Feature Density as an uncertainty estimator method in the binary classification mammography images task for a supervised Deep Learning model

Ricardo Javier Fuentes-Fino[1,2], Saúl Calderón-Ramírez[2], Enrique Domínguez[1,3], Ezequiel López-Rubio[1,3], Marco A. Hernandez-Vasquez[2], and Miguel A. Molina-Cabello[1,3]

[1] Department of Computer Languages and Computer Science
University of Malaga, Málaga, Spain
[2] Instituto Tecnologico de Costa Rica, Costa Rica
[3] Instituto de Investigación Biomédica de Málaga – IBIMA, Málaga, Spain
RicardoFino@estudiantec.cr,{sacalderon,marco.hernandez}@itcr.ac.cr,
{enriqued,ezeqlr,miguelangel}@lcc.uma.es

Abstract. Labeled medical datasets may include a limited number of observations for each class, while unlabeled datasets may include observations from patients with pathologies other than those observed in the labeled dataset. This negatively influences the performance of the prediction algorithms. Including out-of-distribution data in the unlabeled dataset can lead to varying degrees of performance degradation, or even improvement, by using a distance to measure how out-of-distribution a piece of data is. This work aims to propose an approach that allows estimating the predictive uncertainty of supervised algorithms, improving the behaviour when atypical samples are presented to the distribution of the dataset. In particular, we have used this approach to mammograms X-ray images applied to binary classification tasks. The proposal makes use of Feature Density, which consists of estimating the density of features from the calculation of a histogram. The obtained results report slight differences when different neural network architectures and uncertainty estimators are used.

Keywords: Feature Density, Mahalanobis distance, Jensen-Shannon distance, Uncertainty, Deep Learning

1 Introduction

Machine Learning (ML) approaches are trying to be applied in the field of medicine as a tool to help in classification and diagnosis tasks of diseases like cancer and more recently COVID-19 by using medical images [1, 2]. Cancer is the first or second leading cause of premature death and breast cancer remains the leading cause of death in women worldwide, although it can also be diagnosed in men [3]. In 2019, it was estimated that 268,600 new cases of invasive breast cancer were diagnosed among women and approximately 2,670 cases diagnosed in men [4]. To mitigate these numbers, it is necessary an early and accurate diagnosis. The analysis of imaging evaluation such as mammography or histopathological [5, 6] images may supply that diagnosis. Due to this, approaches like ML have been extensively studied to improve classification tasks and apply them to medical diagnosis.

In areas such as medicine, the main problem is the limited data set, its quality and the acquisition process, and it causes that not all approaches are suitable and not all methods provide optimal performance. ML algorithms usually face many problems in real-world deployment environments and several examples of this can be found [7–10]. According to [7] and [8] the labelled dataset can include a limited number of observations for each class, in the context of breast cancer, a more significant number of samples without cancer can be observed than with cancer, which can cause a tendency of the models to classify better (or recognize) the samples of the majority class, this is known as Data Imbalance. Also in [9] mentioned that the test dataset can include observations of patients with other types of pathologies than those observed in the training dataset, this is known as Out-Of-Distribution (OOD) data, and it can be potentially harmful to classifications models performance and cause a degradation in its accuracy. Another well-studied problem [10] is the mismatch distribution of the data. This usually happens when deploying the algorithms to a real-world environment. Training models with a specific dataset does not guarantee that testing the model in another setting (another hospital or clinic, usually called target dataset) will give the same performance results.

Experimental evidence shows that despite accuracy being harmed by the problems mentioned above and in [11] mentions that obtaining models that can generalize the characteristics of breast cancer is complicated since there is significant variability of anomalies which will always limit the efficiency of the algorithms, the ML techniques they remain an attractive approach for the detection, classification or segmentation of different types of anomalies. Hence, it is essential to continue their improvement and investigation.

In ML, uncertainty measures how reliable or accurate a model is in classifying the images in a test data set based on the supervised training that the model has performed. In this work, we evaluate feature density as a measure of uncertainty and compare this method with others proposed in state-of-the-art like Mahalanobis distances. To perform this investigation, we offer the following question: is it possible to obtain a statistically significant improvement between using Feature Histogram to improve the estimation of predictive uncertainty concerning other techniques that assume a Gaussian distribution of the data set?

