Improving Uncertainty Estimations for Mammogram Classification using Semi-Supervised Learning

Saúl Calderón-Ramírez *†, Diego Murillo-Hernández †, Kevin Rojas-Salazar †
Luis-Alexander Calvo-Valverde †, Shengxiang Yang *, Armaghan Moemeni ‡
David Elizondo *, Ezequiel López-Rubio §, Miguel A. Molina-Cabello §

* Centre for Computational Intelligence (CCI), De Montfort University, United Kingdom
†PARMA Research Group, Instituto Tecnológico de Costa Rica, Costa Rica
‡School of Computer Science, University of Nottingham, United Kingdom
§Department of Computer Languages and Computer Science. University of Málaga, Spain
Instituto de Investigación Biomédica de Málaga (IBIMA), Spain

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

***Abstract*—Computer aided diagnosis for mammogram images have seen positive results through the usage of deep learning architectures. However, limited sample sizes for the target datasets might prevent the usage of a deep learning model under real world scenarios. The usage of unlabeled data to improve the accuracy of the model can be an approach to tackle the lack of target data. Moreover, important model attributes for the medical domain as model uncertainty might be improved through the usage of unlabeled data. Therefore, in this work we explore the impact of using unlabeled data through the implementation of a recent approach known as MixMatch, for mammogram images. We evaluate the improvement on accuracy and uncertainty of the model using popular and simple approaches to estimate uncertainty. For this aim, we propose the usage of the uncertainty balanced accuracy metric.**

***Index Terms*—Uncertainty Estimation, Breast Cancer, Mammogram, Semi-Supervised Deep Learning, MixMatch**

I. Introduction

According to the International Agency for Research on Cancer (IARC), breast cancer remains as the leading cause of cancer death in women worldwide [1]. In 2019, for the US alone, it was estimated that around 268,600 women would be diagnosed with invasive breast cancer, and out of them 41,760 were expected to die [2].

Such aggravating statistics contribute to emphasize the importance of an early diagnosis, as this allows a rapid response which highly increases the efficacy in the treatment of the disease [2]. Nevertheless, this is not always possible, as regions with scarce medical professionals and low economic resources lack the ability to appropriately carry out preemptive and uniform mammogram screening tests to their populations.

Over the last decade deep learning has proven to be a viable approach for this problem, leveraging effective computer aided diagnosis tools. Deep learning has been applied for the detection and classification of diseases by the analysis of medical images of patients suffering from conditions like cancer [3] and more recently, COVID-19 [4].

Nevertheless, despite its proven success it is well known that deep learning techniques still face great disadvantages. One of them being the vast amount of correctly labeled data needed in order to train models by classic supervised techniques [3]. This represents an especially difficult challenge in the medical domain, as highly trained professionals are required to spend long periods of time dedicated to carefully labeling the data.

Semi-supervised and self-supervised deep learning techniques have recently gained popularity, as these approaches take advantage of cheaper unlabeled data to train deep learning models [4]–[6]. For medical imaging, still the advantage of semi-supervised learning can be considered under-reported, with authors often focusing only on accuracy gains [4].

Model uncertainty estimation is of special interest in the medical domain, as it is related to model explainability. Enforcing model reliability and accountability is relevant for health professionals. For this reason, it is of great interest to assess the impact that semi-supervised learning can have in model uncertainty estimation.

In this work we focus on the measurement and improvement of uncertainty estimations for deep learning models trained to identify breast cancer on whole mammogram images. Specifically, we aim to improve uncertainty estimations by using unlabeled data under regimes with very limited number of labeled observations for training.

II. State of the Art

A. Uncertainty Estimation

Uncertainty estimation relates to model explainability as it consists on the capability of the model to estimate how reliable is its output. This can be of special interest in the medical domain, where the decisions taken by a model can have direct

consequences in the life of patients. As such, this subject has been widely studied in the literature [7]–[10].

In this work we focus on uncertainty estimation methods that are practical and straightforward to implement. Specifically, the Softmax and Monte Carlo Dropout (MCD) approaches were tested. These solutions do not require model architectures to be modified, nor to re-train models.

For classification related tasks, the usage of a Softmax activation function in the output layer of a deep learning model can serve as a basic method for uncertainty estimation [11]. Given a model f and the array of outputs $\boldsymbol{y}_j = f_{\boldsymbol{\theta}}(\boldsymbol{x}_j)$ and model parameters $\boldsymbol{\theta}$, for a given input observation $\boldsymbol{x}_j$, the Softmax function estimates a density function $\boldsymbol{p}$ of the form:

$$p_i = \frac{\exp(y_{i,j})}{\sum_k \exp(y_{k,j})} \tag{1}$$

Where unit i of the Softmax output for a given input observation x_j can be interpreted as the model "certainty" for class i. The entropy over $\boldsymbol{p}$ can also be used as a less naive estimation. However, the Softmax method alone can lead to poor representations of model uncertainty due to typical overconfidence in neural networks' predictions [12], [13].

The MCD approach aims at having more robust estimations while still being simple to implement [12], when compared to the usage of Softmax for uncertainty estimation. MCD is based on a Bayesian interpretation of the model's parameters. It estimates the distribution of the model outputs through M stochastic evaluations of the model $\boldsymbol{y}_{m,j} = f_{\boldsymbol{\theta'}_m}(\boldsymbol{x}_j)$ for a given input $\boldsymbol{x}_j$. The model parameters $\boldsymbol{\theta'}$ are randomly modified, usually with dropout procedure, with the objective of measuring the stability of the output for each evaluation m. Then, for a given input $\boldsymbol{x}_j$ the dispersion of its output is calculated as:

$$\sigma^2_{model}(\boldsymbol{x}_j) = \frac{1}{M}\sum_{m=1}^{M}\sum_{k=1}^{K}(y_{m,j,k} - \bar{y}_{j,k})^2 . \tag{2}$$

Where $y_{m,j}$ refers to the output of each perturbed model for observation $\boldsymbol{x}_j$ and $\bar{y}_{j,k}$ is the model output average for the M model evaluations. Dispersion estimation can be accumulated for all the output units $k = 1, ..., K$ or alternatively only the unit with the highest output can be used.

