Methodology for the integration of different deep learning-based models for the detection of lesions in screening mammography

David Esteban Ortega Figueroa and Erika Yesenia Suárez Bonilla

Degree Work to opt for the degree of Electronic Engineer

Director

Said Pertuz

Ph.D in Computer Science

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones

Bucaramanga

2023

Acknowledgments

This work has been possible thanks to the collaborative effort of many people. We thank the members of the CPS research group for their guidance and assistance throughout the research process. We also thank our director Said Pertuz for offering us his support, guidance, ideas, and experience, which have been fundamental aspects of this process. Thanks to our professors and the Universidad Industrial de Santander for providing us with the tools and knowledge to achieve this academic objective. Thanks to the "Young Researchers" grant from MINCIENCIAS for financially supporting part of this work. Finally, thanks to our family and friends for their support, encouragement and love that have been our main strength and motivation. Thank you all for being a part of our academic journey.

Contents

List of Figures

List of Tables

Abstract

Title: Methodology for the integration of different deep learning-based models for the detection of lesions in screening mammography [*]

Authors: David Esteban Ortega Figueroa and Erika Yesenia Suárez Bonilla. [**]

Keywords: Screening mammography, breast cancer, artificial intelligence, detection models, deep learning.

Description: Deep learning models have shown potential for breast cancer detection in screening mammography. To date, many learning architectures and training methods exist to perform detection, classification and/or localization tasks. However, the implementation and validation of these models raises serious concerns about the reproducibility of results. In this work, we are interested in the integration of different deep learning-based models for the detection of breast cancer in screening mammography. For this purpose, we build a dataset unrelated to the training data distribution to analyze the reproducibility of the models, and to perform different experiments for the integration methodology. This work includes three important stages. First, a benchmark of 5 screening models for breast cancer detection is performed. Second, a saliency analysis is performed to study the relationship between the detection and lesion localization. And finally, integration methods are tested based on the individual predictions of each model. Our results suggest that saliency detection is not an accurate predictor of an algorithm's performance. For this reason, the proposed integration methodology was based solely on processing the individual predictions of each model. Thus, the best of the results was obtained from logistic regression using cross-validation, with a performance of AUC = 0.841 (95% CI: 0.759 - 0.923) on breast-level predictions.

---

[*] Bachelor's Thesis

[**] Facultad de Ingeníeras Fisicomecánicas, Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones. Director: Said Pertuz, PhD in Computer Science.

Resumen

Titulo: Metodología para la integración de diferentes modelos basados en deep learning para la detección de lesiones en mamografía de cribado. *

Autores: David Esteban Ortega Figueroa y Erika Yesenia Suárez Bonilla. **

Palabras claves: Mamografía, cáncer de mama, inteligencia artificial, modelos de detección, aprendizaje profundo.

Descripción: Los modelos de aprendizaje profundo han demostrado su potencial para la detección del cáncer de mama en mamografías de cribado. Existen muchas arquitecturas de aprendizaje y métodos de entrenamiento que realizar tareas de detección, clasificación y/o localización. Sin embargo, la implementación y validación de estos plantea serias dudas sobre la reproducibilidad de los resultados. Este trabajo, se intereza en la integración de diferentes modelos para la detección de cáncer de mama. Para ello, construimos un conjunto de datos independiente, para analizar la reproducibilidad de los modelos, y realizar diferentes experimentos para su integración. Este trabajo incluye tres etapas. Primero, se realiza una evaluación comparativa de 5 modelos de cribado para la detección. Segundo, se realiza un análisis de saliencia para estudiar la relación entre las capacidades de detección y localización de lesiones de estos. Y por último, se prueban métodos de integración. Los resultados sugieren que el desempeño en saliencia no es un predictor preciso del rendimiento de un algoritmo en la detección de cáncer de mama. Por esto, la integración se basó únicamente en el procesamiento de las predicciones individuales de cada modelo. Así, el mejor de los resultados se obtuvo con regresión logística mediante validación cruzada, con un rendimiento de AUC = 0.841 (IC 95%: 0.759-0.923) a nivel de mama.

---

* Trabajo de grado

** Facultad de Ingenierías Fisicomecánicas, Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones. Director: Said Pertuz, PhD in Computer Science.