2 State of the Art

In [12] they propose to combine two uncertainty measurements. The first one, based on subjective logic [13], $u(p) : p \to \mathbb{R}$, based on the information contained from the probabilistic predictions, while the second, a data closeness measurement $D_m(z) : z \to \mathbb{R}$ following a Mahalanobis approach [14] that measures the distance D_m of a sample to the training distribution cluster. They have observed that the Mahalanobis distance brings a complementary aspect, especially related to out-of-distribution cases [14]. For instance, when a classifier trained on breast images (ID) is fed with outliers from a flower dataset (OOD), the authors saw that the rejection criterion based on the Mahalanobis distance is quite effective. Despite the effectiveness of the combination, further research is required on automatic ways to find the optimal thresholds.

On the other hand, [15] their focus is on uncertainty estimation methods that are practical and straightforward to implement. Specifically, the Softmax and Monte Carlo Dropout (MCD) approaches were tested. 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. 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. Softmax method alone can lead to poor representations of model uncertainty due to typical overconfidence in neural networks' predictions. The MCD approach aims at having more robust estimations while still being simple to implement [16], when compared to the usage of Softmax for uncertainty estimation. MCD is based on a Bayesian interpretation of the model's parameters. According to their results, an improvement with statistical significance was observed for SSDL models over supervised models.

To deal with data imbalance, [8] propose to use the transfer learning approach. Multiple models were trained under different training configurations to evaluate the impact of SSDL on their Transfer learning (a simple Domain adaptation method) and loss function based class-imbalance correction were also tested. Deep learning models were first trained in a supervised manner with complete mammography datasets $D^l_{s,INbreast}$ and $D^l_{s,DDSM}$ in order to obtain source-trained models which were further fine-tuned on their target Costa Rican dataset in a Supervised manner, with limited amounts of labelled observations. In summary, models that were subject to do main adaptation from a source mammography dataset showed improved classification performance results in comparison to other experimental configurations tested there. nd

3 Methods

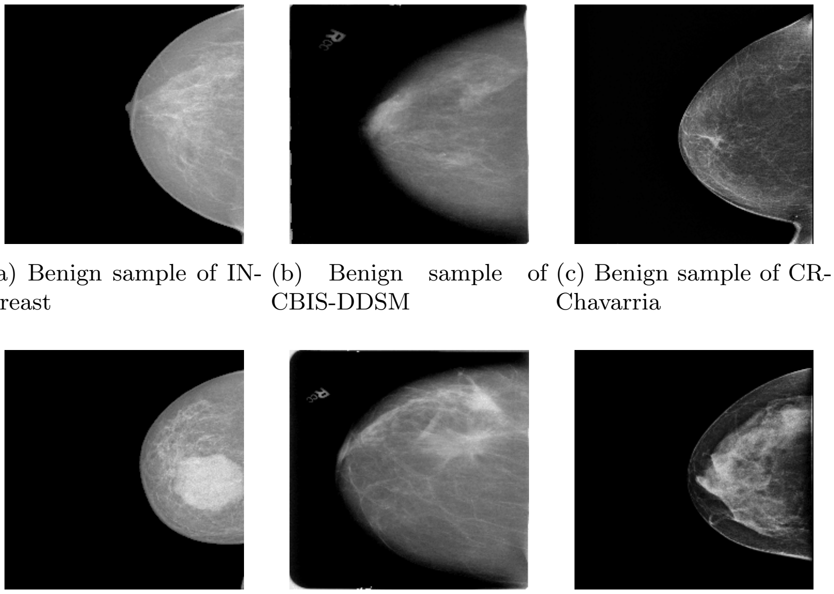
3.1 Mammography Datasets

Three different mammography datasets were used to carry out the experiments. The characteristics of those datasets are summarized in Table 1 and some samples of X-Ray images are illustrated in Figure 1.

INbreast The INbreast dataset introduced in [17] is a dataset containing a wide variety of breast anomalies such as masses, calcifications, architectural distortions, asymmetries and images with multiple anomalies at the same time, and

Table 1: Summary of characteristics of the datasets.

