




# An uncertainty estimator method based on the application of feature density to classify mammograms for breast cancer detection

Ricardo Fuentes-Fino<sup>1</sup> · Saúl Calderón-Ramírez<sup>2</sup> · Enrique Domínguez<sup>1,4</sup> · Ezequiel López-Rubio<sup>1,4</sup> · David Elizondo<sup>3</sup> · Miguel A. Molina-Cabello<sup>1,4</sup> 

Received: 29 September 2022 / Accepted: 14 July 2023 / Published online: 23 August 2023  
© The Author(s) 2023

## Abstract

In the area of medical imaging, one of the factors that can negatively influence the performance of prediction algorithms is the limited number of observations for each class within a labeled dataset. Usually, in order to increase the samples, a second set of unlabeled images is used. However, this set adds two new problems (i) finding patient observations with different pathologies than those observed in the labeled data set and (ii) finding images belonging to a different distribution from the dataset used in the model training process. This way, merging datasets from different sources can have an adverse effect on the distribution of features. Encountering this type of data (better known as out-of-distribution data) within the deployment environments may also lead to varying degrees of performance degradation as can be seen in the different experimental results obtained. In this research, a study of the behavior of Feature Density is made, as a mathematical model for the estimation of predictive uncertainty in supervised classification algorithms, in order to improve the behavior when out-of-distribution data are presented in the dataset. The Feature Density method is based on the estimation of feature density by means of histogram calculation (or Probability Density Function). The advantage of this method over the baseline approach (Mahalanobis distance) is that it does not assume a Gaussian-type distribution of sample characteristics and serves to estimate the uncertainty. This work focuses on the binary classification of mammography X-ray images from three different datasets simulating the condition of a different degree of contamination with out-of-distribution sample. According to the obtained results, the performance of the proposed method depends directly on the architecture of the implemented neural network.

**Keywords** Feature density · Mahalanobis distance · Jensen–Shannon distance · Uncertainty estimation · Deep learning

## 1 Introduction

Nowadays, many fields of knowledge investigate and use Artificial Intelligence models (and their different approaches) for the processing and analysis of data (structured and unstructured). It is one of the most important technologies in the fourth industrial revolution and in the not-too-distant future it will have an influence on people's daily lives [1]. So far, different techniques have been developed (and are still being developed) that mimic a part of human

behavior to solve specific and complex problems. Emerging advances in the application of deep learning in different areas of knowledge have been published [1]. In the field of medicine, the use of different ML (Machine Learning) approaches is being studied as a support tool in tasks of classification and diagnosis of diseases such as cancer in its many facets, tissue abnormalities, and, more recently, with the COVID-19 pandemic [2].

According to [3], in 134 of 183 countries around the world, cancer is the first or second leading cause of premature death. Breast cancer is the most commonly diagnosed cancer in women (although less frequently this disease can also be diagnosed in men). In [4], the authors mention that about 287,850 possible new cases of breast cancer in women will be diagnosed in the US during the

---

Ricardo Fuentes-Fino, Saúl Calderón-Ramírez, Enrique Domínguez, Ezequiel López-Rubio, David Elizondo have contributed equally to this work.

---

Extended author information available on the last page of the article

year 2022 and about 2,710 cases in men. Nowadays, one of the most effective strategies in the fight against breast cancer is routine check-ups, which allow an early diagnosis that can be as accurate as possible. Diagnosis is commonly made through manual evaluation of images such as mammographies. Being a rudimentary process, a certain margin of uncertainty or error can be generated, giving way to poor diagnoses. Due to this, approaches such as ML are being of interest and widely investigated as a support tool in classification tasks and medical diagnosis. The main problems of applying ML or any other AI approach in the area of medicine are the limited number of samples in the dataset (generalization of a pattern), the quality of the data, and the process in which they are acquired. As a consequence of this, not all the proposed models are adequate or provide optimal performance.

Despite the serious problems that ML algorithms still face when deployed in real-world environments, they still remain an attractive approach with many advantages [5–7]. It is usual that within the labeled data set, there is a very limited number of observations that can adequately represent the characteristics of some of the classes (anomalies) of the case of study, while other classes may have a huge number of observations. In our context (breast cancer), it was observed that within the datasets used in the experiments, there is a large number of observations with negative cases in cancer compared to positive cases; a specific example is to compare the number of samples of the BI-RADS 6 classification with the number of BI-RADS 1 samples in each of the sets used. This difference in the number of samples can cause a bias in the classification process since the models will tend to more easily recognize the characteristics of the class with the largest number of samples, a phenomenon known as Data Imbalance. Additionally, sometimes, within the test dataset or the unlabeled dataset (for semi-supervised models), observations with pathologies different from those observed in the training set can be found. This type of data anomaly is known as out-of-distribution data (OOD) and can be harmful to the performance of classification models, causing a degradation in the accuracy value [6]. In these cases with data imbalance, it must be highlighted to not confuse the effect (bad performance of the model) with the cause (data imbalance).

A third well-studied problem in [7] is the mismatch distribution of data or features. This problem usually happens when models are implemented in a different environment than the one in which they were trained or developed; another point of view mentioned in [8] is that, traditionally, Deep Learning models are trained and tested from the same data set, but this is not always true in real-world scenarios due to the complicated process of obtaining data (for instance, in the area of medicine), so it is

necessary to train the models with data that is easier to obtain. For the training of the classification models, a specific dataset is typically used, thus obtaining a specific performance; but when deploying the same model in another environment (usually called target dataset) replication of performance results are not be guaranteed.