B. Semi-Supervised Deep Learning

Semi-supervised learning techniques have proved to be viable solutions to the shortage of labeled data [5]. This has led to such techniques to become especially appealing for applications in contexts like the medical domain, where the lack of publicly available and properly labeled training data continues to be a great issue [14]. Different semi-supervised methods for deep learning architectures have been proposed recently. In [5] a detailed survey on semi-supervised methods can be found, and in [15] a review on its usage for medical applications was developed. MixMatch is among the most recent and successful semi-supervised approaches. As presented by the authors in [16], this approach combines intensive data augmentation, pseudo-labelling, and unsupervised regularization.

More specifically, MixMatch starts by generating pseudo-labels from previously augmented unlabeled training data (with affine transformations, e.g. image flipping and mirroring), by averaging and subsequently sharpening the model "guesses" for the aforementioned augmented observations [16].

MixMatch uses both unlabeled and labeled datasets, $\tilde{S}_u = (X_u, \tilde{Y})$, where X_u correspond to the unlabeled observations and $\tilde{Y} = \{\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_{n_u}\}$ are their pseudo-labels, and $S_l = (X_l, Y_l)$, correspond to the labeled observations along with their respective labels.

Then, both labeled and unlabeled data sets $(S_l, \tilde{S}_u)$ are further augmented with the MixUp algorithm [17]. The technique creates new observations, generated by linearly interpolating already existing data. This process results in the augmented data sets S'_l and $\tilde{S}'_u$.

Finally, the resulting data is used to train a model by minimizing the following loss function [16]:

$$\begin{aligned}\mathcal{L}(S, \theta) = &\sum_{(x_i, y_i) \in S'_l} \mathcal{L}_l(\theta, x_i, y_i) + \\ &\gamma r(\tau) \sum_{(x_j, \tilde{y}_j) \in \tilde{S}'_u} \mathcal{L}_u(\theta, x_j, \tilde{y}_j)\end{aligned} \tag{3}$$

Where $S = (S'_l, \tilde{S}'_u)$, and θ refer to the model weights. $\mathcal{L}_l$ is a supervised loss term, implemented with a cross-entropy loss, and $\mathcal{L}_u$ is an unlabeled loss term, implemented with an Euclidean distance. γ serves as a regularization coefficient for the unlabeled loss term, along with the ramp-up function $r(\tau)$, which increases its value as the epochs τ increase. Our implementation of $r(\tau)$ was set to $\tau/3000$.

It is important to note that unlabeled data also affects the term $\mathcal{L}_l$, as both S_l and $\tilde{S}_u$ are used to generate $\tilde{S}'_u$ as well as S'_l via the MixUp approach [16].

Recent studies have linked the usage of unlabeled data to improvements in the uncertainty of deep learning models. The authors in [18] implemented a regularization approach based on the use of unlabeled data and yielded better uncertainty estimations for structured data. Furthermore, concepts implemented in semi-supervised learning techniques, specifically on MixMatch [16], have been known to improve the uncertainty estimation of models. These include: consistency regularization of self-supervised learning [19] and the MixUp data augmentation algorithm [20], both implemented in the MixMatch approach.

C. Deep Learning for Mammogram Analysis

The analysis of mammograms for breast cancer detection constitutes a popular research topic in the literature [3], [21]. Works in this area can be roughly divided into plain mammogram image classification or in segmentation of specific anomalies (masses, microcalcifications, tissue deformations, etc.). In this work we focus on the former, as we aim to classify images of patients as malign (cancerous) or benign.

Professionals in radiology follow specific standards in the process of examining mammogram images. One of such is using the BI-RADS scale developed by the American College of Radiology [22], in order to report the results of a mammogram as precisely as possible. The BI-RADS scale is based on the level of suspicion of findings in an image. It ranges from 0 to 6 as: 0 for exam not conclusive, 1 for no findings, 2 for benign, 3 for probably benign, 4 for suspicious, 5 for high probability of malignancy and 6 for proved cancer (via biopsy).

Given this standard and how it is common for recent datasets to be labeled according to it, some works have experimented with multiclass classification, as in [23] and in [24]. However, most research seems to focus on binary classification (benign-malign), as BI-RADS classification can be considered prone to label noise, with a high abstraction level of the categories, making the problem particularly challenging [25].

Moreover, research in this area mainly concentrates on the classification of pre-annotated regions of interest, such as anomalies like masses or calcifications [3], [26]. For instance, in [27]–[29] convolutional neural networks for mass classification in mammograms have been implemented, achieving state of the art results [29].

A considerably smaller amount of research has focused on the analysis and classification of whole mammogram images. While this represents a task of great interest in the field of computer aided diagnosis, it is still faced with many challenges. One of them being that the size, number and type of anomalies present in a single image can vary greatly.

In order to tackle these problems, some authors have implemented solutions like the one explained in [21] based on Region Based Convolutional Neural Networks (R-CNN) and their variants (Fast R-CNN and Faster R-CNN). However, these approaches need great amounts of finely labeled data, almost on the level of pixel annotations [26], which are harder to obtain and not usually generated in the every-day work of radiologists.

The authors in [26] propose a similar approach, where deep learning models are trained to classify "patches" of the original mammogram images. More convolutional layers are then added to the trained models in order for them to be able to process full images and are further fine-tuned. This "end-to-end" approach is reported to achieve positive results.