	INbreast [17]	**CBIS-DDSM [18]**	**CR-Chavarria 2020 [8]**
Origin	Portugal	United States	Costa Rica
Year	2011	1997-2016	2020
Cases	115	1522	87
Images	410	3103	282



(a) Benign sample of INbreast (b) Benign sample of CBIS-DDSM (c) Benign sample of CR-Chavarria

(d) Malignant sample of INbreast (e) Malignant sample of CBIS-DDSM (f) Malignant sample of CR-Chavarria

Fig. 1: Mammogram samples from each dataset used according to a binary classification from a CC view (top-down view of the breast).

usual patient samples. This dataset was built from 115 cases of X-ray images originating at Centro Hospitalar de São João at Porto, Portugal. Of the 115 cases, 90 cases have associated two images for each breast, belonging to each of the views (Craniocaudal (CC): which is a top to bottom view of the breast; and Mediolateral oblique (MLO): which is a side view of the breast); that is, 4 images associated with each patient; the remaining 25 cases only have related images for each of the views; giving a total of 410 X-ray images. The resolution of the images varies depending on the size of the patient's breast. In addition, these images were evaluated and classified according to the categories of BI-RADS and according to their density measurement. For this case, the images were acquired digitally (Full-Field Digital Mammography) and stored in a DICOM (Digital Imaging and Communications in Medicine) format.

CBIS-DDSM The Curated Breast Imaging Subset of Digital Database of Screening Mammography (CBIS-DDSM) [18] is an improved version of the Digital Database of Screening Mammography, which contained 2620 cases from different sources. This dataset has X-Ray images with standard samples, benign and malignant cases of breast cancer. The main problem with the original database was that some of the information attached to each case was limited or difficult to access. Due to this, a new dataset is created to improve the quality; to do this, inaccurate images or images that did not meet confidentiality standards are discarded. In [8] it is detailed that CBIS-DDSM contains a total of 3103 digitized images (scanned) belonging to 1566 cases, separated according to the anomaly presented in the X-Ray images (masses or calcifications) and classify according to the category of the BI-RADS system and according to its density measure. By classifying the dataset in a binary way, a total of 1728 images with benign cases were obtained and 1375 images with malignant cases.

CR-Chavarria-2020 Introduced in [8] the dataset from the Dr. Chavarria Estrada Medical Imaging private clinic located in Costa Rica. In [8] this dataset is used as out-of-distribution data as it comes to represent the conditions of a real-world deployment environment for the Machine Learning algorithms. The dataset was built from 87 cases, whose patients have an age range of 40 to 90 years. It contains 341 images, of which only 282 images are used, because in some cases the image does not have optimal quality or the patients have breast implants, which could produce noise in the classification models. When performing the classification in a binary way, the result is that 268 images are negative samples and 14 images are positive samples of cancer, showing a clear data imbalance in its classes. The images belonging to CR-Chavarria-2020 dataset were evaluated and classified according to the BI-RADS categories. Also, the images were acquired digitally form(FFDM).

3.2 Data Preprocessing

As part of the X-Ray image preprocessing from all three datasets described above, it was necessary to perform three operations on the datasets:

- A readjustment of the resolution of each image was performed, resulting in images of $224x224$ pixels, dimensions also used in the state-of-the-art literature in previous experiments, in order to reduce execution time, processing load and amount of disk space used.
- It was also necessary to change the file extension (image format) from DICOM to BMP (Windows Bitmap).
- This work was focused on the binary classification of the samples, because of this it was necessary a reclassification of the available datasets, similar to [8], where mammograms labelled with BI-RADS categories 4, 5 and 6 are defined as positive cases of breast cancer, while mammograms labelled with categories 1 and 2 are defined as negative cases of breast cancer. Image

samples labelled with categories 0 and 3 were discarded due to the peculiarity of their characteristics.

It was necessary to perform a second preprocessing stage on the dataset CBIS-DDSM since the X-ray images belonging to this set were digitized (scanned), thus their images were noisy. The anomalies observed are the following:

- In the pixels surrounding the breast it is observed as a blur (pixels in different shades of grey) similar to a shadow, which could cause the classification algorithms to take those areas as part of the image's characteristics and cause a classification deficiency. To clean up noise, it was used the procedure described in [19].
- Despite the preprocessing that was given to the images described in the previous point, after a visual inspection it was found that in some images there were still remains of annotations of the type of view or data belonging to the X-ray, which could generate a bias within the classification model. To eliminate the remaining noise, it was necessary to make manual annotations of the area with noise and treat them using an algorithm.