The most common strategy to face the aforementioned problems is to increase the labeled data set to obtain a good generalization of the characteristics of the case study; this would allow classifying any sample within a test dataset. However, in an area like medicine, getting a large set of labeled data is very expensive, both financially and professionally. As explained in [9], creating a dataset focused on medical images requires: a large amount of human effort (to manually label the images), financial expenses to hire the necessary professional staff, or in some cases it is necessary to build the infrastructure to collect the information. The authors of [10] also mention that on many occasions for the labeling work, it is necessary for several radiologists to evaluate the dataset individually, compare and discuss the results of these evaluations with each other on a case-by-case basis to arrive at an accurate final conclusion. A viable alternative to the scarcity of data and to prevent overfitting of the models is to perform certain transformations on the base images (for example, rotations, augmentations, cuts, and geometric transformations) in this way the number of images available for training is increased; this technique is called data augmentation. Some novel data augmentation methods, based on feature transformation, are proposed in [11]; also, in [12], the most commonly used techniques for data augmentation are mentioned. Another acceptable option that is being studied is the use of semi-supervised algorithms, which in turn allow the use of large unlabeled data sets combined with smaller labeled datasets, the work carried out in [13] presents several SSDL methods used in classification tasks. Despite the results obtained in different contexts where the unlabeled set helps to improve model training, it does not exempt this new set from facing the same problems described above. It has been shown experimentally that the problems mentioned above affect the accuracy of the models. In [10], they mention that obtaining a good generalization of the characteristics (patterns) of any case study is complicated since there will always be significant variability in the observations, thus limiting the efficiency of the models; this should provide enough motivation to continue improving and investigating new models.

In the context of ML, the uncertainty measure indicates how reliable or accurate a model is in classifying the observations of a given dataset from the supervised training that has been provided to the model. In this research, the method called Feature Density is evaluated as a measure of uncertainty, comparing it with other proposed methods.

This work aims to study the possibility of obtaining a statistically significant enhancement by using a Feature Histogram to achieve a higher performance in the estimation of predictive uncertainty with respect to other techniques that assume a Gaussian distributed dataset. The Mahalanobis distance (baseline approach) is a method used to measure the distance between two points, usually the distance between any point and the mean (or centroid) of the data; similar to the Euclidean distance. In addition, the Mahalanobis method takes into account the covariance of the data. In the context of this research, it is used to measure how similar a test data set (separated by hits and misses) is to the base data set (training data). Feature Density consists of the estimation of the density of features from the calculation of a histogram; this takes the form of a graph that allows to represent the frequency distribution of a variable (in the context of this research would be the features belonging to a set of images). The advantage of this method over the Mahalanobis distance is that it is not necessary to assume that the features follow a Gaussian distribution, being this the main innovation of the present work.

## 2 State of the art

The authors of [14] propose to combine two methods to measure uncertainty estimation. The first is based on subjective logic [15],  $u(p) : p \rightarrow \mathbb{R}$ , which uses probabilistic predictions and the information contained in them. The second is based on the data proximity  $D_m(z) : z \rightarrow \mathbb{R}$  using the Mahalanobis method [16]; this measure  $D_m$  allows to quantify the distance between a sample and a training distribution cluster. Likewise, the authors in [16] determined that by applying the Mahalanobis Distance, out-of-distribution cases can be detected. For example, when training a classification model using images of breasts (ID) and subsequently analyzing outliers such as any other images (OOD), the Mahalanobis Distance rejection criterion is quite effective. Despite the combination of these methods and their effectiveness, in [14], the authors suggest that more research is needed focusing on determining optimal thresholds.

The authors in [17] focus on uncertainty estimation methods that are more practical and straightforward to implement; in their experiments, they use the Monte Carlo Dropout (MCD) and Softmax approaches. For uncertainty estimation, a Softmax activation function can be used in many ways; a basic way is to implement it in the output layer of deep learning models. Another way is to quantify the entropy value over the distribution of all Softmax output values, from an input  $x_j$ . The disadvantage of using

a method such as Softmax is that it can lead to a model that makes poor estimations of uncertainty, due to the excess confidence observed in the estimations made by neural network models. Another approach that has been studied is MCD, where the estimations tend to be more robust and simple to implement [18]. The MCD method makes use of the Bayesian interpretation of the model parameters, resulting in a significant improvement in the SSDL models compared to the supervised ones. There are different lines of research where mathematical models are proposed for the estimation of uncertainty since it is an important issue in the medical area and in decision-making; each option has advantages and disadvantages depending on the context in which they are applied and the models used for image classification.

Regarding the problem of data imbalance, [5] proposes to use the transfer learning approach. In order to validate the effect that Transfer Learning has on SSDL models, multiple models with different configurations for training were implemented, it was also experimented with loss function to solve the imbalance of classes. As a first stage in the experiments, a supervised training of the models was used on a dataset of mammograms ( $D'_{s,INbreast}$  and  $D'_{s,DDSM}$ ), thus obtaining source-trained models that are subsequently fine-tuned with a limited number of labeled samples from the target dataset (CR-Chavarria-2020). As a result of the investigation, an improvement in classification performance was found in the models subjected to domain adaptation compared to other configurations. On the other hand, in [19], they use multiple models to extract features from images (knowledge transfer) and then combine all the features into a single vector that serves as input to the classifier model; the results in this research indicate that it is better to have a layer composed of several feature extractors than to have a single one.

The aim of this research (based on [20]) is to compare the Mahalanobis Distance (baseline model), but unlike [14] it is not combined with any other mathematical model to improve its performance; with the Feature Density method proposed in [7] where it is used as a possible alternative solution to the class imbalance problem. On the one hand, the Mahalanobis Distance is a method used to measure the distance between two points (usually between any point and the centroid of the data); similar to the Euclidean Distance. In the context of this research, it is used to measure how similar a test dataset (correct and incorrect estimates made by a convolutional network) is to the base data set (training data) in this way the value of uncertainty. On the other hand, Feature Density is a method that consists of estimating the density of features from the calculation of a histogram; a graph that allows representing the distribution of frequencies of a variable. In the context of

this research, they would be the characteristics belonging to a set of images. The advantage of the proposed model over the base model is that it is not necessary to assume that the features follow a Gaussian-type distribution.

