Similarly to the aforementioned case, more source unlabelled observations might be available when compared to the number of source labelled observations, thus $n_s^l < n_s^u$.

In this work, the usage of both transfer and semi-supervised learning using two different source datasets is explored: INbreast ($D_{s,\mathrm{IN}}^l$) [43] and CBIS-DDSM ($D_{s,\mathrm{DDSM}}^l$) [37]. The target dataset was obtained from the Costa Rican medical private clinic Imágenes Médicas Dr. Chavarría Estrada (hereafter referred as $D_{t,\mathrm{CR}}^l$). The aim of this research is to experiment the effectiveness of fine-tuning deep learning models in a semi-supervised fashion (using both D_t^u and D_t^l), performing transfer learning from models trained with the source datasets $D_{s,\mathrm{DDSM}}^l$ and $D_{s,\mathrm{IN}}^l$. For this study, the usage of unlabelled data from other source datasets was avoided, as it has been reported that it might decrease the performance of a Semi-supervised Deep Learning (SSDL) model [15, 16]. In this work, we use MixMatch as a semi-supervised learning approach [11], given previously positive results reported for this approach in medical imaging [13, 14].

This work proposes the usage of unlabelled data in fine-tuning with the MixMatch SSDL approach. The fine-tuning approach tested in this work refers to pre-training the model in a source dataset, to later re-train (fine-tune) the model using the target dataset. We compare semi-supervised fine-tuning to supervised fine-tuning (using the same target dataset for both cases). This is done as a mean of improving the performance of deep learning models on the task of binary classification of whole mammogram images under a real-life scenario using a novel target dataset. Evaluations and comparisons are drawn over the performance of deep learning models on the classification of mammogram images obtained in the context of the day-to-day basis of a local medical private clinic of Costa Rica. We test the combination of semi-supervised learning with other common approaches to deal with small labelled datasets, namely data augmentation and transfer learning. As for transfer learning, we test two different source datasets, in order to assess the impact of the source dataset in the performance of the model.

2 State of the Art

2.1 Transfer learning and data augmentation for mammogram classification

CAD of breast cancer via mammogram image classification has been widely studied in the literature. Authors in [1] present a survey of the state of the art in the application of deep learning in the analysis of mammography images for the early detection of breast cancer. The authors summarize open challenges and best practices to follow when dealing with mammogram analysis using deep learning. One of the most frequent short-comings of implementing deep learning for mammogram analysis in a target clinic/hospital is the lack of labelled training data [1]. This can lead to model overfitting to the dataset. Labelling medical images can be particularly expensive, as trained professionals are needed to carry out such specialized tasks [53]. To overcome this challenge, a number mammogram datasets are publicly available. However, different patient populations and image acquisition protocols can limit and hinder the performance of the final model using the target data [36].

Two of the most common approaches to tackle the problem of labelled data scarcity and subsequent model overfitting, are transfer learning and data augmentation [1, 29]. Using pre-trained model parameters from more general tasks often improve the model's performance. Authors in [26] experimented with the multiclass classification of mammograms using transfer learning from ImageNet. Similarly, authors in [46] observed encouraging results in the classification of mammograms when using transfer learning from a chest X-ray dataset of patients with pneumonia.

Applying transfer learning with models trained with observations from the same domain is intuitively an interesting approach. Authors in [4] carried out an exhaustive research for improving the performance of deep learning models in the binary classification of mammogram anomalies by using features previously learned from different mammogram datasets. Authors in [48] also experimented with transfer learning from mammogram datasets for the detection and classification of anomalies in mammogram images. For these cases the more specific term "domain adaptation" can be used, as although images from different datasets can be visually and semantically similar, their distributions might be significantly different, as explained in [20, 54].

As previously mentioned, data augmentation is also an effective approach to tackle data scarcity [1]. Simple augmentations by applying common image transformations like image rotations and flips can improve results [38]. In previous works, more sophisticated and domain-specific data augmentation techniques have been developed [25]. Authors in [19] obtained positive results by implementing elastic deformations for mammogram images, simulating possible different views of the same breast. The augmentation of training data has also been recently achieved by creating artificial observations with generative deep learning models [34, 58]. Alternative approaches to deal with small labelled datsets and meant to regularize deep learning models for mammogram classification, can be found in the literature [59]. For instance in [24] an Euclidian magnitude regularization approach is proposed in a deep learning pipeline for mammogram mass segmentation. More recently, adversarial augmentation combined with graph based regularization [40] has been proposed improve the model's generalisation for mammogram diagnosis.

Other methods to deal with small labelled target datasets such as semi-supervised learning (leveraging unlabelled data), have received comparably less attention in the literature. In this work our contribution can be summarized as the evaluation of common methods to deal with model overfitting in small labelled datasets (fine-tuning, data augmentation) combined with semi-supervised learning. We use a novel labelled dataset from a Costa Rican clinic, showing the practical challenges of using deep learning for mammogram analysis. Therefore, we include a data-centric approach in our proposed pipeline, as we evaluate the usage of different source datasets for transfer learning and further model fine-tuning using semi-supervised learning (along with data augmentation). The evaluation of the different configurations tested in this work, can shed light around the impact of using each one of the tested approaches individually and combined. This along the usage of different data sources and unlabelled data.

2.2 Semi-supervised learning for medical imaging and mammogram analysis

Another approach to deal with small labelled datasets is the usage of SSDL, which leverages unlabelled data to improve the model's performance [20]. In recent years, the usage of the cheaper and larger unlabelled datasets for training deep learning models has proven to be a viable option for handling the lack of labelled data, as well as improving the performance of models [13, 17]. Authors in [20] present a survey of recent literature of semi-supervised learning approaches for medical imaging. The survey shows how unlabelled datasets have been used for improving model training in brain tumor segmentation, detection of vascular lesions, and prostate cancer detection. More recently, the usage of unlabelled data with semi-supervised deep learning has proven to give positive results in the detection of COVID-19 in chest x-ray images [13, 17].

However, research on SSDL approaches for mammogram analysis is still limited. In [52] the authors propose a new semi-supervised architecture for convolutional neural networks, designed to extract information from multiple views of masses from mammogram images for their binary classification. In [6] a semi-supervised setup is proposed for the joint use of weakly labelled data with fully labelled data of mammogram regions in the detection and classification of anomalies. Authors in [53] also proposed a semi-supervised approach based on graphs and convolutional neural networks for the classification of anomalies in mammograms. However, from our knowledge few authors in the literature deal with the classification of mammograms using less expensive whole-image labels only. In [14] the MixMatch approach was tested to improve the accuracy and predictive uncertainty of models applied to the binary classification of whole mammogram images. A target hospital or clinic might not have lower level labels available, to fine-tune and test a deep learning model.

As previously mentioned, analysis of mammograms includes lower level tasks such as: segmentation and detection of anomalies, the higher abstraction of level tasks, the binary classification of images (malign findings with no/benign findings) [1, 29]. It may also include multi-class classification, for instance using the BI-RADS standard [25]. As such, different levels of annotations in the data might be needed for lower level tasks, like pixel-level annotations of the Region of Interest (ROI). When using transfer learning to leverage information from thoroughly annotated source datasets for lower level tasks, fine-tuning on the target data might still be needed [20]. This, similar degrees of annotations would be preferable in the target dataset as well. Therefore, the need to use target data to train or fine-tune a model makes the use of unlabelled data an interesting alternative. Different image acquisition protocols and patient distribution sampled in a dataset source is a frequent real-life scenario that increases the need of model fine-tuning.

2.3 SSDL with MixMatch

In this work, the MixMatch method is used as the semi-supervised learning approach for training models with unlabelled data. This is novel SSDL method, presented by the authors in [11] has shown important accuracy gain against previous SSDL frameworks. Given the performance boost reported by the authors in [11] of MixMatch against other state of the art semi-supervised methods, in this work we chose it to test the impact of semi-supervised learning for mammogram classification. It is mainly based on the use of pseudo-labels, unsupervised regularization and data augmentation. The following corresponds to a brief description of the method.

SSDL makes use of labelled and unlabelled observations X_l, X_u respectively. MixMatch implements data augmentation with affine transformations on both datasets. Pseudo-labels are then generated for each unlabelled observation, sharpening the average of the predictions of a model on each of its augmented "versions". This results in the set $\tilde{Y}$ of pseudo-labels for observations of X_u. Similarly, the set Y_l can be used to represent the labels of observations in X_l.