After a second visual inspection of the images in the CBIS-DDSM dataset, it was possible to observe that in some exceptions the algorithm removed a considerable part of the breast. For these cases, manual cleaning of the image was carried out, similar to item two described above.

3.3 Training Process

For this work, the FastAI implementations of AlexNet and DenseNet architectures were chosen as classification models, were used a pre-trained version of the same and subsequently a Fine-Tuning process was performed on the dataset INbreast and CBIS-DDSM.

Initially, the configuration of hyperparameters used is the default configuration by the FastAI library, i.e. no modification was made to the algorithm to improve its accuracy when classifying images, with that a maximum of 70% accuracy was obtained on classification tasks, to improve that and achieve the accuracy reported in the state-of-art was resorted to using of Adam optimization function and data augmentation technique but was not obtain a statistical improvement.

Since the purpose of this work is not focused on obtaining models with the best possible accuracy in classification tasks, but to try uncertainty techniques, no further modifications were made to the classification models and left the default settings. To a certain extent, it is sought that the models are not perfect and that they make errors, in order to be able to evaluate the uncertainty estimators.

Initially, the models were trained from 857 X-ray images as shown in Table 2 for a maximum of 50 epochs. The selection of these images was done randomly. In order to improve the accuracy of the models, it was also experimented the

Table 2: Composition of images from the training dataset

Dataset	**Number of images**	**Class Balance**
INbreast BI-RADS-1	47	242
INbreast BI-RADS-2	195	
INbreast BI-RADS-4	34	77
INbreast BI-RADS-5	39	
INbreast BI-RADS-6	4	
CBIS-DDSM Benign Calcifications	140	329
CBIS-DDSM Benign Masses	189	
CBIS-DDSM Malignant Calcifications	92	209
CBIS-DDSM Malignant Masses	117	

training of the models with more epochs (e.g. 200 epochs) and tried to use a more balanced training set, but it did not obtain an improvement of the performance.

From the training process, the feature extractor was obtained, which in simple words are all those operations or mathematical processes that the network has used to extract the features of images. The feature extractor is used as part of the uncertainty estimators. The aim is to obtain the features of the correct and incorrect estimations and compare them with the features of the training images.

3.4 Uncertainty Estimation Process

Once the training of the models is finished, the uncertainty estimators were evaluated. For this, 10 test sets were used. Once the network has classified the test images, the confusion matrix and the network's predictions were used to find out the number of correct and incorrect estimations. From this information, representative subsets were created, these sets (correct and incorrect estimations) were subsequently processed by the uncertainty estimator models, together with the other necessary parameters. (similar to data flow shown in Figure 2).

For the Mahalanobis Distance method, it was necessary to calculate the covariance matrix and the vector of means, from the training dataset, these elements are the basis that was used to estimate the uncertainty of the previously built image sets. For each image within the subsets mentioned above, an uncertainty measurement was obtained, thus creating two vectors of uncertainty, i.e. a vector with uncertainties of correct estimations and the other with uncertainties of incorrect estimations. Once this information was obtained, a PDF (Probability Density Function) was created for each of the uncertainty vectors, and it proceeded to calculate the distance between them (Jensen-Shannon Distance). The distance will be compared subsequently with the other estimator method.

For the Feature Density method, it was first necessary to estimate the feature histogram of the training dataset, this histogram is the basis for estimating the uncertainty of the previously constructed image subsets. As in the previous method, for each subset (correct and incorrect estimations) a vector was obtained that contains each one of the uncertainty measurements corresponding to each