### 3 Methods

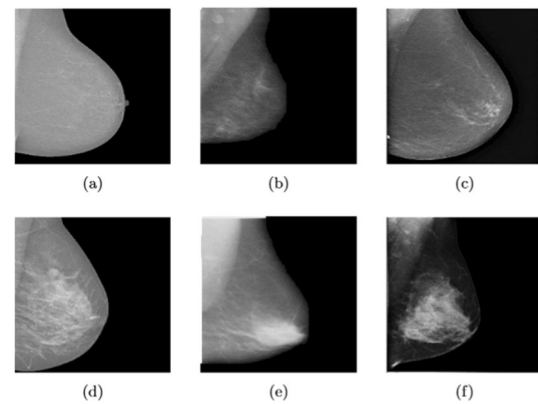
As previously mentioned, the objective of this research is to evaluate the Feature Density method as a measure of uncertainty in image classification. The main advantage of this method, on a comparative basis (with respect to the Mahalanobis distance), is that it does not assume a Gaussian distribution of the features of the samples belonging to the dataset. Feature Density computes Feature Histograms to obtain a measure of uncertainty from the feature distribution of both hits and misses (classified images) committed by CNN models (AlexNet, DenseNet, and MobileNet models). The main metric adopted to measure the performance of the proposed method is the Jensen–Shannon distance. This metric is usually used to quantify the difference that exists between two probability distributions; in the context of this research, the Jensen–Shannon distance of the uncertainty distribution is sought to be as large as possible, as the proportion of out-of-distribution samples in the test set increases. In order to observe a clear trend in the behavior of the proposed method in estimating the uncertainty value, approximately 54000 X-ray images were binary classified for each CNN model. These subsets (hits and misses) were then sent to the uncertainty estimator methods (Feature Density and Mahalanobis Distance) to obtain a measure of uncertainty for each image. Finally, a comparison by means of the Jensen–Shannon distance was done.

#### 3.1 Mammography datasets

Three mammography image datasets were used in this work: INbreast, CBIS-DDSM and CR-Chavarría-2020. The diversity of the samples intends to expose classifiers and uncertainty estimators to multiple observations and characteristics of breast cancer. This seeks to have a generalization in the recognition of the patterns that make up the images. Figure 1 shows some samples.

##### 3.1.1 INbreast

This dataset [21] is made up of 115 patient cases and provides a wide variety of breast cancer anomalies: masses, calcifications, architectural distortions, asymmetries, observations with multiple anomalies, and routine control samples. Mammography images have two views: Cranio-caudal (CC), a top-to-bottom view of the breast; and



**Fig. 1** Mammogram samples from each dataset according to a binary classification from a MLO view: **a** Negative case of INbreast, **b** Negative case of CBIS-DDSM, **c** Negative case of CR-Chavarría, **d** Positive case of INbreast, **e** Positive case of CBIS-DDSM and **f** Positive case of CR-Chavarría

Mediolateral oblique (MLO), a lateral view of the breast. From the 115 cases that make up this data set, 90 cases have an image for each view (CC and MLO), from each of the breasts; the 25 missing cases only have one associated image for each view and for each breast. These samples were taken from the mammography exams. Each of the 410 X-ray images was classified using the 6 BI-RADS categories and the density measurement, and their resolution depends on the size of the patient’s breast.

##### 3.1.2 CBIS-DDSM

According to [22], access to the information contained in the pre-existing dataset Digital Database of Screening Mammography was very complicated, and its quality was not very good; so they decided to create an improved version: Curated Breast Imaging Subset of Digital Database of Screening Mammography (CBIS-DDSM). The aim of this work was to contribute and improve the quality of the information of the collected cases, in addition to being easier to access. Another action taken by the authors was to remove inaccurate observations or observations that did not satisfy confidentiality standards. Finally, the authors of [22] structured all the images by separating them into other subsets intended for training and testing; this in turn allowed for a binary type classification of the images according to the type of anomaly present.

On the other hand, the authors in [5] mention some features of the CBIS-DDSM set. It is made up of 3,103 digitized (scanned) images from 1,566 patients. Abnormalities that can be detected on mammograms are masses and calcifications. Of the total number of images, it was determined that 1,728 images have benign anomalies and 1,375 images have malignant anomalies of breast cancer. Due to the number of images within this dataset and the



memory it uses, it was decided to only use the images belonging to the test subset.

### 3.1.3 CR-Chavarria-2020

The CR-Chavarria-2020 dataset [5] is a relatively new set compared to the INbreast and CBIS-DDSM sets (both of which are widely used in related research). The images belonging to this dataset come from the private clinic of Dr. Chavarria Estrada Medical Imaging located in Costa Rica. This set is made up of 87 cases; the age range of patients is between 40 and 90 years old. In this study, of the total of 341 images, only 282 images are used; the rest of the observations were discarded because they did not satisfy optimal quality or in some cases, the patients had breast implants, which would produce noise in the classification stage. If a binary-type classification is performed on the dataset, 268 images of negative cases are obtained, and only 14 images with positive cases. Each of the images was classified using the BI-RADS categories and was acquired digitally (FFDM). In [5] and other related research, CR-Chavarria-2020 is generally used as out-of-distribution data, since it has several characteristic conditions (such as data imbalance in its classes) of a dataset target and represents a real challenge for ML algorithms.