Introduction

Cancer is one of the leading causes of death in the world, which makes it one of the most dangerous and frightening diseases with great relevance in the field of medicine. In fighting cancer, the main aim is to enhance cancer control and reduce avoidable deaths by emphasizing health promotion, timely diagnosis, and access to medical care (WHO, 2022). In this regard, in 2015, the United Nations adopted the objective of sustainable development in health and well-being, which aims to “ensure healthy lives and promote well-being for all at all ages" (SDG, 2019), where one of the goals for the fight against noncommunicable diseases such as cancer is that by 2030, there should be a one-third reduction in premature mortality. (Hák et al., 2016).

Among the wide variety of cancers, breast cancer is one of the most common with high cancer-related mortality and morbidity rates in women. In 2020, it was the most diagnosed type of cancer with more than 2.2 million new cases and 685.000 deaths globally. According to data provided by the World Health Organization, about 1 in 12 women is expected to develop breast cancer in her lifetime, with the majority of cases and deaths occurring in low- and middle-income countries (WHO, 2021).

Early detection is a pivot step in the fight against breast cancer, allowing to start a timely treatment and increasing patient’s survival probabilities (Humphrey et al., 2002). A variety of studies support that screening mammography is a very effective method for the diagnosis of breast cancer (Wei et al., 2011), allowing the identification of abnormal masses by a radiologist, who is in charge of indicating the procedure to follow depending on his interpretation of the image. Within

this principle, the role of artificial intelligence and medical imaging is based on the development of screening models that allow the identification of lesions or signs of the disease from mammograms (Mendelson, 2019). This way, fewer cancers could be missed because an AI algorithm is not affected by fatigue or subjective factors, and timely diagnosis could be possible without the need for additional medical tests (Stadnick et al., 2021).

Currently, there are multiple deep learning models in screening mammography for which promising results have been obtained. They vary widely in topology, training methods, and validation. However, before the adoption of artificial intelligence in clinical practice, there are some limitations that prevent these screening tools from being available to radiologists, such as methodological flaws and/or underlying biases that result in difficulties in the reproducibility and generalization processes (Varoquaux and Cheplygina, 2022),(Freeman et al., 2021). Specifically, these difficulties include the bias in small datasets that are not fully representative, the variability of large internationally sourced datasets, the incorrect integration of data, the necessity for clinicians and data analysts to work side-by-side, among others (Roberts et al., 2021),(Mongan et al., 2020).

AI research in screening mammography has attempted to tackle this set of issues. In each case, turning to the implementation of different strategies that can potentially improve the characteristics and quality of radiomics studies and, consequently, obtain better performance and robustness of the models. Along the same lines, this research work seeks to analyze whether merging the individual results of a group of different detection models into a single one, can lead to better results on a mammogram dataset that is not related to the original image samples used in the development of the models.