Further data augmentation is applied to the datasets S_l and $\tilde{S}_u$, with $S_l = (X_l, Y_l)$ and $\tilde{S}_u = (X_u, \tilde{Y})$, by using linear interpolation of the data with the MixUp algorithm, as mentioned in [11]. This way, the sets of augmented data $\tilde{S}'_u$ and S'_l are obtained and finally used to train a model by minimizing the compound loss function shown in equation 1.

$$\begin{aligned}\mathcal{L}(S,\theta) = &\sum_{(x_i,y_i)\in S'_l} \mathcal{L}_l(\theta, x_i, y_i) + \\ &\gamma r(\tau) \sum_{(x_j,\tilde{y}_j)\in \tilde{S}'_u} \mathcal{L}_u(\theta, x_j, \tilde{y}_j)\end{aligned} \tag{1}$$

This loss function is formed by the respective supervised and unsupervised loss terms $\mathcal{L}_l$ and $\mathcal{L}_u$. In this work, the supervised loss term is implemented as a cross-entropy loss, while the unsupervised term is implemented as an Euclidean distance, with the regularization coefficient γ and the rampup function $r(\tau) = \tau/3000$, as recommended in [13]. We refer the reader to the original publication in [11] for more details.

2.4 Class imbalance correction

A major factor that must be taken into account in the process of implementing a model for classification tasks, specially in the medical domain, is the distribution of classes in a dataset [1]. For medical conditions, it is common for observations depicting a disease or a "positive" case, to be fairly less frequent in comparison to normal or healthy observations [17]. Training a model with imbalanced data can lead to the final model being biased towards the majority classes, while ignoring the minorities.

Multiple approaches to tackle the problem of imbalanced class distributions in datasets can be found in the literature [17]. Two of the most straightforward techniques used include under-sampling and over-sampling [56]. These techniques, although fairly simple and intuitive, might not prove to be the best choice, as they can lead respectively to information loss and over fitting [56]. Other common approaches used towards imbalanced class distributions in datasets involve the so called "cost sensitive learning" [56]. One implementation of this approach is to give weights to each class inside the cross-entropy loss function to correct for class imbalance. In the case of semi-supervised learning, authors in [17] proposed a similar technique called Pseudo-label based Balance Correction (PBC). This technique applies class-balance correction both to the labelled and unlabelled data in the MixMatch SSDL approach. Given its reported positive results, we implement the class imbalance correction approach tested in [17] in our work.

2.5 Classification Metrics for Imbalanced Data

Class-imbalanced datasets and its impact on the implementation of classification models has long been a subject of study in the literature [35]. Using metrics that account for class imbalance is an important aspect, specially for CAD systems used under real-life conditions. The most frequent and almost customary method for evaluating the classification performance of models consists in the traditional classification accuracy [51]. Despite its wide usage, traditional accuracy is not an adequate metric for imbalanced test data settings [2]. This metric does not take into account the possible differences between the distribution of both classes, and thus can mislead to optimistic results, as illustrated by authors in [21].

Basic and widely known classification metrics that also derive from the confusion matrix scheme are the recall, specificity, and precision [2]. These metrics offer more information about the model's classification performance and have been used in the literature to provide more complete analysis in cases with imbalanced data settings [2, 33].

Precision, sensitivity and specificity measures provide values in the interval $[0, 1]$, where higher is better. While these metrics can be studied individually to analyse different dimensions of the performance of a model, other metrics can be used to summarize them into a single score or value. As discussed by the authors in [21], currently there is no consensus in the machine learning community on the ideal classification metric to use, specially in cases with imbalanced data.

Two of the most widely used classification metrics, besides traditional accuracy, are the F-1 Score and Area Under the Receiver Operating Characteristic Curve (AUROC). These metrics are commonly used in contexts prone to data imbalance, such as information retrieval [47] and the medical domain [51], although they are not always adequate for such cases [41]. The F-1 score corresponds to the harmonic mean between recall and precision. This metric is most useful in contexts where the main focus of a problem is the positive class, and the detection of the negative class is less relevant [51]. It offers a balanced score of the rate of true positives (recall) and the rate of correctly predicted positives (precision).

Nevertheless, multiple works and studies point out the deficiencies of this metric and discourage its use as a standalone measure for the classification performance of a model [21, 27, 41, 47], specially in cases of high class imbalance. Namely, one of the problems commonly pointed out is the fact that the F-1 Score weights the false positives (FP) the same as the false negatives (FN). To address this short-coming in imbalanced data scenarios is the F-2 score [23].

The AUROC is another single score metric that summarises the trade-off between the rate of true positives and the rate of false positives given multiple decision thresholds for the classification performance of a model. It provides a deeper insight of the model's behavior, when compared to the accuracy. However, it still faces many problems that are pointed out by a number of authors in the literature [10, 21, 30], some related to the impact of highly imbalanced data.

Other classification metrics that have been proposed and explored in the literature for data imbalance scenarios are the balanced accuracy and the G-Mean [2, 33, 35, 50]. Both of these metrics summarize the recall and specificity, offering a single score that balances the model's capacity to correctly classify observations belonging to both the majority (negative) and the minority (positive) classes. Both metrics rely solely on the recall and the specificity of a model. The balanced accuracy consists of the arithmetic mean of both metrics, while the G-Mean is their geometric mean. They can be useful in cases of imbalanced data, as values closer to 1 imply that a model has a high predictive power for both classes.

It can be noted that, while both metrics are similar, due to its mathematical properties, the G-Mean is less sensitive to outliers [2]. An example can be a model that achieves a perfect specificity of 1 by correctly classifying all negative samples, but with a low recall of 0.1. Here, the balanced accuracy would be 0.55, while the G-Mean would be 0.31. This shows how the balanced accuracy can be over-optimistic. In this work, the usage of the

G-mean as a metric is implemented as it takes into account the rate of true positives and true negatives for malign cases, as its the most under-represented class.

A wide variety of other classification metrics can be used for cases of imbalanced data, like the Matthews correlation coefficient [21]. This metric corresponds to a correlation coefficient between the observed and predicted classifications. Other metrics include the Youden's index and the Discriminant Power [51]. These metrics, although useful, are not as popular or widely used as the other mentioned classification metrics and might not be as intuitive to understand.

3 Methods

3.1 Experimental Setup

For this purpose several experimental configurations were analysed and carried out, as illustrated in Figure 1. Multiple models were trained under different training configurations to evaluate the impact of SSDL on their classification performance on a target dataset. Transfer learning (a simple "Domain adaptation" method) and loss function based class-imbalance correction were also tested. This was done as means for dealing with common difficulties of the implementation of classification models for real-life use cases, such as limited amounts of data and extreme class imbalance (further detailed in section 3.2.2).

Deep learning models were first trained in a supervised manner with complete mammography datasets $D^l_{s,\mathrm{IN}}$ and $D^l_{s,\mathrm{DDSM}}$ in order to obtain source-trained models, which were further fine-tuned on our target Costarrican dataset in a Supervised (Config. **S+FT**) or Semi-Supervised (Config. **SSDL+FT**) manner, with limited amounts of labelled observations n^l_t.

The performance of source-trained models, without fine-tuning on the target dataset, was also evaluated (Config. **S+No-FT**). The performance of models directly trained on the target dataset using SSDL, without domain adaptation from a source mammography dataset (Config. **SSDL**) was also tested. Class imbalance correction of the loss function with the PBC method developed in [17] was also used as part of the experiments of Configurations **SSDL+FT**, **S+FT** and **SSDL**. The empirical results obtained in this study showed a considerable impact of its usage for correcting data imbalance. Therefore, we included it to train all of the tested SSDL models. Finally, all models were evaluated on test images from our novel target Costarrican dataset.

Due to the extreme data imbalance present in the target dataset (95% of observations belong to the negative class and 5% to the positive class), specific classification metrics, aside from traditional accuracy, were evaluated as performance indicators. Following the research presented in Section 2.5, the G-Mean was chosen as main classification metric. This metric was used to provide insight related to the accuracy of the models on the positive class, without ignoring their predictive power at classifying the negative class. Other metrics including F-2 Score, accuracy, recall, specificity, and precision are also reported.

Table 1: Summary of datasets used in this work

	INbreast [43]	**CBIS-DDSM** [37]	**Target CR Dataset**
Origin	Portugal	USA	Costa Rica
Year	2011	1997-2016	2020
Number of cases	115	1566	87
Number of images	410	3103	341
Views	CC MLO	CC MLO	CC MLO
Image mode	Full-field digital	Digitised screen-film	Full-field digital
Categories	BI-RADS ACR Density	BI-RADS ACR Density Verified Pathology	BI-RADS
ROI annotations	Yes	Yes	No

Deep data set Dissimilarity Measures (DeDiMs) following the novel approach presented by authors in [15] were also evaluated, to provide a more thorough analysis of the impact of the choice of source datasets. This method consists in a simple and practical approach to compare different datasets by measuring their dissimilarity in the feature space of a generic deep learning classification model. We aim to quantitatively assess the similarity between the tested datasets and correlate it with the yielded results.

3.2 Mammography Datasets

Three different mammography datasets were used to carry out the experiments depicted in this work, summarized in Table 1. Sample images are shown in Figure 9. The selected datasets correspond to two popular and publicly available "source" datasets, used solely for model training: the INbreast ($D^l_{s,\mathrm{IN}}$) and CBIS-DDSM ($D^l_{s,\mathrm{DDSM}}$). A third novel "target" dataset $D^l_{t,\mathrm{CR}}$ comprised of mammogram images gathered from a private medical clinic of Costa Rica was also used.