## 3.2 Data processing

As part of the preprocessing actions of the X-ray images belonging to the three data sets described above, the following operations were performed:

- Readjust the resolution of each image, resulting in images with dimensions of 224x224 pixels. These dimensions have been used in previous experiments to reduce the execution time of algorithms, reduce the processing load on the GPUs, and decrease the amount of disk space used by images.
- A file extension conversion process (image format) from DICOM to BMP.
- Being an investigation that focuses on the binary classification of the samples, a reclassification of the datasets was necessary, similar to how it is done in [5], where they are defined as positive cases of breast cancer those mammograms belonging to BI-RADS categories 4, 5 and 6. On the other hand, mammograms belonging to BI-RADS categories 1 and 2 were labeled as negative cases of breast cancer. All mammograms within BI-RADS categories 0 and 3 were eliminated due to the particularity of their characteristics.

Due to the peculiarity of the CBIS-DDSM data set (noisy and digitized X-ray images), a second preprocessing stage was performed on this set. In general, the observed

anomalies on the images are a blur or shadow (pixels in different shades of gray) in the pixels surrounding the breast (this could cause a deficiency in the classification of the images since there is a possibility that the classification algorithms take this noise as part of the image), and different annotations (metadata) in the images (this information describes the image taken of the breast, such as the type of view or relevant data for the doctor).

Using the procedure described in [23], the images of the CBIS-DDSM set were cleaned. To verify the effectiveness of the process, the images were reviewed by means of a visual inspection. In some cases, remains of annotations were found in the images; also in some exceptions the algorithm removed a considerable part of the pixels belonging to the breast. Both problems can cause the models to carry out a faulty classification. To solve this, the images were treated manually by means of annotations of the affected areas and the execution of their own algorithm.

## 3.3 Cross-validation process

The cross-validation method is widely used to ensure that experimental results do not depend solely on the partition selected for training and testing, thus giving a more substantial value to the results obtained. Initially, samples from the source data sets are classified using the BI-RADS category (in the case of INbreast and CR-Chavarria-2020) or by the present abnormalities, masses, or calcifications malignant/benign (in the case of CBIS-DDSM). In order to provide a balanced data set for the training and testing process of the models used in the experiments, the following steps were performed to create the different experimental image segments:

1. Each of the source data sets contains folders labeled according to the categories into which these sets have been classified.
2. The images inside each of the aforementioned folders were split into  $N = 10$  image sections. This will ensure, to some extent, that each of the future subsets (for training and testing) contains images of the different anomalies present and that the models will have the ability to generalize the features, as well as to avoid bias in the classification of the images.
3. Two steps are then performed simultaneously:
  - A binary classification of the categories, as explained in the previous section.
  - A clustering between the  $N_i$  image sections, i.e., shuffling all sections with the same index/position.

Visually, it could be said that there are now two folders: a first one representing the benign samples and a second one representing the malignant samples. In

turn, these folders contain  $N = 10$  subfolders, where the images from the INbreast and CBIS-DDSM sets have been mixed stratified.

4. Now, from these  $N = 10$  segments,  $w_{te} = 2$  (e.g.,  $N_1$  and  $N_2$ ) are selected to be taken as the test subset,  $w_v = 2$  other image segments will be used as the validation subset (e.g.,  $N_3$  and  $N_4$ ) and the rest  $w_{tr} = 6$  will be used for model training ( $N_5, N_6, \dots, N_{10}$ ).
5. As mentioned above, a characteristic shared by the data sets used in the experiments is that there are a larger number of benign samples than malignant ones. This leads to an imbalance in the classes (in the context of binary classification), making it necessary a balancing process in the classes. To create the training and validation subset,  $x$  number of benign samples equal to the number of malignant samples ( $y$ ) was randomly selected from each of the  $N$  image segments designated for that purpose. The remaining images become part of the designated test segments.
6. In the case of the test subsets,  $M = 10$  folds are created for each degree of contamination with OOD sample and the number of images shown in Table 1.
7. The previous steps are repeated using a stride  $j = 1$  to the right, i.e., now the  $N_2$  and  $N_3$  segments will be used for testing purposes,  $N_4$  and  $N_5$  for validation and the rest for training. This is until each segment has been used for all 3 purposes.

It is necessary to clarify that a similar process is performed with the CR-Chavarria-2020 set, intended as a data contamination set.

### 3.4 Experimental design

Figure 2 shows an overview of each of the components used in the evaluation process of Feature Density as an uncertainty estimator method. Each CNN model is trained and validated on a dataset that gathers images from the different categories coming from INbreast and CBIS-DDSM. Each subset is as balanced as possible (same

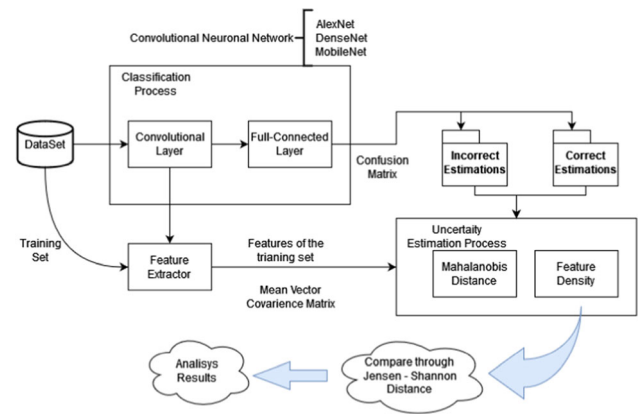


Fig. 2 Schema of the estimation of uncertainty

amount of benign and malignant samples). The CNNs, by means of their convolutional layers, extract features from the images and learn to identify/classify the anomalies present. In the testing stage, batches of images that the network has never seen are sent for classification; based on the hits and misses, two subsets of data are created (represented by folders in Fig. 2) and the uncertainty measure is calculated for each image within these subsets. Both the Mahalanobis method and the method proposed by this research are used here. Finally, the two methods are compared by means of the Jensen–Shannon distance metric. Details of the whole process are given in later sections.