Works presented by the authors in [23] and [24] have achieved substantial results in multiclass classification of whole mammograms. This while relying heavily on pre-processing and data augmentation techniques.

Research on Semi-supervised Deep Learning (SSDL) for mammogram classification is far more scarce. In [30] a semi-supervised model using transductive learning is implemented for the patch-wise classification of masses in two classes, benign and malign. The authors used the Digital Database for Screening Mammography database (DDSM). No study on the impact of the labeled dataset size or comparison with other methods was done.

More recently, in [31], a semi-supervised segmentation model for mammogram lesions was proposed. The model is based on a co-training approach for pseudo-label estimation using two models based on a support vector machine. No comparison with recent SSDL approaches was done, with no statistical analysis of the results of the other two approaches tested (region growing and watersheds).

In this work we test simple off-the-shelf classification models, to limit the interaction of additional aspects on the matter at hand in this work. To our knowledge, this work is the first to apply the MixMatch semi-supervised approach to this domain and to study the impact of its usage in uncertainty estimation. We focus on the classification of whole mammogram images with BI-RADS labels, as in real-world clinics this type of data is considerably easier to get. Implementing a model to a specific clinic with a particular mammogram sampling equipment and technicians can result in a better model performance than using other larger datasets with data acquired with different equipment and technicians. This problem is referred as domain adaptation [32], where in this case the domain might remain the same, but the distribution of the datasets might change. Finally, for the comparison of the tested uncertainty estimation approaches, we propose the usage of the uncertainty balanced accuracy as a metric.

III. Contributions

Contributions to the state of the art are made through this work in the following aspects:

- We explore the impact of semi-supervised deep learning approach known as MixMatch, on the accuracy and reliability of uncertainty estimations for whole mammogram binary classification, using common deep learning architectures.
- We evaluate and compare the performance of three different uncertainty estimation techniques for both supervised and semi-supervised models for the binary classification of whole mammogram images.
- We propose the uncertainty balanced accuracy to ease the comparison of uncertainty estimation approaches in real world usage scenarios.

IV. Proposed Method

This work proposes the improvement of uncertainty estimation for deep learning models through the usage of unlabeled data with the MixMatch approach (as depicted in section II-B). The study involves the binary classification of whole mammogram images.

Three simple and popular uncertainty estimation methods were tested as follows:

- The Softmax function as depicted in Equation 1 is used in this work for uncertainty estimation (also known as model calibration). The maximum Softmax value for the output layer of a model serves as uncertainty estimation u_j for a given input x_j. This was implemented as $u_j = \underset{i}{\operatorname{argmax}}\ p_i$, were p is the corresponding Softmax distribution output.

- The complete set of values for a Softmax output given an input x_j can also be used for uncertainty estimation. This is done by calculating the entropy over the corresponding output distribution p of Softmax. This was implemented as $u_j = -\frac{1}{\alpha}\sum_i p_i \log(p_i)$, were α is a normalization coefficient.
- MCD as described in Equation 2, uses the distribution from multiple evaluations of dropout-enabled models for the same input x_j. Either the entropy of such distribution or its second statistical moment (standard deviation) can be used. The latter was chosen and implemented as $u_j = \sigma_{model}(x_j)$.

A similar approach to the one in [33] was implemented to evaluate and compare the results for the tested uncertainty estimators. Authors in [33] proposed several quantitative metrics based on an approach known as "uncertainty confusion matrix" that consists on grouping uncertainty estimations for each of a model's predictions according to their "correctness" and "confidence". The correctness of a prediction can be attained by comparing it with its ground truth label, while the confidence can be estimated by comparing the prediction's uncertainty estimate with a threshold.

Thus, predictions can be classified into one of four groups: correct and certain (true certainty (TC)), incorrect and uncertain (true uncertainty (TU)), correct and uncertain (false uncertainty (FU)), and incorrect and certain (false certainty (FC)). According to these groups the following metrics can then be calculated, as described in [33]:

- Uncertainty sensitivity or U_{Sen}, corresponds to the percentage of uncertain predictions out of all incorrect predictions. A value closer to 1 means a lower model confidence on misclassified observations. It is calculated as:

$$U_{\text{Sen}} = \frac{\text{TU}}{\text{TU} + \text{FC}} \tag{4}$$

- Uncertainty Specificity or U_{Spe}, is the percentage of certain predictions out of all correct predictions. Values closer to 1 stand for a higher model confidence on correctly classified observations. It is calculated as follows:

$$U_{\text{Spe}} = \frac{\text{TC}}{\text{TC} + \text{FU}} \tag{5}$$

- Uncertainty Precision or U_{Pre}, corresponds to the percentage of incorrect predictions out of all uncertain predictions. A value closer to 1 means a lower model classification accuracy on uncertain predicted observations. Its calculation is depicted as follows:

$$U_{\text{Pre}} = \frac{\text{TU}}{\text{TU} + \text{FU}} \tag{6}$$

- Uncertainty Accuracy or U_{Acc}, corresponds to the percentage of favourable outcomes out of all predictions. It is calculated as follows:

$$U_{\text{Acc}} = \frac{\text{TU} + \text{TC}}{\text{TU} + \text{TC} + \text{FU} + \text{FC}} \tag{7}$$

In addition to these, we propose the implementation of another metric:

- Uncertainty Balanced Accuracy or U_{BAcc}, corresponds to an average of U_{Sen} and U_{Spe} for a given model. It offers a better insight to the overall confidence of a model in both its correct and incorrect predictions, and can be useful to detect situations in which a high U_{Sen} or U_{Spe} value is reached due to a model being, respectively, certain or uncertain about almost every prediction. It is calculated as follows:

$$U_{\text{BAcc}} = \frac{U_{\text{Sen}} + U_{\text{Spe}}}{2} \tag{8}$$

As such U_{BAcc} represents an average of a model's confidence in its correct and incorrect predictions. This metric can be of special interest in the medical domain, as a value closer to 1 would mean that uncertainty estimations for a model could be better trusted. This is of high importance to decide if a prediction should be submitted for a second opinion by a human professional or not, as high uncertainty could be associated with incorrect predictions and low uncertainty with correct predictions. Additionally, we note that in some cases the use of U_{BAcc} can be preferred to the single use of U_{Acc}, as the latter could be prone to optimistic results if the proportion of correct-incorrect predictions is unbalanced. This is inherently the case in models with high classification accuracy.