3.2.1 Third-party Source Datasets

Introduced in [43], the INbreast dataset is a mammographic database comprised of multiple full-field digital

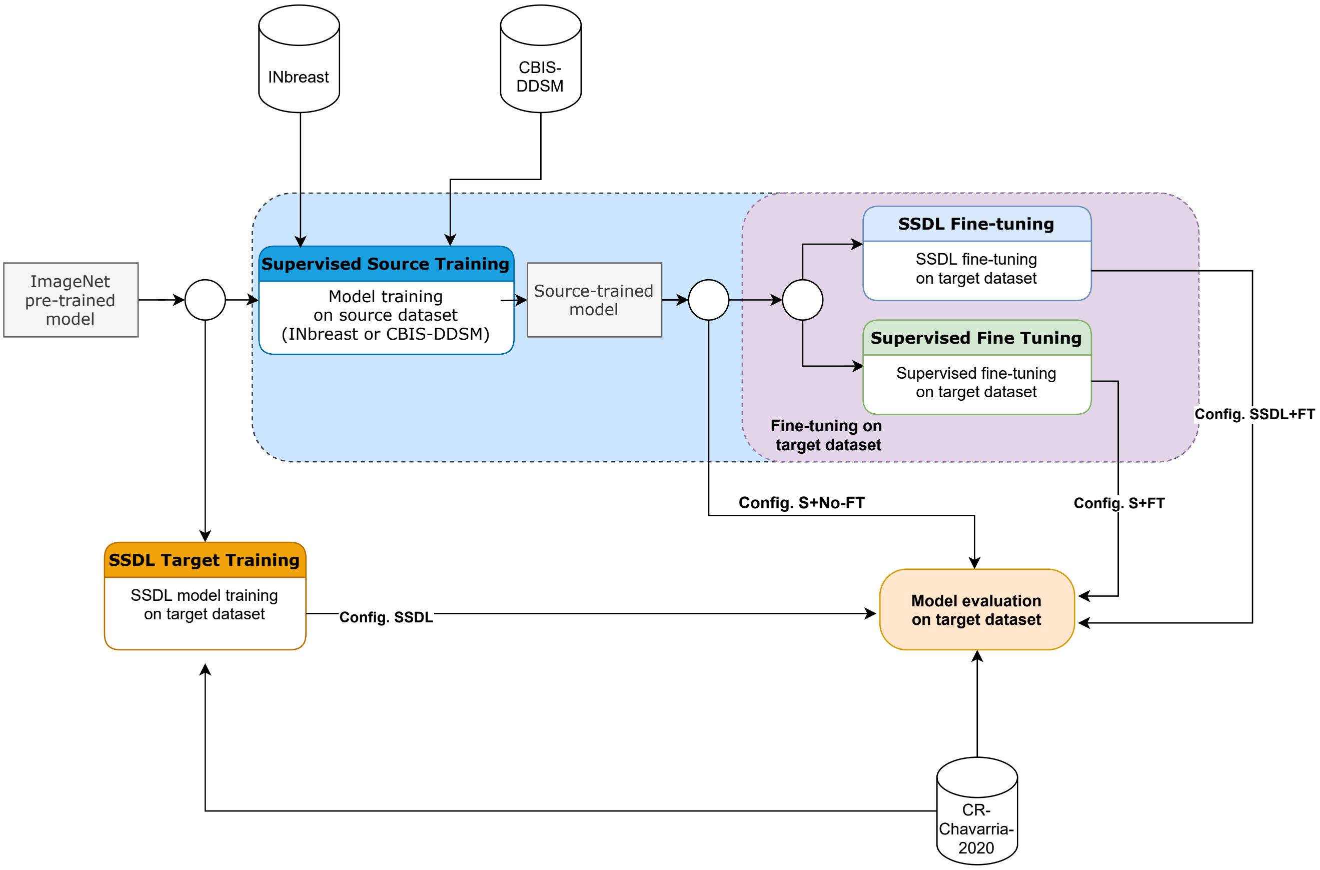

INbreast
CBIS-DDSM
ImageNet pre-trained model
Supervised Source Training
Model training on source dataset (INbreast or CBIS-DDSM)
Source-trained model
SSDL Fine-tuning
SSDL fine-tuning on target dataset
Supervised Fine Tuning
Supervised fine-tuning on target dataset
Fine-tuning on target dataset
Config. SSDL+FT
Config. S+No-FT
Config. S+FT
SSDL Target Training
SSDL model training on target dataset
Config. SSDL
Model evaluation on target dataset
CR-Chavarria-2020

Fig. 1: Diagram of experimental configurations presented in this work

mammograms of patients with a wide variety of anomalies like masses and calcifications. Each image is labelled according to the BI-RADS scale from categories 1 to 6 and their density measure with the American College of Radiology (ACR) standard. The dataset is composed of 410 images in total, collected from 115 different cases.

Since this work is focused on the binary classification of mammograms (i.e. according to the presence of breast anomalies), images from the INbreast dataset were divided into 2 groups. Similar to [48], mammograms labelled with BI-RADS categories 1 and 2 are defined as negative (benign) observations, and the ones labelled with categories 4, 5 and 6 are defined as positive (malign) observations. Mammograms labelled with categories 0 (non-conclusive) and 3 (probably benign) are ignored. For the INbreast dataset, this process results in 287 negative and 100 positive observations.

The Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) dataset, presented in [36] was made publicly available by the The Cancer Imaging Archive (TCIA) [22]. It corresponds to a curated and standardized version of the DDSM dataset [31]. The dataset comprises a total of 3103 digitised screen-film mammography images gathered from 1566 cases, labelled according to the type of anomalies present (masses or calcifications), their BI-RADS category, their ACR density measure and their verified pathology as benign (1728 images) or malign (1375 images). The dataset presents an overlap between cases that are classified as containing masses or calcifications, as some patients presented both. The total number of images detailed here represents the overall total of both mass and calcification cases, as obtained from [37] and subsequently used for model training.

3.2.2 Clínica Chavarría's 2020 Mammogram Target Dataset

The *CR-Chavarria-2020* dataset consists of a novel collection of full-field digital mammograms obtained from the Costa Rican medical private clinic Imágenes Médicas Dr. Chavarría Estrada, over a period of one year (referred as CR-Chavarria-2020 in Figure 1). The images are completely anonymized. Specifically, these images correspond to mammograms taken as a result of routinely medical appointments for patients of the clinic across the year 2020. The entire dataset is available for researchers, along with documentation of its distribution, annotations, and extra images that were discarded in the process of constructing the dataset. If the reader is interested in using our collected dataset, please make contact via email with the first author, as we plan to make the dataset publicly available in the future [1].

We highlight the value of this dataset as target data for the evaluation of deep learning models in the medical domain, as it is highly representative of the operation conditions that production-implemented models would have to deal with, in a medium sized clinic. The complete dataset, referred as $D^l_{t,\mathrm{CR}}$, consists of a set of BI-RADS-labelled images. These are also annotated in a similarly manner as the source datasets, with their respective anonymous patient id, gender, age, type of view, and depicted breast.

The complete $D^l_{t,\mathrm{CR}}$ dataset contains a total of 341 labelled images from 87 patients. Similarly to the INbreast dataset, images from $D^l_{t,\mathrm{CR}}$ were also subject to the same "binarization" process described above. This resulted in the binary-labelled target dataset $D^b_{t,\mathrm{CR}} \subset D^l_{t,\mathrm{CR}}$, with a total of 282 images; 268 negative and 14 positive observations from 68 and 4 patients, respectively.

Figures 2 and 3 illustrate the distribution of both BI-RADS and binary labels for $D^l_{t,\mathrm{CR}}$ and $D^b_{t,\mathrm{CR}}$ respectively. Here, the extreme class imbalance of observations can be better appreciated, being one of the most frequent and troublesome situations that arise in the implementation of machine learning models in the medical domain. In addition, figures 4, 5, 6, 7 and 8 show the distribution of other dimensions of both $D^l_{t,\mathrm{CR}}$ and $D^b_{t,\mathrm{CR}}$, like the depicted view and breast in each mammogram, along with the age of patients. These aspects show more balanced distributions, as is the case with most mammogram datasets, and that the regular age span for patients varies from 40 to almost 90 years old.

Along with the complete $D^l_{t,\mathrm{CR}}$ dataset, a set of discarded images has also been made available. These images were retrieved from the clinic, but were discarded due to low image quality or artifacts (i.e. patients with breast implants). Nevertheless, these could prove to be useful on further investigations, surrounding the robustness of models to domain-specific noise or corruptions in images [32].