From the images designated for testing, the experiments in charge of evaluating the performance of the uncertainty estimation methods were established. For each degree of contamination with OOD sample (starting from 0% contamination, progressively increasing by 12.5% contamination with OOD images, until reaching 100% contamination)  $K = 10$  experimental batches are created; in order to be able to observe a clear trend in the behavior of the evaluated methods. Each experimental batch consisted of 60 images that were randomly selected. It was sought that each batch be as balanced as possible between benign and malignant samples (this becomes impossible in the last degrees of contamination), as shown in detail in

**Table 1** Composition of negative/positive cases in breast cancer for the different experimental batches

Exp.	INbreast and CBIS-DDSM	CR BI-RADS 1	CR BI-RADS 2
100% IOD	30/30	–	–
87.5% IOD - 12.5% OOD	26/26	2/2	2/2
75% IOD - 25% OOD	22/23	4/5	4/2
62.5% IOD - 37.5% OOD	19/18	6/10	5/2
50% IOD - 50% OOD	14/16	8/12	8/2
37.5% IOD - 62.5% OOD	6/16	12/12	12/2
25% IOD - 75% OOD	8/7	15/12	16/2
12.5% IOD - 87.5% OOD	3/4	19/12	20/2
100% OOD	–	23/12	23/2

Table 1. Since these images are used for testing, it is not necessary to ensure that a sample is included for each of the anomalies present in the data sets.

### 3.5 Training process

For this research, the AlexNet [24], DenseNet [25] and MobileNet [26] architectures were used for the classification tasks. These architectures are from the FastAI library and are widely used for image classification. The details of the architecture and the diagram of each network can be found in their respective papers.

To speed up the training process, the pre-trained versions of the architectures were used. In addition, the Fine-Tuning technique was used to train the architectures on the data sets that were defined as IOD (INbreast and CBIS-DDSM). To avoid overfitting in the first training iterations of the architectures, data augmentation techniques were used. No hyperparameter configuration of the proposed architectures was made since the purpose of this research is not focused on testing models that obtain high precision, but on testing techniques for estimating uncertainty, for which models that are not perfect are needed.

The architectures in the different iterations of experiments were trained from approximately 440 X- mammograms covering as many anomalies as possible. In contrast, the validation of the training was carried out with approximately 150 images completely different from those used in the training process. The selection of these images was performed using the cross-validation technique described in the previous section. In order to see if the accuracy of the proposed architectures could be improved, it was decided to carry out some brief experiments where the number of training epochs was increased (for example, 200 epochs); but this only made the architectures overfit faster, given the configuration selected for this investigation.

After training, the feature extractor was obtained. This corresponds to all of the mathematical processes that the architecture used to obtain the features of the images. This element is used as one of the parameters in the uncertainty estimation methods, with the aim of making a comparison between the characteristics of the predictions and the characteristics of the training set.

### 3.6 Uncertainty estimation process

After the model training stage, the Feature Density method is evaluated as an uncertainty estimator together with the comparison base method. For this, 10 test sets were used for each level of contamination in the data. From the confusion matrix and the classification predictions made by the model, the correct and incorrect estimates were

separated. They were later processed by the uncertainty estimation methods in another stage, together with the parameters requested by each method. Figure 2 shows a schema of this process.

Concerning the base comparison method (Mahalanobis Distance), the covariance matrix and the vector of means must be estimated from the mammograms used in the training. The characteristics of these observations are used to estimate the uncertainty of the image subsets built in the previous stage. For each mammogram of the test set classified by the model, an uncertainty measure is computed and stored in an uncertainty vector according to whether it was a correctly classified estimate or not. As a last step, the estimation of the Jensen–Shannon Distance (JS) as a comparison metric is carried out by estimating the probability density function with the histogram. That is, the absolute frequency distribution for each uncertainty vector is measured, and it is subsequently normalized. This value is later used to compare with a similar value obtained from the proposed method. Regarding the proposed method of Feature Density, the histograms of features of the training data set must first be estimated, these represent the distribution of the features of the images and that is the basis for the calculation of uncertainty. The authors of [7] explain that to calculate the feature histogram for each sample (images of the training set)  $x \in \mathbb{R}^n$ , being  $n$  the number of dimensions is done by means of the feature extractor (usually refers to the last convolutional layer of the CNN architectures) this element calculates the feature vector, whose dimensions will depend on the neural network architecture that is being implemented; this creates a set of features given by  $H_l$ . From  $H_l$  the feature histogram is created; for each of the dimensions in the Feature Space, the normalized histogram must be calculated in order to approximate the density function  $P_r^l$  within  $H_l$ , thus producing a set of feature density functions. For more in-depth knowledge on the feature density process, please refer to the respective article. As in the previous method, for each image belonging to the subset (correct and incorrect estimates), an uncertainty value is obtained and stored in the corresponding vector. Again a frequency distribution is built for each uncertainty vector and the distance between them is computed.

This JS distance value allows a direct comparison between the base method and the proposed method; the value of the JS distance of the uncertainty distribution must be as large as possible, indicating which method is more suitable for estimating uncertainty.