V. Experiments

A. Dataset

The INbreast dataset was used for the experiments depicted in this work. Published at [34], INbreast is a mammographic database comprised of multiple full-field digital mammograms of a wide variety of patients with anomalies like masses and calcifications. Each image labeled according to the BI-RADS scale, from category 1 to 6.

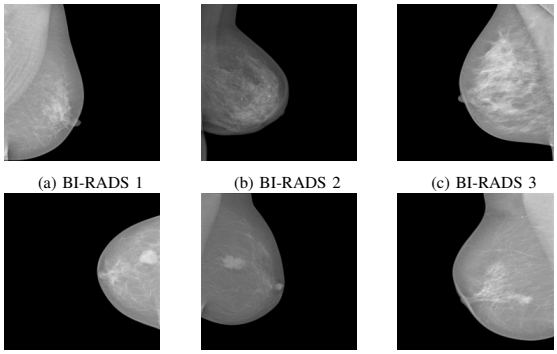
In our work, we focused on binary classification (malign, benign) of mammograms. Therefore, the BI-RADS labeled images from INbreast were divided into 2 groups. Following a similar approach to [26], categories 1 and 2 are defined as negative observations (benign). Categories 4, 5 and 6 are defined as positive observations (malign). Category 3 (probably benign) is ignored. This results in 287 negative and 100 positive observations.

B. Experiments

All the experiments described in this work were implemented in Python using the FastAI and PyTorch libraries, based on the MixMatch implementation found in [4].

The PyTorch implementation of the VGG 19-layer with batch normalization was chosen as the architecture for both supervised and semi-supervised models. Transfer learning with pre-trained weights from ImageNet was used as in [24], [26], [29].

Both supervised and SSDL models were trained and tested across a total of 10 different randomly-generated subsets of



(a) BI-RADS 1 (b) BI-RADS 2 (c) BI-RADS 3

(d) BI-RADS 4 (e) BI-RADS 5 (f) BI-RADS 6

Fig. 1: Examples of mammograms according to their BI-RADS classification in INbreast

data, with an average distribution of 70% of images for training and 30% for testing on each subset, with observations from different patients for training and for testing. Both types of models were trained on each data subset with $n_l = 10, 20, 40, 60$, with n_l number of labeled observations balanced with 50% for each class. Supervised models were trained only with labeled images, while the SSDL models used the remaining training images as unlabeled data.

Data augmentation was implemented in the train data in the form of random flips and rotations through the FastAI library, for both supervised and SSDL models. Due to memory constraints, images were resized to $224x224$ pixels, after being converted from DICOM format to PNG.

Models were trained for 50 epochs each, with early stopping in order to avoid overfitting. A learning rate of 0.00002, a weight decay of 0.001 and a batch size of 10 images were used. Hyper-parameters for MixMatch were set following the authors' recommendations in [16]: $K = 2$ transformations, a sharpening temperature of $T = 0.25$, an alpha mix value of $\alpha = 0.75$ and unsupervised $\gamma = 200$.

Furthermore, the Softmax-based and MCD uncertainty estimation methods were evaluated for each model, using the test data from their respective random subset of data. The results for each individual input were averaged to obtain "per-model" results. These values were then averaged across the 10 data subsets in order to obtain "per-amount-of-labeled-observations" results. For the MCD method, a total of $M = 100$ evaluations were made for each test input, using a dropout value of $p = 0.5$.

Finally, the described per-model results for the evaluated uncertainty estimation methods were grouped according to their "correctness" and "confidence" by, respectively comparing them to their ground truth labels and to a specific threshold. This was done in order to calculate the quantitative uncertainty metrics described in the previous section.

Due to the different value ranges obtained for the uncertainty estimation methods, normalization was applied to the obtained results. This way the same threshold value required for the metrics depicted in Equations 4, 5, 6, 7 and 8 could be employed. MCD values were normalized using the maximum and minimum values obtained over all test configurations. The max Softmax values were also normalized to be in a range from 1 to 0, and then their complements were calculated.

VI. Results

Results of the described uncertainty estimation tests are presented in Tables II, III and IV, as the mean and standard deviation across each of the 10 random data subsets, for the correct and wrong estimations of the models as well as the difference between these, along with the number of labeled observations used for training. As a comparison reference, model accuracy is similarly reported in Table I, averaged across the 10 randomly-generated data subsets for each number of labeled observations used for training.

As seen in Table I, SSDL models yielded moderately better accuracy results in comparison to supervised models. This occurs especially when using a very reduced number of labeled observations, as higher accuracy gains were obtained with the smallest number of labeled observations. The accuracy gain decreases with larger values of n_l.

A similar trend can be observed in results using the evaluated uncertainty estimation methods. Tables II and III suggest better uncertainty estimation results when using SSDL with very few labeled observations. For instance, in Table II the difference between the sample averages for both the MCD uncertainty estimator with the SSDL and supervised model decreases as the number of labeled observations n_l increase. A similar behavior can be seen for the rest of tested uncertainty estimators in Tables III and IV. The sample averages difference for the correct and incorrect samples are calculated in Tables II, III and IV. For this attribute, the following trend becomes evident: a higher incorrect to correct uncertainty difference is seen with the SSDL model, when compared to the supervised model. This happened for all the three tested uncertainty estimation methods. However, as the standard deviation is significant, we need to make further assessment of such results using specific uncertainty estimation metrics.