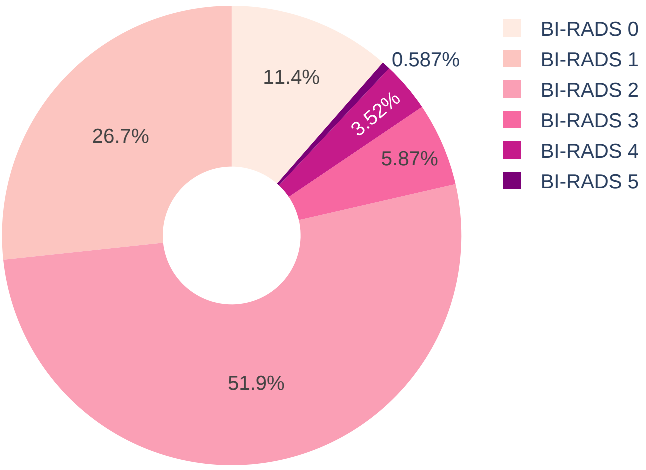


Fig. 2: BI-RADS categories distribution for $D^l_{t,\mathrm{CR}}$

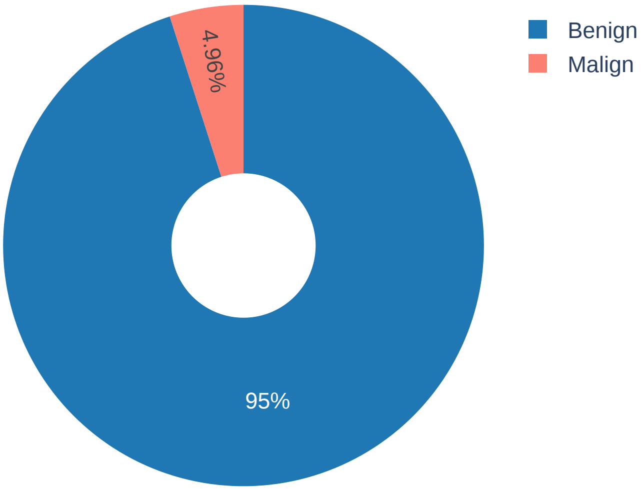


Fig. 3: Binary categories distribution for $D^b_{t,\mathrm{CR}}$

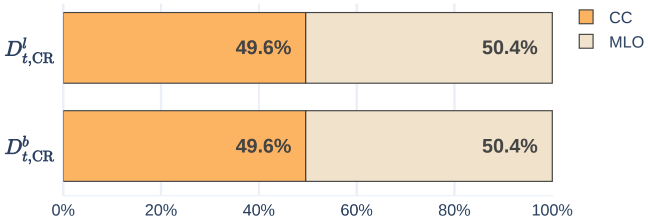


Fig. 4: Craniocaudal (CC) and Mediolateral Oblique (MLO) views distribution for complete and binary-labelled target datasets

3.2.3 Data Preprocessing

Mammograms from all three described datasets originally possessed considerably high image resolutions. In order to avoid memory constraints, all image files were resized to 224 × 224 pixels, after being converted from the DICOM format to the BMP one. Standardization was applied to all images. The mean and standard deviation, according to the respective dataset employed for training, were calculated (complete INbreast, complete CBIS-DDSM or the corresponding training partition of

[1] The authors using our novel dataset are required to cite this paper, for instance as: Calderon-Ramirez, S., Murillo-Hernandez, D., Rojas-Salazar, K., Elizondo, D. A., Yang, S., Moemeni, A., Molina-Cabello, M. (2021). A Real Use Case of Semi-Supervised Learning for Mammogram Classification in a Local Clinic of Costa Rica.

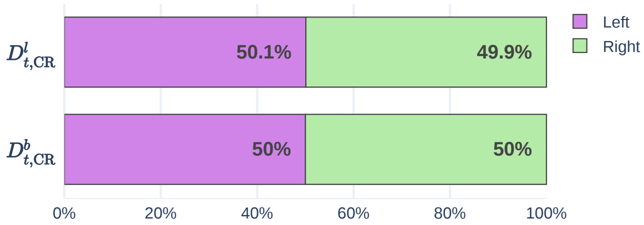


Fig. 5: Depicted breast distribution for complete and binary-labelled target datasets

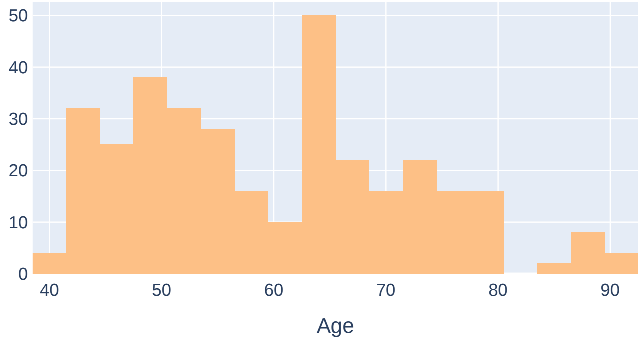


Fig. 6: Age distribution for patients in $D^l_{t,\mathrm{CR}}$

each of the data subsets of the target dataset). Then, for each image, the channel-wise pixel values were subtracted by the mean and divided by the standard deviation. Standardization is done for each training batch.

Additionally, through visual inspection of the images in CBIS-DDSM dataset, it can be noted that several mammograms contain multiple forms of noise, mainly due to the digitization process of the screen-film. Physical labels, orientation tags and scanning artifacts are some of the types of noise inducing elements that can be found in mammogram images, as illustrated in [44]. To minimize the effects of these types of noise, a similar approach to the one described in [8] was implemented and applied to images from the CBIS-DDSM dataset. This is shown in figure 11. Authors in [8] describe the implemented preprocessing pipeline in this work, designed for background removal in mammograms. The process consists mainly on the application of a rolling ball algorithm with $radius = 5$. This is followed by the application of Huang's fuzzy thresholding and morphological transformations of erosion and dilation. This process results in a binary map that can be used to remove background noise from an image. Such image preprocessing pipeline is implemented in this work, which makes use of the base code made available by the authors of [8] and algorithm implementations from the OpenCV library.

3.2.4 Experiments

All experiments described in this work were implemented in Python using the FastAI and PyTorch libraries, based on the MixMatch implementation described in [13] [2]. The PyTorch implementation of the VGG 19-layer with batch normalization was chosen as the main architecture for the models of all experiments. Additionally, experiments of Configurations **SSDL+FT** and **S+FT** were also carried out using PyTorch implementations of ResNet-152 and EfficientNet-b0. The complete results of experiments with these architectures are presented in the supplementary material. Transfer learning with pre-trained weights from ImageNet was used for the initial models of all experimental configurations. All depicted experiments were executed employing a total of 10 different randomly generated subsets $D^b_{i,t,\mathrm{CR}} | i = 1, ..., 10$ of the binary-labelled target Costarrican dataset $D^b_{t,\mathrm{CR}}$. Each with an average distribution of 70% of images for training and 30% for testing, with observations from different patients for training and for testing. Therefore, around 198 training images (including both labelled and unlabelled), and 82 test images were used.

The models for the configurations **SSDL+FT**, **S+FT** and **SSDL** were trained on each data subset $D^b_{i,t,\mathrm{CR}}$, with $n^l_t = 20, 40$ and 60 amounts of labelled observations, with 95% of observations corresponding to the negative class (benign) and 5% to the positive class (malign). Class-imbalance correction of the loss function was implemented, respectively, as a weighted cross-entropy loss for the supervised models and as the PBC technique [17] for the SSDL models. Supervised models were trained only with the specified n^l_t images from the corresponding training partition of the $D^b_{i,t,\mathrm{CR}}$ target data subset as D^l_t. The SSDL models also used the remaining training images in $D^b_{i,t,\mathrm{CR}}$ as unlabelled data D^u_t.