## 4 Experiment results

To evaluate the performance and to be able to observe a clear trend in the results of the uncertainty estimation models, 54,000 mammography images were processed, divided into 10 folds containing 90 experimental batches each. For every 10 experimental batches, a degree of contamination was covered with the OOD data. Figure 3 shows the trend of the accuracy value in the training stage of each of the architectures, according to the results, the MobileNet architecture is the one that obtains the best precision in the validation stage when classifying the images; this behavior continues in the first experimental stages (first 5 degrees of contamination), but it degrades in the last 4 degrees of contamination (where there is more presence of OOD data than IOD), which indicates that the MobileNet architecture faces difficulties when classifying images that do not belong to the data set used in training, i.e., OOD data, as can be seen in Fig. 3. The architecture that maintains a similar behavior throughout the different experiments is the AlexNet. But this does not mean that it has the best Feature Extractor for uncertainty estimation. These accuracy values obtained by the implemented architectures differ, and to a certain extent are lower when compared to other values published in other articles, due to the fact that several datasets are used for training, validation, and testing in this research. Another possible cause of the degradation of the precision values may be due to the quality of the images of the CBIS-DDSM dataset, which is much lower compared to the other sets used.

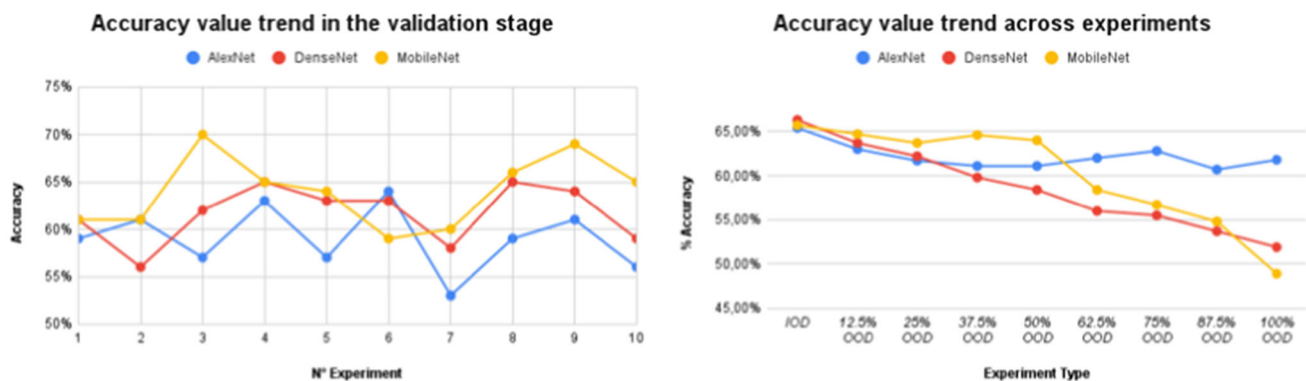
Despite not obtaining high accuracy values, this was not an impediment to continuing with the experiments, since this research focuses on analyzing methods for estimating uncertainty and not on obtaining the highest accuracy values. Another important aspect that must be taken into consideration when deciding which method is better is the

execution time that the experiments took; the more complex the structure of the convolutional layers (operations), the more processing time it takes to obtain the feature extractor. In order to reduce the processing load and the execution time, data parallelism was used through GPU. Despite this, each architecture took a considerable amount of time to train, obtain the base parameters and process all the experiments.

Figure 4 shows the trend for the average and the median value of the JS distance, through the experiments carried out. As can be seen, using any of the three Feature Extractors, the Feature Density method obtains higher values for the JS distance; being the AlexNet architecture where a greater difference is appreciated. Another deduction is that the greater the degree of contamination with OOD data, the greater the value of the JS distance. Therefore, it can be deduced that the performance of Feature Density is linked to the architecture that has been used for the classification of the images and it is not necessary to use models with complex convolutional layers to obtain good results.

## 5 Conclusion

The purpose of this work was to evaluate the Feature Density method as an estimation model of uncertainty, in the context of binary classification of mammograms. The neural network architectures used for the experiments were AlexNet, DenseNet, and MobileNet. According to the results obtained in the different experiments, there is no statistically significant difference between the Feature Density method and the Mahalanobis Distance as uncertainty estimator models when using the DenseNet and MobileNet convolutional network architectures. The opposite case can be seen when using an AlexNet

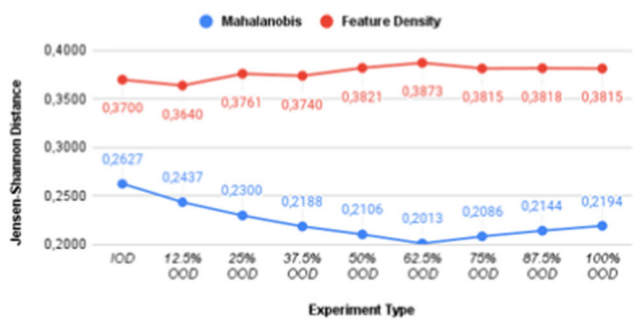


**Fig. 3** Accuracy values in the validation and testing stages for the different implemented architectures. In the left image, it can be seen that the values of each iteration are connected between them with lines to better compare the results, but this does not mean that the

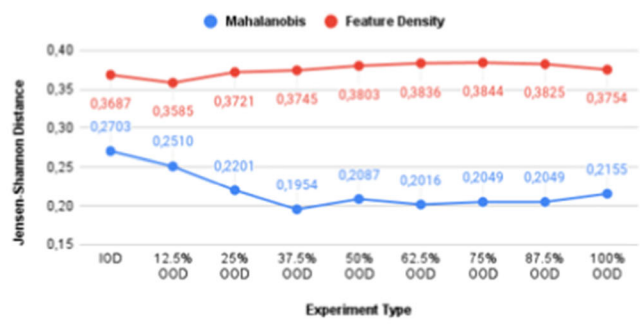
results are related. Analogously, in the right image, the values of each experiment type are connected between them with lines to better compare the results, but this does not mean that the results are related