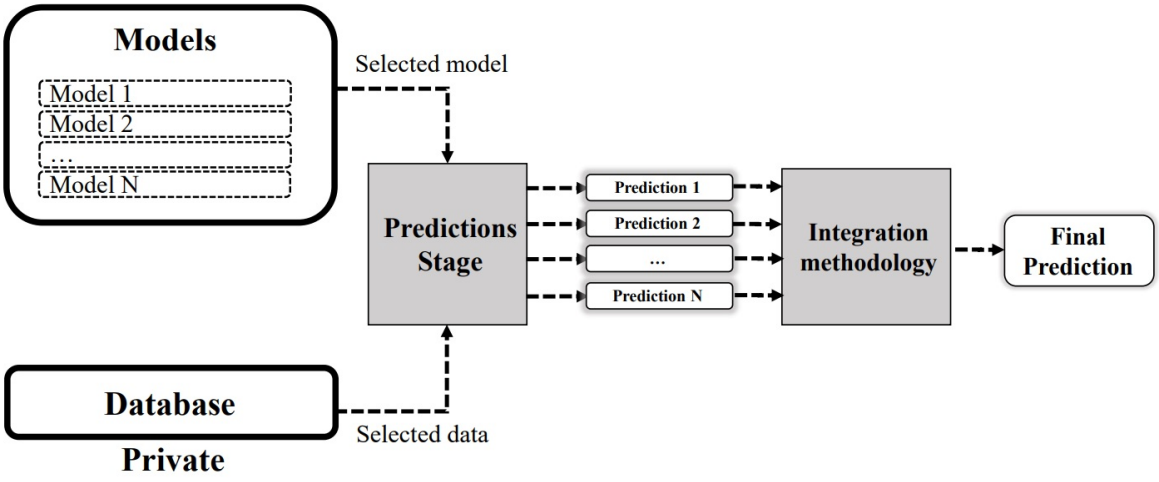
1. Problem statement

Currently, advances in artificial intelligence allow its application in medical imaging, being considered one of the most promising and innovative areas in the health sector (Lakhani et al., 2018). This has led to an increase in the number of published research articles, and thus a large number of screening models have been developed and studied in an attempt to make the process of detecting abnormal breast masses easier and faster (Pesapane et al., 2018). Many of these models, are not suitable for clinical use, since they do not satisfy the criteria to be used in clinical practice.

As reproducibility is so important in clinical practice, recent efforts seek to accelerate the process, making available and easier to validate models through model-repositories (Stadnick et al., 2021). These repositories provide access to their codes and make it possible to reproduce them in order to observe and evaluate the results of their performance in different databases, whose characteristics may vary widely from those used by their developers.

Figure 1

Process scheme



As many models are available, beyond validation, the combination of different models can be evaluated. However, there is no a clear way regarding how to combine them. This is why this research work seeks to propose a methodology for the integration of results from different models, as can be seen in Figure 1, where from an initial stage of predictions, involving a set of unrelated models with a private database, a single final prediction is obtained with the help of the proposed methodology.

2. Objectives

2.1. General Objective

To propose, implement and evaluate the performance of a methodology for integrating independent screening mammography lesion classifiers based on deep learning.

2.2. Specific Objectives

- To prepare a mammography dataset for screening mammography.
- To select and benchmark screening mammography classifiers.
- To propose a methodology to combine the considered screening mammography classifiers.
- To implement and evaluate the proposed methodology on the selected models and database.

3. Literature review

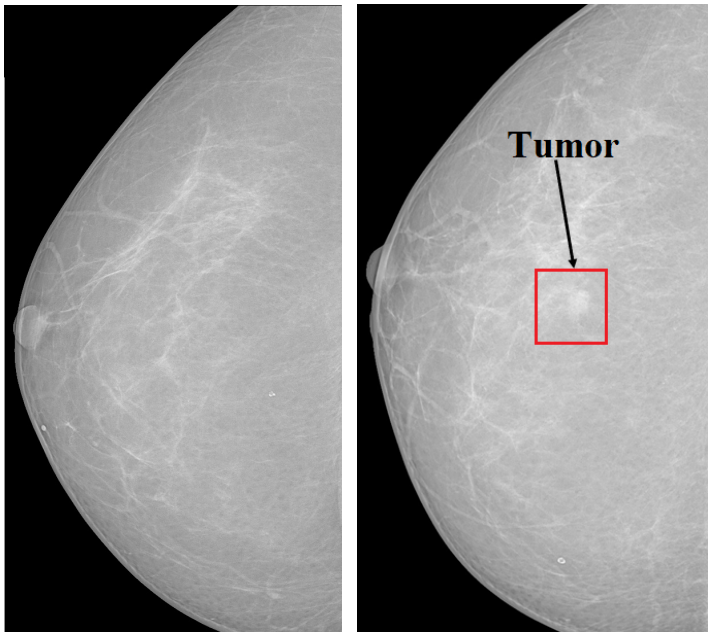
Preliminary research has demonstrated the relevance of early detection of breast cancer lesions. Therefore, an overview of different machine learning algorithms for the detection and localization of lesions in screening mammography is shown in this section. In addition, algorithms and studies related to ensemble methods are described.

3.1. Review of classification models

Mammography is an imaging modality in which breast tissue is visualized as shown in figure 2, using a low-dose x-ray device and a digital detector. It is fast, noninvasive and requires no recovery time. In this examination, a medical specialist can see the presence (Figure 2(b)) or absence (Figure 2(a)) of an abnormality in the breast tissue. A variety of studies support that mammography is a very effective method for the early detection of breast cancer (Gastounioti et al., 2016). In addition, it has been shown that an early diagnosis substantially improves survival and reduces mortality due to this disease (Byrne et al., 1994).