Data augmentation was implemented for the training dataset as random flips and rotations through the FastAI library, for both supervised and SSDL models. All models were trained for 50 epochs each, with early stopping to avoid overfitting. We used the G-Mean as a criterion for keeping the model from the epoch with the best score after training. A learning rate of 0.00002, a weight decay of 0.001 and a batch size of 10 images were used. The hyper-parameters for MixMatch were set as: $K = 2$ transformations, a sharpening temperature of $T = 0.25$, an alpha mix value of $\alpha = 0.75$ and unsupervised coefficient $\gamma = 200$, following the authors' recommendations in [11]. The G-Mean, F2-Score, traditional accuracy, recall, specificity, and precision were

[2] `https://towardsdatascience.com/a-fastai-pytorch-implementation-of-mixmatch-314bb30d0f99`

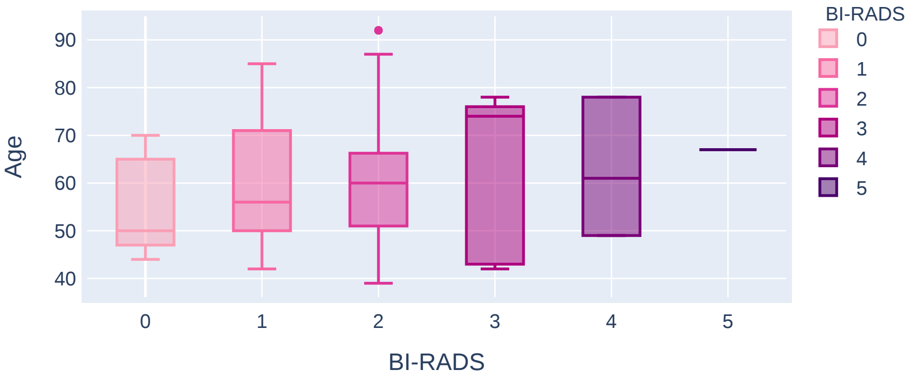


Fig. 7: Age distribution according to BI-RADS categories for patients in $D^l_{t,\mathrm{CR}}$

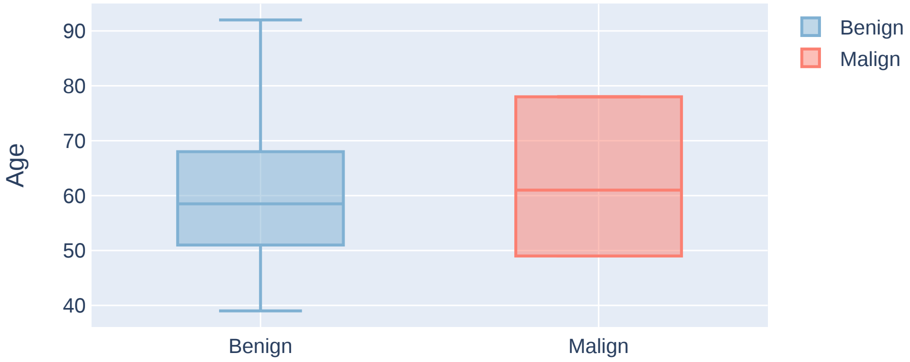
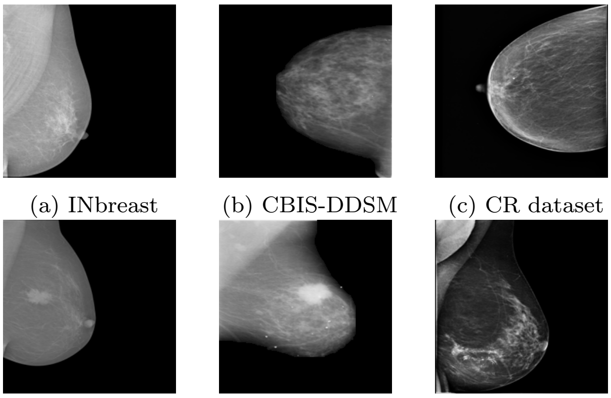


Fig. 8: Age distribution according to binary categories for patients in $D^b_{t,\mathrm{CR}}$



(a) INbreast (b) CBIS-DDSM (c) CR dataset

(d) INbreast (e) CBIS-DDSM (f) CR dataset

Fig. 9: Examples of benign (top) and malign (bottom) mammogram images from each dataset

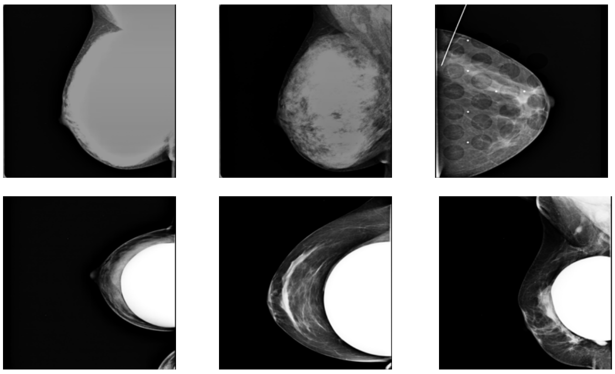


Fig. 10: Examples of images from original CR data discarded due to image quality (top) or patients with breast implants (bottom)

evaluated for each model, using the test data from their respective $D^b_{i,t,\mathrm{CR}}$. Results from these metrics were then reported as averages across the 10 target data subsets.

The dissimilarities between the complete source datasets $D^l_{s,\mathrm{DDSM}}$ and $D^l_{s,\mathrm{IN}}$, and the binary-labelled target dataset $D^b_{t,\mathrm{CR}}$ were evaluated following the approach presented in [15]. The cosine distance d_C was chosen as the dissimilarity measure, given its reported behavior in [15]. This was evaluated in the feature space of a generic Wide-ResNet model pre-trained on ImageNet, with the cosine distance calculated between the distributions of two datasets on each feature of the feature space and then summed [15]. We used 10 randomly selected batches of 40 observations to calculate the feature distribution distances, as suggested in [15].

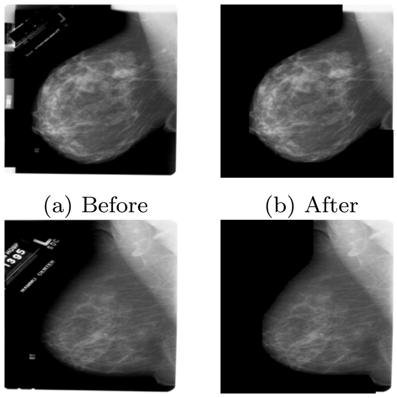


Fig. 11: Examples of images with background noise from CBIS-DDSM dataset, before and after being preprocessed

Table 2: Classification performance for models of Configuration **S+No-FT**, using the VGG-19 architecture

Metric	INbreast models		CBIS-DDSM models	
	$\bar{x}$	s	$\bar{x}$	s
G-Mean	0.3773	0.1043	0.3476	0.2534
F2-Score	0.1882	0.0625	0.1347	0.1148
Accuracy	0.2183	0.0602	0.7379	0.0678
Recall	0.7667	0.2509	0.2333	0.1876
Specificity	0.1901	0.0558	0.7639	0.0707
Precision	0.0470	0.0160	0.0517	0.0467

Table 3: Classification performance for models of Configuration **SSDL**, using the VGG-19 architecture

Metric	$n_t^l = 20$		$n_t^l = 40$		$n_t^l = 60$	
	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
G-Mean	0.4798	0.1936	0.5720	0.1257	0.6413	0.0929
F2-Score	0.2169	0.1194	0.2683	0.1168	0.3038	0.0889
Accuracy	0.5786	0.2212	0.6482	0.2172	0.6869	0.1412
Recall	0.5167	0.2687	0.5750	0.2648	0.6333	0.2297
Specificity	0.5815	0.2404	0.6518	0.2354	0.6904	0.1544
Precision	0.1189	0.1551	0.1079	0.0754	0.1096	0.0491

4 Results and Discussion

The results of each of the described experimental configurations are presented in tables 4, 2, 3, 5, and 6, as the mean and standard deviation of the corresponding classification metrics, evaluated across each of the 10 random data subsets of the target dataset. Results are also presented accordingly to the number of n_t^l that were used for training (Configs. **SSDL+FT**, **S+FT** and **SSDL**).

The classification performance on the target dataset of source-trained-only models appears to be rather poor, with no clear advantages between the source datasets, as seen in Table 2. The low average G-Mean values yielded by models trained on each of the source datasets show a deficient ability to correctly discriminate between both classes. This situation is confirmed by the yielded average recall and specificity values, which show a clear imbalance of the discrimination accuracy for each class. Low average F2-Score values also reinforce this conclusion, showing a relatively high number of FP in proportion to true positives (TP) predictions. The "accuracy paradox" can also be seen in the yielded average accuracy scores of Table 2. Models trained on $D_{s,\mathrm{IN}}^l$ scored notably lower accuracy values in comparison to models trained on $D_{s,\mathrm{DDSM}}^l$. However, further analysis suggests that the higher accuracy scores of the latter models were due to their relatively high specificity scores. This shows a clear bias on the accuracy scores for the majority class (negative cases).

Table 3 shows the classification performance results of models trained with SSDL on the target dataset, without domain adaptation from a source mammography dataset. Considerably high standard deviations are observed for the majority of the results. Despite this, the average values of both G-Mean and F2-Score show steady improvements as the number of n_t^l increases. It is only logical that these models are able to make a better use of an increased number of labelled observations for training. This is mainly due to the fact that they do not possess previous domain-knowledge from a source dataset.