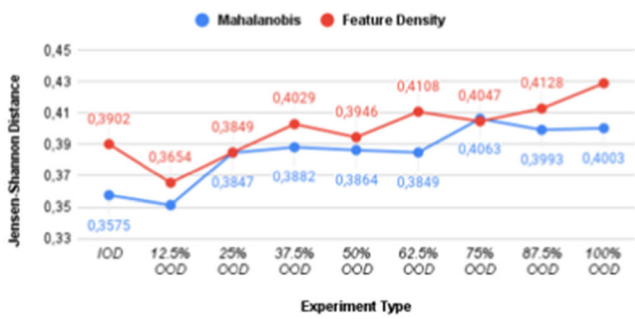
Mean value of the Jensen-Shannon distance for an AlexNet architecture



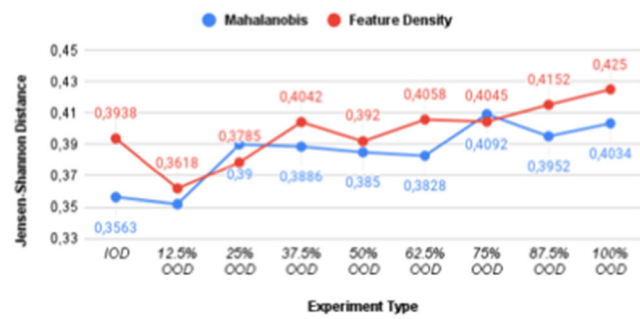
Median value of the Jensen-Shannon distance for an AlexNet architecture



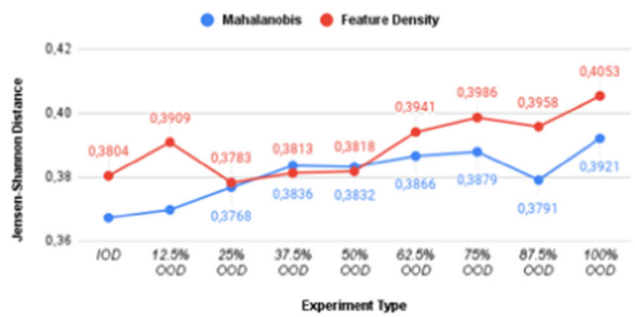
Mean value of the Jensen-Shannon distance for an DenseNet architecture



Median value of the Jensen-Shannon distance for an DenseNet architecture



Mean value of the Jensen-Shannon distance for an MobileNet architecture



Median value of the Jensen-Shannon distance for an MobileNet architecture

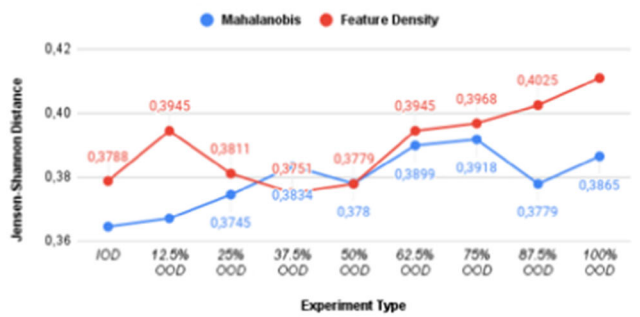


Fig. 4 Performance of the JS Distance for the different tested architectures. Each row reports the performance of AlexNet, DenseNet, and MobileNet models, respectively. The values of each

experiment type are connected between them with lines to better compare the results, but this does not mean that the results are related

architecture. Despite this, with the AlexNet architecture, there is no significant difference between the values of the Jensen–Shannon distance when using only IOD images and when using only OOD images; differences, however, can be seen in the results using the other two architectures; therefore, it cannot be concluded that the AlexNet network combined with the Feature Density method is the best option as an uncertainty estimator model.

the convolutional layers of the architecture, the execution time and computational cost will increase. Despite the results proposed by this work, the Feature Density method should not be left aside as an estimation model of uncertainty, there may be contexts where its performance is better.

Considering the execution time and the computational cost that must be used to estimate the uncertainty using the Feature Density method, in some contexts it may be more feasible to use the Mahalanobis Distance method. It is necessary to highlight that depending on the complexity of

As recommendations for future lines of work, it is proposed to perform experiments with other data sets, looking for a context where Feature Density performance might be better; and to investigate mathematical models that can be used as uncertainty estimation methods and that their dependence is not directly related to the type of network used for image processing.

**Acknowledgements** This work was partially supported by the Autonomous Government of Andalusia (Spain) under project UMA20-FEDERJA-108, the Ministry of Science and Innovation of Spain grant number PID2022-136764OA-I00, and the University of Málaga (Spain) under grants B4-2022, B1-2019\_01, B1-2019\_02, B1-2021\_20 and B1-2022\_14. The authors acknowledge the grant of the Universidad de Málaga and the Instituto de Investigación Biomédica de Málaga y Plataforma en Nanomedicina-IBIMA Plataforma BIO-NAND. Funding for open access charge: Universidad de Málaga/CBUA.

**Funding** Funding for open access publishing: Universidad Málaga/CBUA.

**Data availability** Datasets used in the manuscript can be found at: <https://doi.org/10.1016/j.acra.2011.09.014>, <https://doi.org/10.1038/sdata.2017.177>, <https://doi.org/10.1007/s11517-021-02497-6> All other data are available from the authors upon reasonable request.