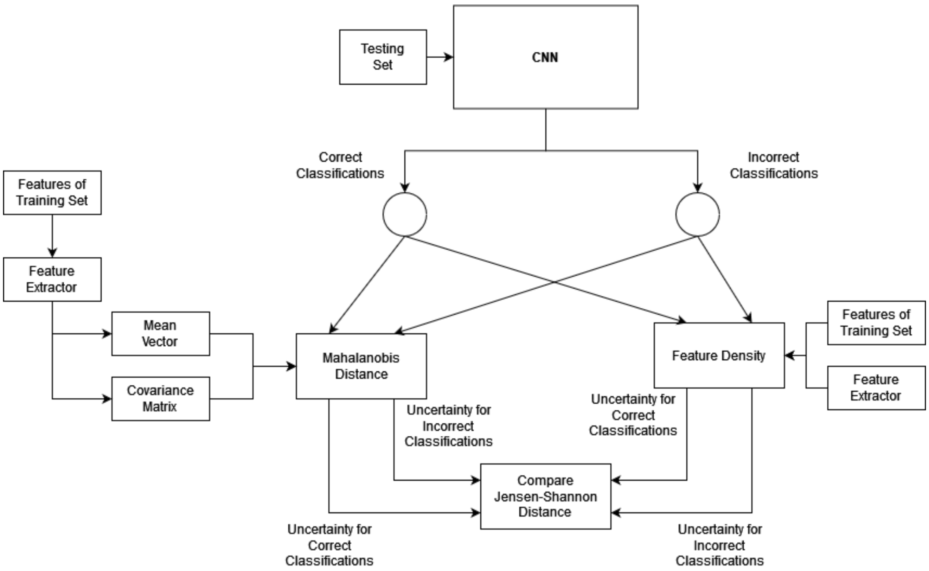

Testing Set
CNN
Correct Classifications
Incorrect Classifications
Features of Training Set
Feature Extractor
Mean Vector
Covariance Matrix
Mahalanobis Distance
Feature Density
Features of Training Set
Feature Extractor
Uncertainty for Incorrect Classifications
Uncertainty for Correct Classifications
Compare Jensen-Shannon Distance
Uncertainty for Correct Classifications
Uncertainty for Incorrect Classifications

Fig. 2: Schema of the estimation of uncertainty

image. Again, another PDF was created for each of the uncertainty vectors and the distance between them was calculated.

Once the Jensen-Shannon distance of the uncertainty vectors has been measured using each of the methods, a direct comparison was made as to which method is more accurate. As mentioned above, the Jensen-Shannon distance of the uncertainty distribution is intended to be as large as possible.

4 Experiment results

To evaluate the performance of the uncertainty estimator models, 10 experiments (batches) were used, each of the test sets had 60 randomly selected X-ray images, covering each of the types of images available. It is important to mention that the network had never seen the images of test sets previously. In the first five experiments were used in-of-distribution images, i.e. images that belonged to the INbreast and CBIS-DDSM datasets with which the network was trained. In the remaining five experiments, different degrees of out-of-distribution data contamination were used, as shown in Table 3, belonging to the CR-Chavarria-2020 dataset.

The first experimental stage it was necessary to train the AlexNet architecture with the INbreast and CBIS-DDSM dataset with the number of images detailed in Table 2, 20% of the total images were used as a validation set. The neuronal network was trained for 50 epochs. The maximum accuracy obtained in the train validation was 70%.

Table 3: Evaluation experiments for the uncertainty estimation methods

Experiments without contamination			**Experiments with contamination**		
N° of Exp.	**Number of images**	**Distribution Percentage**	**N° of Exp.**	**Number of images**	**Distribution Percentage**
1	60	100% IOD	6	60	75% IOD 25% OOD
2	60	100% IOD	7	60	50% IOD 50% OOD
3	60	100% IOD	8	60	50% IOD 50% OOD
4	60	100% IOD	9	60	25% IOD 75% OOD
5	60	100% IOD	10	60	100% OOD

Table 4: Number of correctly and incorrectly classified images, using INbreast and CBIS-DDSM as IOD data and CR-Chavarria as OOD data, with an Alexnet architecture for classification.

Experiments without contamination				**Experiments with contamination**			
N° of Exp.	**Correct. Estimations**	**Incorrect. Estimations**	**Acc**	**N° of Exp.**	**Correct. Estimations**	**Incorrect. Estimations**	**Acc**
1	33	27	0,5500	6	31	28	0,5254
2	31	29	0,5167	7	31	29	0,5167
3	32	28	0,5333	8	33	27	0,5500
4	33	27	0,5500	9	40	20	0,6667
5	28	32	0,4647	10	45	15	0,7500

Despite not obtaining high accuracy in the classification tasks, it was not taken as an impediment to continue with the experiments, since a perfect classification model was not sought. Table 4shows the number of correct and incorrect estimations made by the neural network over the test dataset, as well as the accuracy with which it was made.