Significant improvements can be perceived in the classification performance of the source-trained models after fine-tuning on the target dataset, as depicted by tables 5 and 6. Wilcoxon signed-rank tests were applied to these results in order to identify statistically significant (p-values < 0.05) differences between the performance of the models fine-tuned either in a supervised manner or with the SSDL method. Therefore, the null hypothesis is defined as that there is no statistically significant difference of using semi-supervised learning against using conventional supervised learning. The alternative hypothesis, refers to the statistically significant difference between using semi-supervised learning against using supervised learning. Table 5 shows the results of the models first trained on $D_{s,\mathrm{IN}}^l$ and then fine-tuned on the target dataset. The results with other architectures are depicted in the supplementary material. Models fine-tuned with SSDL generally yielded moderately better average G-Mean and F2-Score results in comparison to models fine-tuned using a supervised manner. This happens specially when using a reduced number of labelled observations for training ($n_t^l = 20, 40$), as the perceived gains decrease with a higher value of n_t^l. With more labels, the results tend to reveal less statistical significance with p-values > 0.05. Therefore we reject the previously stated null hypothesis, when few labels are used ($n_t^l = 20, 40$).

When comparing the models performance of the configurations **SSDL**, and **SSDL+FT**, described in tables 3, 5 and 6, we can see two different scalability trends, with respect to n_t^l. The **SSDL** configuration (with no fine-tuning), yields considerably lower performance scores, when compared to the **SSDL+FT** configuration. However, it scales better, when n_t^l increases. This suggests that the **SSDL+FT** configuration, with initial knowledge on the target task (mammogram classification), is less benefited when the number of labels grows.

The results shown in Table 6 correspond to the models that were first trained on $D^l_{s,\mathrm{DDSM}}$ and then fine-tuned on the target dataset. Considerably higher average G-Mean and F2-Score values were yielded by models fine-tuned with SSDL. They show statistical significance when employing lower amounts of labelled observations ($n_t^l = 20, 40$), specially for the models that used the VGG19 architecture. For these models, the ones that were fine-tuned in a supervised fashion scored higher average specificity values. However, by observing their respective average recall values it is clear that their rate of correct predictions is unbalanced for both classes. These models appear to be biased to the majority class. However, the models with SSDL can be considered to be less biased, according to the yielded results. Their average recall and specificity show a more stable behavior. Models with supervised fine-tuning also achieved generally higher average accuracy values, when compared to the no fine-tuned models.

In summary, models that were subject to domain adaptation from a source mammography dataset showed improved classification performance results in comparison to the other experimental configurations tested in this work. However, the choice of source dataset and deep learning model architecture are shown to be important factors in the yielded results. Models that used the CBIS-DDSM as source dataset showed better overall results, with more evident trends and noticeable improvements by the use of SSDL. Models that used the INbreast as source dataset scored relatively worse results, with no significant differences between the performance of supervised and SSDL models. Additionally, the performance of supervised models does not change significantly across the different number of labelled observations tested. These models achieved seemingly converging G-Mean values with fairly balanced recall and specificity values from a lower number of n_t^l. This was observed on all tested model architectures.

Regarding the poor performance of configuration **S+No-FT**, we found that the measurement of the DeDiMs can be an useful warning of choosing one unlabelled data source over another. The dissimilarity between $D^l_{s,\mathrm{IN}}$ and $D^b_{t,\mathrm{CR}}$ was measured as $\mathbf{31.10} \pm 1.56$, while for the dissimilarity between $D^l_{s,\mathrm{DDSM}}$ and $D^b_{t,\mathrm{CR}}$ was $\mathbf{26.21} \pm 2.31$, both results with p-values < 0.05. These results indicate that the feature distributions (using a generic ImageNet pre-trained model) between both source datasets and the target dataset are significantly different. This can explain the poor results of Configuration **S+No-FT** as a high dissimilarity is accurately suggesting that some sort of domain adaption is needed. At the same time, a lower dissimilarity between $D^l_{s,\mathrm{DDSM}}$ and $D^b_{t,\mathrm{CR}}$ might indicate that the former could be better suited to be used as a source dataset, as seen in the yielded performance behavior for both datasets in tables 5 and 6. The reasons behind a higher dissimilarity between two datasets need to be explored further.

Table 4 summarizes the performance of the models with the lowest number of labels. The average G-Mean scores are shown for models fine-tuned with the lowest number of labelled observations. The results in Table 4 show how the model architecture constitutes an important factor in the yielded performance of the models. As seen previously, SSDL models show better performance in comparison to supervised ones. However, the improved gains are stronger for the more complex models (i.e. architectures with more trainable parameters).

Overall, SSDL models without domain adaptation show significantly lower performance than models with domain adaptation either supervised or with SSDL (Configs. **S+FT** and **SSDL+FT**). Low average precision and F2-Score values are observed for models of all experimental configurations. As it was mentioned, for a binary classification task, this implies a considerably high number of false positives in relation to the number of true positives. Nonetheless, it must be taken into account that the target dataset suffers from extreme class imbalance. This causes the calculation of the precision to be highly sensitive to the number of false positives.

Table 4: Summary of G-Mean scores for models of Configs. **SSDL+FT** and **S+FT**, using $n_t^l = 20$ labelled observations. The corresponding number of trainable parameters for the PyTorch-implementation of each architecture is also shown

Model Architecture	INbreast				CBIS-DDSM				Trainable Parameters
	SSDL		Supervised		SSDL		Supervised		
	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	
VGG19_bn	**0.6764**	0.1084	0.6682	0.0770	**0.7313**	0.0742	0.5163	0.2826	139.5 Million
ResNet-152	**0.6774**	0.1167	0.6767	0.1021	**0.6575**	0.1075	0.5857	0.0598	58.1 Million
EfficientNet-b0	**0.6512**	0.1081	0.6393	0.0603	**0.5982**	0.0753	0.5824	0.0489	4 Million

Table 5: Results of Configurations **SSDL+FT** and **S+FT**, using **INbreast** as source dataset with the VGG-19 architecture

n_t^l	Metric	SSDL		Supervised	
		$\bar{x}$	s	$\bar{x}$	s
20	G-Mean	**0.6764**	0.1084	0.6682	0.0770
	F2-Score	**0.3506**	0.0973	0.3133	0.0673
	Accuracy	**0.7812**	0.0727	0.7014	0.0793
	Recall	0.5917	0.1687	**0.6500**	0.1748
	Specificity*	**0.7907**	0.0755	0.7048	0.0876
	Precision	**0.1436**	0.0636	0.1074	0.0335
40	G-Mean	**0.7017**	0.0932	0.6656	0.0877
	F2-Score	**0.3650**	0.0899	0.3484	0.1112
	Accuracy	**0.7742**	0.0659	0.7224	0.1590
	Recall	**0.6417**	0.1715	0.6417	0.2081
	Specificity	**0.7810**	0.0693	0.7262	0.1721
	Precision	0.1380	0.0373	**0.1837**	0.1708
60	G-Mean	**0.6689**	0.0957	0.6604	0.0876
	F2-Score	0.3278	0.0958	**0.3415**	0.1116
	Accuracy	0.7211	0.1169	**0.7432**	0.1374
	Recall	**0.6250**	0.1318	0.6000	0.1748
	Specificity	0.7267	0.1230	**0.7510**	0.1466
	Precision	0.1226	0.0565	**0.1822**	0.1704

Table 6: Results of Configurations **SSDL+FT** and **S+FT**, using **CBIS-DDSM** as source dataset with the VGG-19 architecture

n_t^l	Metric	SSDL		Supervised	
		$\bar{x}$	s	$\bar{x}$	s
20	G-Mean*	**0.7313**	0.0742	0.5163	0.2826
	F2-Score	**0.3910**	0.0909	0.2892	0.1797
	Accuracy*	0.7455	0.1115	**0.8333**	0.0710
	Recall*	**0.7333**	0.1459	0.3917	0.2292
	Specificity*	0.7460	0.1201	**0.8554**	0.0709
	Precision	0.1480	0.0551	**0.1602**	0.1289
40	G-Mean*	**0.7264**	0.0909	0.5743	0.2308
	F2-Score*	**0.3917**	0.1124	0.3070	0.1597
	Accuracy*	0.7588	0.1041	**0.8286**	0.0476
	Recall*	**0.7083**	0.1632	0.4417	0.2189
	Specificity*	0.7612	0.1110	**0.8482**	0.0453
	Precision	**0.1520**	0.0630	0.1458	0.0899
60	G-Mean	**0.7142**	0.0717	0.6466	0.1462
	F2-Score	**0.3779**	0.1001	0.3436	0.1506
	Accuracy*	0.7197	0.1445	**0.8132**	0.0723
	Recall	**0.7333**	0.1459	0.5333	0.2297
	Specificity*	0.7190	0.1559	**0.8271**	0.0779
	Precision	0.1435	0.0623	**0.1539**	0.0834

5 Conclusions

In this work we discussed the impact of using target datasets with scarce labelled data for the implementation of deep learning models for detection of malign cases using mammogram images. As presented in [7], the determination and study of an appropriate dataset size is an open challenge. It is clear that under real-life conditions medical imaging implementation of deep learning systems is still challenging, namely due to problems with labelled data scarcity and class-imbalance.