As previously mentioned in Section IV, we measured U_{Sen}, U_{Spe}, U_{Pre}, U_{Acc}, and U_{BAcc} with SSDL and supervised models, as depicted in Tables V and VI. The quantitative metrics for uncertainty estimation were measured for all methods with a threshold of 0.3, as advised in [33]. These results are presented in Tables V and VI, with the mean and standard deviation of each metric across each of the 10 random data subsets, for every number of labeled observations used.

Tables V.a and V.b depict the uncertainty metrics for the SSDL and supervised model with $n_l = 10$ and $n_l = 20$ number of labeled observations. For $n_l = 10$ number of labeled observations, the highest values were yielded by the SSDL model, with the exception of the model uncertainty sensitivity, which is dramatically better for the supervised model. The balanced uncertainty accuracy U_{BAcc} averages the U_{Sen} and the U_{Spe}, and still shows an advantage of the SSDL model over the supervised approach. With $n_l = 20$, the U_{Pre} gets worse for the SSDL model when compared with the

supervised model, however the U_{BAcc} still is higher for the SSDL model.

For the models trained with more labeled observations, Tables VI.a and VI.b show the results with $n_l = 40$ and $n_l = 60$ number of labeled observations, respectively. With $n_l = 40$, still the SSDL presents higher sample averages than the supervised model for U_{BAcc}, U_{Acc} and U_{Spe}, and lower sample means for the rest of the metrics. With $n_l = 60$, the SSDL model presents lower sample means for U_{BAcc}, U_{Acc} and U_{Spe} when compared to the supervised model, decreasing the gain of SSDL over the supervised model.

Moreover, relatively high standard deviations are observed for the majority of these results, thus Wilcoxon signed-rank tests were applied in order to identify statistically significant differences between results for SSDL and supervised models. Tables V and VI show a visible trend where statistically significant advantage of using SSDL is yielded as the number of labeled observations approach to $n_l = 10$.

The comparison between the tested uncertainty estimation approaches using the proposed metric U_{BAcc} and the rest of the measurements performed, shows no significant difference between the three tested approaches.

Differences between the reliability of each uncertainty estimation method are mixed, with no clear winner between them. Overall results of metrics with MCD are quite similar to those with Max Softmax. Results with normalized entropy of Softmax tend to higher U_{Sen} but lower U_{Spe} values than the other methods, which can suggest a higher perceived uncertainty over both correct and wrong predictions. Our proposed metric, the U_{BAcc} shows very similar values with no statistically significant difference between the three tested uncertainty estimation methods.

n_l	d.e.	No SSDL	SSDL
10	$\bar{x}$	0.6477	**0.7007**
	s	0.0776	0.0558
20	$\bar{x}$	0.6606	**0.7298**
	s	0.0639	0.0679
40	$\bar{x}$	0.6910	**0.7524**
	s	0.0434	0.0488
60	$\bar{x}$	**0.7245**	0.7015
	s	0.0252	0.1331

TABLE I: Accuracy results of the supervised and semi-supervised models for data subsets with different number of observations n_l.

VII. Conclusions

In this work, we tested the impact on accuracy and uncertainty estimation for deep learning models by using unlabeled data through the MixMatch SSDL approach. For this, three uncertainty estimation methods were evaluated: MCD, Normalized entropy of Softmax and maximum value of Softmax. An improvement with statistical significance was observed for SSDL models over supervised models, for lower number of labeled observations used for training. This was first suggested by the improvements over the differences in confidence between correct and incorrect predictions in SSDL models that were detected.

n_l	d.e.	No SSDL			SSDL		
		Correct	Wrong	Diff.	Correct	Wrong	Diff.
10	$\bar{x}$	0.1758	0.1766	0.0008	0.0675	0.0854	**0.0179**
	s	0.0120	0.0120		0.0416	0.0448	
20	$\bar{x}$	0.1676	0.1681	0.0005	0.1145	0.1227	**0.0082**
	s	0.0089	0.0110		0.0399	0.0332	
40	$\bar{x}$	0.1538	0.1714	0.0175	0.0938	0.1131	**0.0193**
	s	0.0224	0.0164		0.0377	0.0376	
60	$\bar{x}$	0.1493	0.1685	**0.0192**	0.1200	0.1270	0.0070
	s	0.0245	0.0159		0.0260	0.0194	

TABLE II: MCD statistics of the supervised and semi-supervised models for data subsets with different number of labeled observations n_l. Lower values indicate higher model confidence.

n_l	d.e.	No SSDL			SSDL		
		Correct	Wrong	Diff.	Correct	Wrong	Diff.
10	$\bar{x}$	0.9380	0.9461	0.0081	0.3644	0.4599	**0.0955**
	s	0.0843	0.0819		0.2029	0.2146	
20	$\bar{x}$	0.9106	0.9172	0.0066	0.6979	0.7415	**0.0436**
	s	0.1429	0.1319		0.2780	0.2461	
40	$\bar{x}$	0.7180	0.7952	0.0772	0.5335	0.6352	**0.1017**
	s	0.2342	0.1778		0.2468	0.2361	
60	$\bar{x}$	0.6679	0.7342	**0.0663**	0.7379	0.7814	0.0434
	s	0.2399	0.1911		0.1949	0.1620	

TABLE III: Normalized entropy of Softmax statistics of the supervised and semi-supervised models for data subsets with different number of labeled observations n_l. Lower values indicate higher model confidence.