Not in all experiments can the capacity of the neural network to classify OOD data be determined with such precision, although experiment 10 of Table 4 can be taken as a basis, where there is 100% of OOD data and the model adequately classified 75% of the samples. In Table 5 and 6 the averages of the uncertainty measurements were compiled for the ten experiments carried out in this stage.

Despite being hardly noticeable, when analyzing the averages of the uncertainty values, there are two tendencies:

- The difference between the uncertainty measurements for the correct and incorrect estimations is minimal in the case of the Mahalanobis Distance, whereas with the Feature Density method the uncertainty measurements for

Table 5: Average of uncertainty measurements over the correct and incorrect estimations, using INbreast and CBIS-DDSM as IOD data.

N° of Exp.	**Mahalanobis Distance**		**FD Method**	
	Correct. Estimations	**Incorrect. Estimations**	**Correct. Estimations**	**Incorrect. Estimations**
1	9,7627	7,6000	388,2441	386,3513
2	7,9117	6,3012	384,3943	394,4933
3	8,9966	7,2569	336,9922	381,4414
4	7,6128	8,4158	385,8245	395,2873
5	9,4562	7,2151	385,8129	394,8537

Table 6: Average of uncertainty measurements over the correct and incorrect estimations, using INbreast and CBIS-DDSM as IOD data and CR-Chavarria as OOD data.

N° of Exp.	**Mahalanobis Distance**		**FD Method**	
	Correct. Estimations	**Incorrect. Estimations**	**Correct. Estimations**	**Incorrect. Estimations**
6	9,0266	5,7338	416,1937	423,9887
7	8,6063	8,1578	491,3473	465,5490
8	8,0021	7,0746	459,0135	478,2890
9	9,2823	6,6258	520,9386	505,8273
10	11,5599	6,6558	548,6212	554,1428

the incorrect estimations are a little greater than the uncertainty measurements for the correct estimations.

- The uncertainty measurements for the experiments with OOD data are a little greater than the uncertainty measurements for the experiments without OOD data, the most noticeable difference could be seen with the Feature Density method.

The observations above are not always met, especially using the Mahalanobis Distance method. Thus, it is necessary more experiments to determine the causes. All information about the comparison between both methods are showed in Table 7.

One aspect in which there is a big difference between both estimating methods is in the execution time and computational cost. With a convolutional layer belonging to the AlexNet architecture, the Mahalanobis Distance method takes an average of 0.3 milliseconds to process an experimental batch, while with the Feature Density method it takes an average of 41 seconds. The big difference between the execution times is due to the calculation of the Feature Histogram for each one of the dimensions of the training set when it is processed by the Feature Extractor. To calculate the execution time using the Mahalanobis Distance method, the computation time of the covariance matrix and the vector of

Table 7: Jensen-Shannon distance between the uncertainties of correct and incorrect estimations, using INbreast and CBIS-DDSM as IOD data and CR-Chavarria as OOD data. Classification architecture: AlexNet.

Experiments without contamination			**Experiments with contamination**		
N° of Exp.	**JS Distance with the Mahalanobis Method**	**JS Distance with the FD Method**	**N° of Exp**	**JS Distance with the Mahalanobis Method**	**JS Distance with the FD Method**
1	0,3639	0,3579	6	0,3865	0,3011
2	0,3883	0,3409	7	0,3639	0,3480
3	0,3573	0,3158	8	0,3469	0,4000
4	0,4419	0,3069	9	0,3666	0,3079
5	0,2932	0,4647	10	0,3896	0,5324
Avg	0,3689	0,3573		0,3707	0,3779
Std	0,0481	0,0566		0,0157	0,0849

means plus the batch processing time are added. In the case of Feature Density, the time it takes to calculate the Feature Histogram of the training set is added plus the batch processing time.

As a second experimental stage, a DenseNet architecture was used, the process of both training, validation and testing was similar to that used with the AlexNet architecture.