To tackle these challenges on the binary classification of mammograms, a combination of transfer learning from source datasets and semi-supervised learning to leverage unlabelled target data has been proposed and tested. In the experiments carried out in this work, it was found that this combination can achieve significant improvements on the classification performance of deep learning models. This surpasses the performance of models without transfer learning or without the use of unlabelled target data. The experiments depicted in this work also reveal the importance of using transfer learning from source datasets. Still, the highest yielded performance of the SSDL model with fine-tuning have a large room for improvement. Enforcing further supervision with small labelled datasets (pixel-wise labelling of the regions of interest), with other forms of weak or self-supervision [55] and/or domain adaptation [49], along with more complex data augmentation approaches as in [25], might improve the overall model performance. This must be done without raising too much the need of expensive labelling.

The target dataset used in this work for the evaluation of the models in the classification of mammograms is made available for other interested researchers. The dataset built for this work shows real-life conditions for the deployment of a deep learning based CAD system. Highly imbalanced data, along with the significant distribution mismatch with the source datasets are important and frequent aspects of real-world test data for medical imaging based CAD.

The dissimilarity between source and target datasets was found to be significant with the use of the DeDiMs

*Statistic significance (p-values < 0.05) for average differences between results of SSDL and supervised models measures. This was shown to be the case even though images from datasets can be considered as semantically and visually similar. Related to this, the choice of the source dataset was found to be an important factor in the yielded improvements in the performance of models, as well as model complexity. The measured DeDiMs can be considered a generic and simple data quality metric, similar to the data heterogeneity metric proposed in [42]. In general, specific data quality metrics for deep learning models to solve medical imaging challenges is still a very under-developed topic in the literature. We plan to contribute in such data-oriented metric development in the medical imaging analysis field in the future. In future work, we aim to explore computationally efficient and informative data quality metrics for deep learning architectures. Feature space based quality metrics can be explored in more recent deep learning architectures such as transformers [39]. Additionally further evaluation of model-oriented properties of deep learning models such as robustness and predictive uncertainty, as recommended in [45], is also a future work-line to develop.

Acknowledgements This research was partially supported by a machine allocation on Kabré supercomputer at the Costa Rica National High Technology Center. This work is partially supported by by the Ministry of Science, Innovation and Universities of Spain under grant number RTI2018-094645-B-I00, project name Automated detection with low cost hardware of unusual activities in video sequences. It is also partially supported by the Autonomous Government of Andalusia (Spain) under project UMA18-FEDERJA-084, project name Anomalous behaviour agent detection by deep learning in low cost video surveillance intelligent systems. All of them include funds from the European Regional Development Fund (ERDF). The authors thankfully acknowledge the computer resources, technical expertise and assistance provided by the SCBI (Supercomputing and Bioinformatics) center of the University of Malaga. The authors also acknowledge the funding from the Instituto de Investigación Biomédica de Málaga – IBIMA and the Universidad de Málaga. Lastly, the authors acknowledge the contribution and support from Luis Chavarría as a member of the executive board of the private clinic Imágenes Médicas Dr. Chavarría Estrada, for granting permission to the access and usage of mammogram images from the clinic.

Compliance with ethical standards

Conflicts of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Research involving human participants The authors declare that both this study and the data gathering process of these images comply with the Declaration of Helsinki for medical research, as this study was entirely observational, did not directly involve any human subjects. Furthermore, no identifiable human material nor data were used. The data was already acquired during regular clinical practice. Additionally, data from the INbreast and CBIS-DDSM datasets was gathered from previous third-party studies and made available for academic research.

The authors count with explicit permission from the Chavarría Clinic executive board for the usage of their images for academic purposes. Additionally, since the data was collected from patients of the clinic in 2020, it was already gathered by the beginning of this study and was ultimately provided to the research team.

Author Biographies

Saúl Calderón-Ramírez received his B.Sc. in computer science and his M.Sc. in electrical engineering from the University of Costa Rica, Costa Rica. He is currently a Ph.D student at De Montfort University, U.K.

Diego Murillo-Hernández received the B.Sc. degree in computer engineering from the Costa Rica Institute of Technology, Costa Rica, in 2021. He is currently working in the private industry as a data scientist.

Kevin Rojas-Salazar currently an undergraduate student of Computer Engineering at the Costa Rica Institute of Technology, Costa Rica. He works as a research assistant of deep learning applications.

David Elizondo (Senior Member, IEEE) received his P h.D. in computer science, U.of Strasbourg, France. He is currently a professor at De Montfort University, U.K.

Shengxiang Yang (Senior Member, IEEE) received the Ph.D. degree from Northeastern University, China. He is currently Director of the Centre for Computational Intelligence, De Montfort University, U.K.

Armaghan Moemeni received the Ph.D. degree in computer science from De Montfort University. She is currently an Assistant Professor in computer science with the University of Nottingham.

Miguel A. Molina-Cabello received the Ph.D. degree in computer engineering from the University of Málaga. He works at the University of Málaga, where he holds a teaching and researching position.

References

1. Abdelhafiz, D., Yang, C., Ammar, R., Nabavi, S.: Deep convolutional neural networks for mammography: advances, challenges and applications. BMC bioinformatics **20**(11), 1–20 (2019)
2. Akosa, J.: Predictive accuracy: A misleading performance measure for highly imbalanced data. In: Proceedings of the SAS Global Forum, vol. 12 (2017)
3. Alfaro, E., Fonseca, X.B., Albornoz, E.M., Martínez, C.E., Ramrez, S.C.: A brief analysis of u-net and mask r-cnn for skin lesion segmentation. In: 2019 IEEE International Work Conference on Bioinspired Intelligence (IWOBI), pp. 000123–000126. IEEE (2019)
4. Alkhaleefah, M., Ma, S.C., Chang, Y.L., Huang, B., Chittem, P.K., Achhannagari, V.P.: Double-shot transfer learning for breast cancer classification from x-ray images. Applied Sciences **10**(11), 3999 (2020)
5. American Cancer Society: Breast cancer facts & figures 2019-2020. American Cancer Society, Inc. pp. 1–44 (2019)
6. Bakalo, R., Goldberger, J., Ben-Ari, R.: Weakly and semi supervised detection in medical imaging via deep dual branch net. Neurocomputing **421**, 15–25 (2021). DOI https://doi.org/10.1016/j.neucom.2020.09.037
7. Balki, I., Amirabadi, A., Levman, J., Martel, A.L., Emersic, Z., Meden, B., Garcia-Pedrero, A., Ramirez, S.C., Kong, D., Moody, A.R., et al.: Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. Canadian Association of Radiologists Journal **70**(4), 344–353 (2019)
8. Beeravolu, A.R., Azam, S., Jonkman, M., Shanmugam, B., Kannoorpatti, K., Anwar, A.: Preprocessing of breast cancer images to create datasets for deep-cnn. IEEE Access (2021)
9. Bermudez, A., Calderon-Ramirez, S., Thang, T., Tyrrell, P., Moemeni, A., Yang, S., Torrents-Barrena, J.: A first glance to the quality assessment of dental photostimulable phosphor plates with deep learning. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–6. IEEE (2020)
10. Berrar, D., Flach, P.: Caveats and pitfalls of roc analysis in clinical microarray research (and how to avoid them). Briefings in bioinformatics **13**(1), 83–97 (2012)
11. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. In: Advances in Neural Information Processing Systems, pp. 5049–5059 (2019)
12. Calderon-Ramirez, S., Fallas, F., Zumbado, M., Tyrrell, P.N., Stark, H., Emersic, Z., Meden, B., Solis, M.: Assessing the impact of the deceived non local means filter as a preprocessing stage in a convolutional neural network based approach for age estimation using digital hand x-ray images. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 1752–1756. IEEE (2018)
13. Calderon-Ramirez, S., Giri, R., Yang, S., Moemeni, A., Umana, M., Elizondo, D., Torrents-Barrena, J., Molina-Cabello, M.A.: Dealing with scarce labelled data: Semi-supervised deep learning with mix match for covid-19 detection using chest x-ray images. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 5294–5301. IEEE (2021)
14. Calderon-Ramirez, S., Murillo-Hernandez, D., Rojas-Salazar, K., Calvo-Valverde, L.A., Yang, S., Moemeni, A., Elizondo, D., Lopez-Rubio, E., Molina-Cabello, M.: Improving uncertainty estimations for mammogram classification using semi-supervised learning. In: Institute of Electrical and Electronics Engineers (2021)
15. Calderon-Ramirez, S., Oala, L.: More than meets the eye: Semi-supervised learning under non-iid data. arXiv preprint arXiv:2104.10223 (2021)
16. Calderon-Ramirez, S., Oala, L., Torrents-Barrena, J., Yang, S., Moemeni, A., Samek, W., Molina-Cabello, M.A.: Mixmood: A systematic approach to class distribution mismatch in semi-supervised learning using deep dataset dissimilarity measures. arXiv preprint arXiv:2006.07767 (2020)
17. Calderon-Ramirez, S., Shengxiang-Yang, Moemeni, A., Elizondo, D., Colreavy-Donnelly, S., Chavarria-Estrada, L.F., Molina-Cabello, M.A.: Correcting data imbalance for semi-supervised covid-19 detection using x-ray chest images (2020)
18. Calvo, I., Calderon-Ramirez, S., Torrents-Barrena, J., Muñoz, E., Puig, D.: Assessing the impact of a preprocessing stage on deep learning architectures for breast tumor multi-class classification with histopathological images. In: Latin American High Performance Computing Conference, pp. 262–275. Springer (2019)
19. Castro, E., Cardoso, J.S., Pereira, J.C.: Elastic deformations for data augmentation in breast cancer mass detection. In: 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), pp. 230–234. IEEE (2018)
20. Cheplygina, V., de Bruijne, M., Pluim, J.P.: Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical Image Analysis **54**, 280–296 (2019). DOI https://doi.org/10.1016/j.media.2019.03.009
21. Chicco, D., Jurman, G.: The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC genomics **21**(1), 1–13 (2020)
22. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al.: The cancer imaging archive (tcia): Maintaining and operating a public information repository. Journal of Digital Imaging **26**(6), 1045–1057 (2013). DOI https://doi.org/10.1007/s10278-013-9622-7
23. Devarriya, D., Gulati, C., Mansharamani, V., Sakalle, A., Bhardwaj, A.: Unbalanced breast cancer data classification using novel fitness functions in genetic programming. Expert Systems with Applications **140**, 112866 (2020)
24. Dhungel, N., Carneiro, G., Bradley, A.P.: Deep learning and structured prediction for the segmentation of mass in mammograms. In: International Conference on Medical image computing and computer-assisted intervention, pp. 605–612. Springer (2015)
25. Domingues, I., Abreu, P.H., Santos, J.: Bi-rads classification of breast cancer: a new pre-processing pipeline for deep models training. In: 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 1378–1382. IEEE (2018)
26. Falconí, L., Pérez, M., Aguilar, W., Conci, A.: Transfer learning and fine tuning in mammogram bi-rads classification. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), pp. 475–480. IEEE (2020)
27. Forman, G., Scholz, M.: Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. Acm Sigkdd Explorations Newsletter **12**(1), 49–57 (2010)
28. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
29. Hamidinekoo, A., Denton, E., Rampun, A., Honnor, K., Zwiggelaar, R.: Deep learning in mammography and