n_l	d.e.	No SSDL			SSDL		
		Correct	Wrong	Diff.	Correct	Wrong	Diff.
10	$\bar{x}$	0.5994	0.5934	0.0060	0.8913	0.8580	**0.0333**
	s	0.0730	0.0705		0.0715	0.0847	
20	$\bar{x}$	0.6157	0.6118	0.0039	0.7488	0.7249	**0.0238**
	s	0.0936	0.0916		0.1300	0.1230	
40	$\bar{x}$	0.7282	0.6929	0.0353	0.8314	0.7819	**0.0495**
	s	0.1312	0.1102		0.0918	0.0988	
60	$\bar{x}$	0.7544	0.7270	**0.0274**	0.7308	0.7108	0.0200
	s	0.1295	0.1162		0.1145	0.1004	

TABLE IV: Max Softmax statistics of the supervised and semi-supervised models for data subsets with different number of labeled observations n_l. Higher values indicate higher model confidence.

Quantitative metrics based in "uncertainty confusion matrices" were implemented in order to further assess the performance of the three uncertainty estimation methods evaluated. In addition, a new metric was proposed and evaluated, the U_{BAcc}. The U_{BAcc} takes into account the very imbalanced nature of the number of correct and incorrect predictions, as the number of correct predictions tend to be higher when the model accuracy increases, as is the case with the SSDL model over the supervised one. The U_{BAcc} is generally lower than the measured U_{Acc} in such cases, and represents a more realistic measurement under heavy imbalanced situations. Nevertheless, the SSDL models got a higher U_{BAcc}, with the exception of the model trained with the highest number of labeled observations ($n_l = 60$), and also correlates well with the higher correct to incorrect average uncertainties difference, described in Tables

Method	Metric	SSDL		NO SSDL	
		$\bar{x}$	s	$\bar{x}$	s
MCD	U_{Sen}	0.3960	0.2993	**0.9803**	0.0525
	U_{Spe}	**0.7073**	0.2703	0.0158	0.0456
	$U_{\text{Pre}}*$	**0.3715**	0.1299	0.3514	0.0779
	U_{Acc}	**0.6142**	0.1301	0.3540	0.0803
	U_{BAcc}	**0.5517**	0.0622	0.4980	0.0063
Max Softmax	U_{Sen}	0.3974	0.2978	**0.9729**	0.0856
	U_{Spe}	**0.7204**	0.2514	0.0357	0.0999
	$U_{\text{Pre}}*$	**0.3889**	0.0826	0.3549	0.0791
	U_{Acc}	**0.6227**	0.1172	0.3625	0.0868
	U_{BAcc}	**0.5589**	0.0502	0.5043	0.0081
Normalized Entropy of Softmax	U_{Sen}	0.5531	0.2543	**0.9958**	0.0132
	U_{Spe}	**0.5622**	0.2922	0.0029	0.0092
	$U_{\text{Pre}}*$	**0.3762**	0.0804	0.3520	0.0774
	U_{Acc}	**0.5602**	0.1411	0.3523	0.0776
	U_{BAcc}	**0.5576**	0.0336	0.4994	0.0020

(a) $n_l = 10$

Method	Metric	SSDL		NO SSDL	
		$\bar{x}$	s	$\bar{x}$	s
MCD	$U_{\text{Sen}}*$	0.7080	0.3043	**0.9607**	0.0866
	$U_{\text{Spe}}*$	**0.3091**	0.3358	0.0552	0.1166
	U_{Pre}	0.2858	0.0624	**0.3442**	0.0682
	$U_{\text{Acc}}*$	**0.4445**	0.1688	0.3563	0.0845
	$U_{\text{BAcc}}*$	**0.5086**	0.0517	0.5080	0.0189
Max Softmax	$U_{\text{Sen}}*$	0.7601	0.3335	**0.9210**	0.1714
	$U_{\text{Spe}}*$	**0.2908**	0.3590	0.0884	0.1890
	U_{Pre}	0.2943	0.0563	**0.3431**	0.0673
	$U_{\text{Acc}}*$	**0.4454**	0.1780	0.3596	0.0903
	$U_{\text{BAcc}}*$	**0.5254**	0.0462	0.5047	0.0110
Normalized Entropy of Softmax	$U_{\text{Sen}}*$	0.8587	0.2311	**0.9760**	0.0610
	$U_{\text{Spe}}*$	**0.2005**	0.3114	0.0325	0.0686
	U_{Pre}	0.2941	0.0535	**0.3415**	0.0640
	$U_{\text{Acc}}*$	**0.3984**	0.1644	0.3486	0.0723
	$U_{\text{BAcc}}*$	**0.5296**	0.0440	0.5043	0.0179

(b) $n_l = 20$

TABLE V: Uncertainty metrics for all estimation methods, with $n_l = 10$ and $n_l = 20$, using a threshold of 0.3. Rows marked with $*$ indicate no statistic relevance in differences between results of SSDL and supervised models.