The results obtained for the Jensen-Shannon distance are shown in Table 8. As can be seen when using a feature extractor belonging to the DenseNet network, there is a more notable difference between both estimating methods; In this case, the Feature Density method is the one with the highest value for both the IOD and the OOD samples. This would indicate that the performance of the method is related to the type of Feature Extractor that is used.

When using a more complex Feature Extractor, the execution time and the computational cost increased significantly for both methods. For the Mahalanobis method the average time in the execution of the experiments was 3.6047 seconds, while for the Feature Density estimator it was 1763.3704 seconds (approximately 30 minutes), this difference between the times is due to the fact that with the Feature Extractor produced from the DenseNet architecture, 1024 dimensions are obtained as a result, at which The Feature Histogram must be calculated from the training data set. Therefore, the little gain obtained by estimating the uncertainty is overshadowed by the execution time invested.

5 Conclusions and recommendations

This research was carried out to evaluate the feature density method as an uncertainty estimator, applied to the binary classification of X-ray images (mammograms), using the AlexNet and DenseNet neural network architectures.

Table 8: Jensen-Shannon distance between the uncertainties of the correct and incorrect estimations, using INbreast and CBIS-DDSM as IOD data and CR-Chavarria as OOD data. Classification architecture: DenseNet.

Experiments without contamination			**Experiments with contamination**		
N° of Exp.	**JS Distance with the Mahalanobis Method**	**JS Distance with the FD Method**	**N° de Exp**	**JS Distance with the Mahalanobis Method**	**JS Distance with the FD Method**
1	0,2934	0,3479	6	0,1076	0,4151
2	0,2722	0,4098	7	0,3647	0,3779
3	0,2234	0,3988	8	0,3710	0,4193
4	0,3476	0,5553	9	0,4280	0,4163
5	0,3105	0,5180	10	0,3798	0,4209
Avg	0,2894	0,4460		0,3296	0,4099
Std	0,0412	0,0778		0,1135	0,0161

Based on the results of this work, no statistically significant improvement was found between the feature density method concerning the Mahalanobis Distance as an uncertainty estimator method when using an AlexNet architecture. In the case of the DenseNet architecture, a more notable difference can be observed, but the results are not entirely conclusive. This way, more experiments are needed to reach a more accurate answer.

If the execution time and the computational cost invested in estimating the uncertainty using both methods are taken into consideration, it can even be thought that the Mahalanobis Distance has some advantage from that perspective. It is necessary to emphasize that the execution time and computational cost is closely related to the type of architecture selected for the experiments.

Despite the conclusions reached in this research, this does not mean that the feature density method should be discarded entirely as an estimator of uncertainty. Like everything in Artificial Intelligence, more experiments must be carried out to reach an accurate conclusion about which method has a better performance.

As recommendations to continue with the work raised in this research, it proposes:

- Perform more experiments, with a more significant number of images for both training and testing. As there are few images and tests, no conclusive trend regarding improvement can be observed. Another recommendation is to experiment with data augmentation approaches and find the optimal combination of transformations on the images.
- Use other convolutional network architectures to investigate if there are architectures (and thus their feature extractor) where the performance of the feature density method might be better.

- Experiment with the hyperparameters of the architectures until finding an optimal configuration, which can reach the accuracy proposed in [15] and experiment if there is a variation in the estimation of the uncertainty.
- Experiment with other datasets of medical images, with the possibility that in different contexts, a significant improvement is obtained, since not necessarily when getting a low or high performance in a specific context means that it must work in the same way in others.

Acknowledgments

This work is partially supported by the following Spanish grants RTI2018-094645-B-I00 and UMA18-FEDERJA-084. All of them include funds from the European Regional Development Fund (ERDF). The authors acknowledge the computer resources, technical expertise and assistance provided by the SCBI (Supercomputing and Bioinformatics) center of the University of Malaga. The authors acknowledge the funding from the Instituto de Investigación Biomédica de Málaga – IBIMA and the Universidad de Málaga.