## Declarations

**Conflict of interest** The authors declare that they have no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.


## References

- Iliadis L, Magri L (2022) Special issue on deep learning modeling in real life: anomaly detection, biomedical, concept analysis, finance, image analysis, recommendation. *Neural Comput Appl* 34:19397–19400
- Calderon-Ramirez S, Yang S, Moemeni A, Colreavy-Donnelly S, Elizondo DA, Oala L, Rodríguez-Capitán J, Jiménez-Navarro M, López-Rubio E, Molina-Cabello MA (2021) Improving uncertainty estimation with semi-supervised deep learning for Covid-19 detection using chest x-ray images. *IEEE Access* 9:85442–85454
- Wild C, Weiderpass E, Stewart B (2020) World cancer report: cancer research for cancer prevention. International Agency for Research on Cancer, Lyon, France
- Society AC, Society (2022) Breast cancer facts and figures 2022. American Cancer Society, Atlanta
- Calderón Ramírez S, Murillo-Hernández D, Rojas-Salazar K, Elizondo D, Moemeni A, Molina-Cabello MA (2022) A real use case of semi-supervised learning for mammogram classification in a local clinic of Costa Rica. *Med Biol Eng Comput* 60(4):1159–1175
- Calderon-Ramirez S, Oala L, Torrents-Barrena J, Yang S, Moemeni A, Samek W, Molina-Cabello MA (2020) Mixmood: A systematic approach to class distribution mismatch in semi-supervised learning using deep dataset dissimilarity measures. arXiv preprint [arXiv:2006.07767](https://arxiv.org/abs/2006.07767)
- Calderon-Ramirez S, Yang S, Elizondo D, Moemeni A (2021) 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](https://arxiv.org/abs/2109.00889)
- Weiss K, Khoshgoftaar TM, Wang D (2016) A survey of transfer learning. *J Big Data* 3(1):1–40
- Oliver A, Odena A, Raffel C, Cubuk ED, Goodfellow IJ (2018) Realistic evaluation of deep semi-supervised learning algorithms. *CoRR* abs/1804.09170
- Sun W, Tseng B, Zhang J, Qian W (2016) Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Comput Med Imaging Graph*. <https://doi.org/10.1016/j.compmedimag.2016.07.004>
- Nanni L, Paci M, Brahnam S, Lumini A (2022) Feature transforms for image data augmentation. *Neural Comput Appl* 34(24):22345–22356
- Shorten C, Khoshgoftaar TM (2016) A survey on image data augmentation for deep learning. *Comput Med Imaging Graph*. <https://doi.org/10.1016/j.compmedimag.2016.07.004>
- van Engelen JE, Hoos HH (2019) A survey on semi-supervised learning. *Mach Learn* 109:373–440
- Tardy M, Scheffer B, Mateus D (2019) Uncertainty measurements for the reliable classification on mammograms. In: Springer: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp 495–503
- Jøsang A (2016) Subjective logic: a formalism for reasoning under uncertainty. International series of monographs on physics. Springer, Cham, Switzerland
- Denouden T, Salay R, Czarnecki K, Abdelzad V, Phan B, Vernekar S (2018) Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. *CoRR* abs/1812.02765
- Calderón-Ramírez S, Murillo-Hernández D, Rojas-Salazar K, Calvo-Valverd L-A, Yang S, Moemeni A, Elizondo D, López-Rubio E, Molina-Cabello MA (2021) Improving uncertainty estimations for mammogram classification using semi-supervised learning. In: 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, pp 1–8
- Gal Y, Ghahramani Z (2016) Dropout as a Bayesian Approximation: representing model uncertainty in deep learning. In: International conference on machine learning 2016 Jun 11 (pp 1050-1059). PMLR
- Bansal M, Kumar M, Sachdeva M, Mittal A (2021) Transfer learning for image classification using vgg19: Caltech-101 image data set. *J Ambient Intell Hum Comput*. <https://doi.org/10.1007/s12652-021-03488-z>
- Fuentes-Fino RJ, Calderón-Ramírez S, Domínguez E, López-Rubio E, Hernandez-Vasquez MA, Molina-Cabello MA (2022) Feature density as an uncertainty estimator method in the binary classification mammography images task for a supervised deep learning model. In: International Work-Conference on Bioinformatics and Biomedical Engineering. Springer, pp 375–388
- Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS (2012) Inbreast: toward a full-field digital mammographic database. *Acad Radiol* 19(2):236–248
- Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin D (2017) A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data* 4(1):1–9
- Beeravolu AR, Azam S, Jonkman M, Shanmugam B, Kannoopatti K, Anwar A (2021) Preprocessing of breast cancer images to create datasets for deep-CNN. *IEEE Access* 9:33438–33463

24. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Curran Associates Inc
25. Gao H, Liu Z, van der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. IEEE Computer Society. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp 4700-4708)
26. Howard A, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. Google Inc., Menlo Park

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Authors and Affiliations

Ricardo Fuentes-Fino<sup>1</sup> · Saúl Calderón-Ramírez<sup>2</sup> · Enrique Domínguez<sup>1,4</sup> · Ezequiel López-Rubio<sup>1,4</sup> · David Elizondo<sup>3</sup> · Miguel A. Molina-Cabello<sup>1,4</sup> 

✉ Miguel A. Molina-Cabello  
miguelangel@lcc.uma.es

Ricardo Fuentes-Fino  
RicardoFino@uma.es

Saúl Calderón-Ramírez  
sacalderon@itcr.ac.cr

Enrique Domínguez  
enriqued@lcc.uma.es

Ezequiel López-Rubio  
ezeqlr@lcc.uma.es

David Elizondo  
elizondo@dmu.ac.uk

<sup>1</sup> Department of Computer Languages and Computer Science, University of Málaga, Málaga, Spain

<sup>2</sup> Pattern Recognition and Machine Learning, Instituto Tecnológico de Costa Rica, Cartago, Costa Rica

<sup>3</sup> School of Computer Science and Informatics, De Montfort University, Leicester, UK

<sup>4</sup> Instituto de Investigación Biomédica de Málaga y Plataforma en Nanomedicina-IBIMA Plataforma BIONAND, Málaga, Spain