Method	Metric	SSDL		NO SSDL	
		$\bar{x}$	s	$\bar{x}$	s
MCD	U_{Sen}	0.6236	0.2925	**0.8672**	0.1387
	U_{Spe}	**0.5149**	0.2984	0.2442	0.2342
	$U_{\text{Pre}}*$	0.3103	0.0680	**0.3477**	0.0585
	U_{Acc}	**0.5448**	0.1535	0.4363	0.1303
	$U_{\text{BAcc}}*$	**0.5692**	0.0325	0.5557	0.0593
Max Softmax	$U_{\text{Sen}}*$	0.6050	0.3138	**0.7766**	0.2349
	U_{Spe}	**0.5115**	0.3412	0.3192	0.2983
	$U_{\text{Pre}}*$	0.3067	0.0818	**0.3512**	0.0738
	$U_{\text{Acc}}*$	**0.5395**	0.1814	0.4607	0.1492
	$U_{\text{BAcc}}*$	**0.5582**	0.0440	0.5479	0.0598
Normalized Entropy of Softmax	$U_{\text{Sen}}*$	0.7550	0.2673	**0.9057**	0.1242
	U_{Spe}	**0.3460**	0.3113	0.1846	0.1882
	$U_{\text{Pre}}*$	0.2815	0.0585	**0.3358**	0.0501
	$U_{\text{Acc}}*$	**0.4489**	0.1691	0.4074	0.1056
	$U_{\text{BAcc}}*$	**0.5505**	0.0502	0.5451	0.0540

(a) $n_l = 40$

Method	Metric	SSDL		NO SSDL	
		$\bar{x}$	s	$\bar{x}$	s
MCD	$U_{\text{Sen}}*$	0.7136	0.2172	**0.8245**	0.1496
	$U_{\text{Spe}}*$	**0.3266**	0.2488	0.2752	0.2396
	$U_{\text{Pre}}*$	**0.3156**	0.1443	0.3120	0.0525
	$U_{\text{Acc}}*$	**0.4757**	0.1164	0.4273	0.1392
	$U_{\text{BAcc}}*$	0.5201	0.0469	**0.5498**	0.0634
Max Softmax	$U_{\text{Sen}}*$	**0.7946**	0.2031	0.6949	0.2631
	$U_{\text{Spe}}*$	0.2739	0.2606	**0.3799**	0.2935
	$U_{\text{Pre}}*$	**0.3234**	0.1335	0.3081	0.0399
	$U_{\text{Acc}}*$	0.4561	0.1370	**0.4681**	0.1446
	$U_{\text{BAcc}}*$	0.5343	0.0398	**0.5374**	0.0421
Normalized Entropy of Softmax	$U_{\text{Sen}}*$	**0.9406**	0.1296	0.8738	0.1282
	$U_{\text{Spe}}*$	0.1060	0.1802	**0.2191**	0.2285
	$U_{\text{Pre}}*$	**0.3104**	0.1329	0.3076	0.0487
	$U_{\text{Acc}}*$	0.3679	0.1428	**0.4001**	0.1361
	$U_{\text{BAcc}}*$	0.5233	0.0318	**0.5464**	0.0603

(b) $n_l = 60$

TABLE VI: Uncertainty metrics for all estimation methods, with $n_l = 40$ and $n_l = 60$, using a threshold of 0.3. Rows marked with $*$ indicate no statistic relevance in differences between results of SSDL and supervised models.

II, III and IV. The U_{BAcc} can be used to measure in a more fair manner models with a considerably higher number of correct predictions over incorrect ones.

In general, SSDL models achieved statistically significant improvements in these metrics over a small number of labeled observations ($n_l = 10$), suggesting more accurate model confidence estimations over supervised models. Nevertheless, fluctuating results over U_{Sen} and U_{Spe} values for different amount labeled observations indicate that models find it difficult to simultaneously assert ideal high and low confidence levels for correct and wrong predictions respectively. Thus resulting in either high or low overall model uncertainties.

As practical conclusions of our work, the results obtained through the experiments performed suggest a sensitive uncertainty estimation accuracy gain of using SSDL, as long as there is also an important accuracy gain when using unlabeled data. The model trained with the highest number of labeled observations ($n_l = 60$), showed neither an accuracy gain or better uncertainty estimation.

In the context of mammogram analysis, the usage of unlabeled data to feed an SSDL algorithm as MixMatch, can be attractive when a very small labeled target dataset is available (for instance a small dataset from the clinic or hospital which plans to use the deep learning model for mammogram classification). Including a different and larger dataset can be done through transfer learning or including it as part of the labeled dataset. The former might be a better option, as a possible distribution mismatch between the labeled and the unlabeled dataset might lead to an accuracy decrease for the SSDL model, as suggested in [35]. This leads us to our future work, where the interaction between transfer-learning, SSDL with distribution mismatch and uncertainty estimation accuracy for mammogram analysis can be explored. Also, as future work, we plan to analyze the impact of data imbalance for mammogram classification using SSDL for both model accuracy and uncertainty estimation. In our tests, data imbalance revealed to have an important influence in the accuracy of the SSDL model, as suggested in [36]. Also, the impact of a distribution mismatch between the labeled and unlabeled dataset in the uncertainty estimations can be explored, as this is a frequent challenge in real world medical imaging analysis.

VIII. Acknowledgments

This research was partially supported by a machine allocation on Kabré supercomputer at the Costa Rica National High Technology Center. It was also partially supported by the Multidisciplinary Research Program eScience of the In-

stituto Tecnológico de Costa Rica. This work is also partially supported by the Ministry of Economy and Competitiveness of Spain under grant PPIT.UMA.B1.2017. It is also partially supported by the Ministry of Science, Innovation and Universities of Spain under grant RTI2018-094645-B-I00. Additionally, it is also partially supported by the Autonomous Government of Andalusia (Spain) under grant UMA18-FEDERJA-084. 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 Málaga. The authors also thankfully acknowledge the grant of the Universidad de Málaga and the Instituto de Investigación Biomédica de Málaga (IBIMA).

REFERENCES

[1] C. Wild, E. Weiderpass, and B. Stewart, "World cancer report: cancer research for cancer prevention," *Lyon: International Agency for Research on Cancer*, 2020.

[2] American Cancer Society, "Breast cancer facts & figures 2019-2020," *American Cancer Society, Inc.*, pp. 1–44, 2019.

[3] A. Hamidinekoo, E. Denton, A. Rampun, K. Honnor, and R. Zwiggelaar, "Deep learning in mammography and breast histology, an overview and future trends," *Medical image analysis*, vol. 47, pp. 45–67, 2018.