References

1. S. Calderon-Ramirez, S. Yang, A. Moemeni, S. Colreavy-Donnelly, D. A. Elizondo, L. Oala, J. Rodríguez-Capitán, M. Jiménez-Navarro, E. López-Rubio, and M. A. Molina-Cabello, "Improving uncertainty estimation with semi-supervised deep learning for covid-19 detection using chest x-ray images," *Ieee Access*, vol. 9, pp. 85 442–85 454, 2021.
2. 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," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 5294–5301.
3. C. Wild, E. Weiderpass, and B. Stewart, "World cancer report: Cancer research for cancer prevention," *Lyon,France: International Agency for Research on Cancer*, 2020.
4. A. C. Society, "Breast cancer facts & figures 2019-2020," *Atlanta: American Cancer Society*, 2019.
5. M. A. Molina-Cabello, C. Accino, E. López-Rubio, and K. Thurnhofer-Hemsi, "Optimization of convolutional neural network ensemble classifiers by genetic algorithms," in *International Work-Conference on Artificial Neural Networks*. Springer, 2019, pp. 163–173.
6. M. A. Molina-Cabello, J. A. Rodríguez-Rodríguez, K. Thurnhofer-Hemsi, and E. López-Rubio, "Histopathological image analysis for breast cancer diagnosis by ensembles of convolutional neural networks and genetic algorithms," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.
7. S. Calderon-Ramirez, S. Yang, A. Moemeni, D. Elizondo, S. Colreavy-Donnelly, L. F. Chavarría-Estrada, and M. A. Molina-Cabello, "Correcting data imbalance for semi-supervised covid-19 detection using x-ray chest images," *Applied Soft Computing*, vol. 111, p. 107692, 2021.

8. S. Calderón Ramírez, D. Murillo-Hernández, K. Rojas-Salazar, D. Elizondo, A. Moemeni, and M. A. Molina-Cabello, “A real use case of semi-supervised learning for mammogram classification in a local clinic of costa rica,” *Medical & Biological Engineering & Computing*, 2022.
9. S. Calderon-Ramirez, L. Oala, J. Torrents-Barrena, S. Yang, A. Moemeni, W. Samek, and M. A. Molina-Cabello, “Mixmood: A systematic approach to class distribution mismatch in semi-supervised learning using deep dataset dissimilarity measures,” *arXiv preprint arXiv:2006.07767*, 2020.
10. S. Calderon-Ramirez, S. Yang, D. Elizondo, and A. Moemeni, “Dealing with distribution mismatch in semi-supervised deep learning for covid-19 detection using chest x-ray images: A novel approach using feature densities,” *arXiv preprint arXiv:2109.00889*, 2021.
11. W. Sun, B. Tseng, J. Zhang, and W. Qian, “Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data,” *Computerized Medical Imaging and Graphics*, 2016.
12. M. Tardy, B. Scheffer, and D. Mateus, “Uncertainty measurements for the reliable classification on mammograms,” *Springer: International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 495-503)*, 2019.
13. A. Jøsang, *Subjective Logic: A Formalism for Reasoning Under Uncertainty*, ser. International series of monographs on physics. Cham, Switzerland : Springer, 2016.
14. T. Denouden, R. Salay, K. Czarnecki, V. Abdelzad, B. Phan, and S. Vernekar, “Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance,” *CoRR*, vol. abs/1812.02765, 2018. [Online]. Available: http://arxiv.org/abs/1812.02765
15. S. Calderón-Ramírez, D. Murillo-Hernández, K. Rojas-Salazar, L.-A. Calvo-Valverd, S. Yang, A. Moemeni, D. Elizondo, E. López-Rubio, and M. A. Molina-Cabello, “Improving uncertainty estimations for mammogram classification using semi-supervised learning,” in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.
16. Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” 2016.
17. I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, “Inbreast: Toward a full-field digital mammographic database,” *Academic radiology, 19(2), 236–248*, 2012.
18. R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. Rubin, “A curated mammography data set for use in computer-aided detection and diagnosis research,” *Scientific Data*, vol. 4, 2017.
19. A. R. Beeravolu, S. Azam, M. Jonkman, B. Shanmugam, K. Kannoorpatti, and A. Anwar, “Preprocessing of breast cancer images to create datasets for deep-cnn,” *IEEE Access*, vol. 9, pp. 33 438–33 463, 2021.