breast histology, an overview and future trends. Medical image analysis **47**, 45–67 (2018)

30. Hand, D.J.: Measuring classifier performance: a coherent alternative to the area under the roc curve. Machine learning **77**(1), 103–123 (2009)
31. Heath, M., Bowyer, K., Kopans, D., Kegelmeyer, P., Moore, R., Chang, K., Munishkumaran, S.: Current status of the digital database for screening mammography. In: Digital mammography, pp. 457–460. Springer (1998). DOI https://doi.org/10.1007/978-94-011-5318-8_75
32. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations (2019)
33. Johnson, J.M., Khoshgoftaar, T.M.: Deep learning and thresholding with class-imbalanced big data. In: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), pp. 755–762. IEEE (2019)
34. Korkinof, D., Rijken, T., O'Neill, M., Yearsley, J., Harvey, H., Glocker, B.: High-resolution mammogram synthesis using progressive generative adversarial networks (2019)
35. Kubat, M., Matwin, S., et al.: Addressing the curse of imbalanced training sets: one-sided selection. In: Icml, vol. 97, pp. 179–186. Citeseer (1997)
36. Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L.: A curated mammography data set for use in computer-aided detection and diagnosis research. Scientific Data **4**(1) (2017). DOI https://doi.org/10.1038/sdata.2017.177
37. Lee, R.S., Gimenez, F., Hoogi, A., Rubin, D.: Curated breast imaging subset of ddsm. The cancer imaging archive (2016). DOI https://doi.org/10.7937/K9/TCIA.2016.7O02S9CY
38. Lévy, D., Jain, A.: Breast mass classification from mammograms using deep convolutional neural networks. arXiv preprint arXiv:1612.00542 (2016)
39. Li, G., Xu, S., Liu, X., Li, L., Wang, C.: Jersey number recognition with semi-supervised spatial transformer network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1783–1790 (2018)
40. Li, H., Chen, D., Nailon, W.H., Davies, M.E., Laurenson, D.I.: Signed laplacian deep learning with adversarial augmentation for improved mammography diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 486–494. Springer (2019)
41. Maratea, A., Petrosino, A., Manzo, M.: Adjusted f-measure and kernel scaling for imbalanced data learning. Information Sciences **257**, 331–341 (2014)
42. Mendez, M., Calderon, S., Tyrrell, P.N.: Using cluster analysis to assess the impact of dataset heterogeneity on deep convolutional network accuracy: A first glance. In: Latin American High Performance Computing Conference, pp. 307–319. Springer (2019)
43. Moreira, I., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M., Cardoso, J.: Inbreast: Toward a full-field digital mammographic database. Academic radiology **19**, 236–48 (2011). DOI 10.1016/j.acra.2011.09.014
44. Mustra, M., Grgic, M., Rangayyan, R.M.: Review of recent advances in segmentation of the breast boundary and the pectoral muscle in mammograms. Medical & biological engineering & computing **54**(7), 1003–1024 (2016)
45. Oala, L., Fehr, J., Gilli, L., Balachandran, P., Leite, A.W., Calderon-Ramirez, S., Li, D.X., Nobis, G., Alvarado, E.A.M., Jaramillo-Gutierrez, G., et al.: Ml4h auditing: From paper to practice. In: Machine Learning for Health, pp. 280–317. PMLR (2020)
46. Pardamean, B., Cenggoro, T.W., Rahutomo, R., Budiarto, A., Karuppiah, E.K.: Transfer learning from chest x-ray pre-trained convolutional neural network for learning mammogram data. Procedia Computer Science **135**, 400–407 (2018). DOI https://doi.org/10.1016/j.procs.2018.08.190. The 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018) : Empowering Smart Technology in Digital Era for a Better Life
47. Powers, D.M.: What the f-measure doesn't measure: Features, flaws, fallacies and fixes. arXiv preprint arXiv:1503.06410 (2015)
48. Shen, L., Margolies, L.R., Rothstein, J.H., Fluder, E., McBride, R., Sieh, W.: Deep learning to improve breast cancer detection on screening mammography. Scientific reports **9**(1), 1–12 (2019)
49. Shen, R., Yao, J., Yan, K., Tian, K., Jiang, C., Zhou, K.: Unsupervised domain adaptation with adversarial learning for mass detection in mammogram. Neurocomputing **393**, 27–37 (2020)
50. Shi, Q., Zhang, H.: Fault diagnosis of an autonomous vehicle with an improved svm algorithm subject to unbalanced datasets. IEEE Transactions on Industrial Electronics (2020)
51. Sokolova, M., Japkowicz, N., Szpakowicz, S.: Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In: Australasian joint conference on artificial intelligence, pp. 1015–1021. Springer (2006)
52. Sun, L., Wen, J., Wang, J., Zhao, Y., Xu, Y.: Classification of mammography based on semi-supervised learning. In: 2020 IEEE International Conference on Progress in Informatics and Computing (PIC), pp. 104–111 (2020). DOI 10.1109/PIC50277.2020.9350835
53. Sun, W., Tseng, T.L.B., Zhang, J., Qian, W.: Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. Computerized Medical Imaging and Graphics **57**, 4–9 (2017). DOI https://doi.org/10.1016/j.compmedimag.2016.07.004. Recent Developments in Machine Learning for Medical Imaging Applications
54. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J.: On the necessity of fine-tuned convolutional neural networks for medical imaging. In: Deep Learning and Convolutional Neural Networks for Medical Image Computing, pp. 181–193. Springer (2017). DOI https://doi.org/10.1007/978-3-319-42999-1_11
55. Tardy, M., Mateus, D.: Looking for abnormalities in mammograms with self-and weakly supervised reconstruction. IEEE Transactions on Medical Imaging (2021)
56. Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., Kennedy, P.J.: Training deep neural networks on imbalanced data sets. In: 2016 International Joint Conference on Neural Networks (IJCNN), pp. 4368–4374 (2016). DOI 10.1109/IJCNN.2016.7727770
57. Wild, C., Weiderpass, E., Stewart, B.: World cancer report: cancer research for cancer prevention. Lyon: International Agency for Research on Cancer (2020)
58. Wu, E., Wu, K., Lotter, W.: Synthesizing lesions using contextual gans improves breast cancer classification on mammograms (2020)
59. Zheng, Q., Yang, M., Yang, J., Zhang, Q., Zhang, X.: Improvement of generalization ability of deep cnn via implicit regularization in two-stage training process. IEEE Access **6**, 15844–15869 (2018)