[4] S. Calderon-Ramirez, R. Giri, S. Yang, A. Moemeni, M. Umana, D. Elizondo, J. Torrents-Barrena, and M. A. Molina-Cabello, "Dealing with scarce labelled data: Semi-supervised deep learning with mix match for covid-19 detection using chest x-ray images," IEEE Press, 2020.

[5] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.

[6] A. Taleb, W. Loetzsch, N. Danz, J. Severin, T. Gaertner, B. Bergner, and C. Lippert, "3d self-supervised methods for medical imaging," *arXiv preprint arXiv:2006.03829*, 2020.

[7] R. Alizadehsani, M. Roshanzamir, S. Hussain, A. Khosravi, A. Koohestani, M. H. Zangooei, M. Abdar, A. Beykikhoshk, A. Shoeibi, A. Zare, *et al.*, "Handling of uncertainty in medical data using machine learning and probability theory techniques: A review of 30 years (1991-2020)," *arXiv preprint arXiv:2008.10114*, 2020.

[8] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, "Leveraging uncertainty information from deep neural networks for disease detection," *Scientific reports*, vol. 7, no. 1, pp. 1–14, 2017.

[9] P. Seeböck, J. I. Orlando, T. Schlegl, S. M. Waldstein, H. Bogunović, S. Klimscha, G. Langs, and U. Schmidt-Erfurth, "Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct," *IEEE transactions on medical imaging*, vol. 39, no. 1, pp. 87–98, 2019.

[10] B. Ghoshal and A. Tucker, "Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection," *arXiv preprint arXiv:2003.10769*, 2020.

[11] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[12] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, pp. 1050–1059, 2016.

[13] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," *arXiv preprint arXiv:1706.04599*, 2017.

[14] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical image analysis*, vol. 54, pp. 280–296, 2019.

[15] A. Chebli, A. Djebbar, and H. F. Marouani, "Semi-supervised learning for medical application: A survey," in *2018 International Conference on Applied Smart Systems (ICASS)*, pp. 1–9, IEEE, 2018.

[16] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Advances in Neural Information Processing Systems*, pp. 5049–5059, 2019.

[17] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[18] A. Chan, A. Alaa, Z. Qian, and M. Van Der Schaar, "Unlabelled data improves bayesian uncertainty calibration under covariate shift," in *International Conference on Machine Learning*, pp. 1392–1402, PMLR, 2020.

[19] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[20] S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, and S. Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," *Advances in Neural Information Processing Systems*, vol. 32, pp. 13888–13899, 2019.

[21] L. Ihnatchenko, "Deep neural networks in image processing," 2020.

[22] C. J. D'Orsi, "The american college of radiology mammography lexicon: an initial attempt to standardize terminology.," *AJR. American journal of roentgenology*, vol. 166, no. 4, pp. 779–780, 1996.

[23] I. Domingues, P. H. Abreu, and J. Santos, "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.

[24] L. Falconí, M. Pérez, W. Aguilar, and A. Conci, "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.

[25] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, S. Jastrzebski, T. Févry, J. Katsnelson, E. Kim, *et al.*, "Deep neural networks improve radiologistśperformance in breast cancer screening," *IEEE transactions on medical imaging*, vol. 39, no. 4, pp. 1184–1194, 2019.

[26] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep learning to improve breast cancer detection on screening mammography," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[27] A. Duggento, M. Aiello, C. Cavaliere, G. L. Cascella, D. Cascella, G. Conte, M. Guerrisi, and N. Toschi, "An ad hoc random initialization deep neural network architecture for discriminating malignant breast cancer lesions in mammographic images," *Contrast media & molecular imaging*, vol. 2019, 2019.

[28] A. Rampun, B. W. Scotney, P. J. Morrow, and H. Wang, "Breast mass classification in mammograms using ensemble convolutional neural networks," in *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pp. 1–6, IEEE, 2018.

[29] D. Lévy and A. Jain, "Breast mass classification from mammograms using deep convolutional neural networks," *arXiv preprint arXiv:1612.00542*, 2016.

[30] N. Zemmal, N. Azizi, M. Sellami, and N. Dey, "Automated classification of mammographic abnormalities using transductive semi supervised learning algorithm," in *Proceedings of the Mediterranean Conference on Information & Communication Technologies 2015*, pp. 657–662, Springer, 2016.

[31] H. Azary and M. Abdoos, "A semi-supervised method for tumor segmentation in mammogram images," *Journal of Medical Signals and Sensors*, vol. 10, no. 1, p. 12, 2020.

[32] R. Shen, J. Yao, K. Yan, K. Tian, C. Jiang, and K. Zhou, "Unsupervised domain adaptation with adversarial learning for mass detection in mammogram," *Neurocomputing*, 2020.

[33] H. Asgharnezhad, A. Shamsi, R. Alizadehsani, A. Khosravi, S. Nahavandi, Z. A. Sani, and D. Srinivasan, "Objective evaluation of deep uncertainty predictions for covid-19 detection," *arXiv preprint arXiv:2012.11840*, 2020.

[34] I. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. Cardoso, and J. Cardoso, "Inbreast: Toward a full-field digital mammographic database," *Academic radiology*, vol. 19, pp. 236–48, 11 2011.

[35] Y. Chen, X. Zhu, W. Li, and S. Gong, "Semi-supervised learning under class distribution mismatch," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 3569–3576, 2020.

[36] S. Calderon-Ramirez, A. Moemeni, D. Elizondo, S. Colreavy-Donnelly, L. F. Chavarria-Estrada, M. A. Molina-Cabello, *et al.*, "Correcting data imbalance for semi-supervised covid-19 detection using x-ray chest images," *arXiv preprint arXiv:2008.08496*, 2020.