In the literature, there are several deep learning-based architectures used for screening mammography with different characteristics that generate a prediction of the presence of a malignant lesion in response, such as: Faster R-CNN (Ribli et al., 2018), this model is based on a popular object detection architecture which identifies suspicious areas in the image and sends them to the classification section where it is identified as benign, malignant or negative. MammoScreen(FDA, Approval document for MammoScreen , 2020), software that identifies suspicious focal findings of breast cancer in FFDM, applying algorithms trained on large databases, has as its output marks on

Figure 2
Diagnostic mammograms



(a) Healthy woman (b) Breast cancer patiens

what was found in the mammography and the level of suspicion. GMIC (Globally-Aware Multiple Instance Classifier) (Shen et al., 2021), first locates regions of interest in the whole image through a global network, then these regions are processed in a local network and finally fuses the information from both branches to perform the prediction. GLAM (Global-Local Activation Maps)(Liu et al., 2021), its architecture is similar to that of GMIC except that its global network is different and it was designed to produce high-resolution segmentation. End2end (Shen et al., 2019), its architecture performs classification on small patches of the image and then extends the classifier to the entire image. Lunit INSIGHT MMG(FDA, Approval document for Lunit INSIGHT MMG, 2021), software based on an artificial intelligence algorithm that helps in the detection, localization and characterization of suspicious areas of breast cancer. DMV-CNN (Deep multi-view convolu-

tional neural network)(Wu et al., 2019), uses in its architecture the four simultaneous views of the image to do its classification, however, a single image version of the model is also available to evaluate in case of missing data.

Another previous research work is the Meta-repository for screening mammography classifiers (Stadnick et al., 2021). This packages the Faster R-CNN (Ribli et al., 2018), GMIC (Shen et al., 2021), DMV-CNN (Wu et al., 2019), GLAM (Liu et al., 2021) and End2end model (Shen et al., 2019) run codes into a Docker image, making it easy to implement each of them with any database and generating as output the simulated model predictions and ROC curve, and the PR curve as a performance measure.

In the literature, there have been studies showing the potential of logistic regression, principal component analysis, k-means clustering, among others to improve the performance of algorithms involving several inputs and one output (Zhu et al., 2019) (Zhang et al., 2018).

3.2. Review of ensemble methods

In this section we review previous works related to ensemble methods in deep learning models and breast cancer. These methods can be used to integrate the predictions generated by the models, thus achieving a global prediction that takes into account the results provided by each one individually.

A previous work on this is "Ensemble deep learning system for early breast cancer detection" (Hekal et al., 2022). In this research a new deep learning system for early breast cancer detection is presented, this uses a set of four convolutional neural networks (CNN) transfer learning. Importantly, this system only uses the suspected nodule regions (SNR) instead of the whole

image, these SNR are extracted using an optimal dynamic thresholding method. By processing only the regions, the system achieves better performance and is able to detect small-sized nodules. A notable highlight of this work is that it includes a binary support vector machine (SVM) that tracks each CNN model and provides a malignant or benign score. A first-order impulse is used to obtain the final decision of the system, which is guided by the training accuracies of the four CNNs. The proposed joining scheme has been tested on region of interest (ROI) images using a public subset of digital medical images, using the public subset of digital medical images in which it has achieved an accuracy of 94% for distinguishing between malignant (M) and benign (B) classes and 95% for distinguishing between malignant (MM) and benign (BM) mass nodules. This validates the accuracy and advantages of this system.

Other works related to ensemble methods are: "Reviewing ensemble classification methods in breast cancer" (Hosni et al., 2019) which shows a general review of the state of the art where several single classification techniques were used for constructing ensemble methods. This review found that in an ensemble classification task, artificial neural networks, decision trees and support vector machines are most frequently used, also, it is highlighted that the most frequently used validation method in conducting experiments is K-fold cross-validation. In "Breast cancer diagnosis using imbalanced learning and ensemble method" (Cai et al., 2018), a model is proposed combining ensemble method and imbalanced learning technique for the classification of breast cancer data, it applies an unbalanced learning algorithm to selected data, then optimizes several classifiers with Bayesian optimization, and finally combines the optimized classifiers for a final decision with an ensemble stacking method. The proposed model shows better performance and adaptabil-

ity than conventional methods in terms of classification accuracy, specificity and AuROC on two breast cancer datasets. In "Design ensemble machine learning model for breast cancer diagnosis" (Hsieh et al., 2012) a combined ensemble model with three schemes (Neural fuzzy, k-nearest neighbor and quadratic classifier) has been constructed for further validations, in this research, the results show that the potential of joint learning with respect to individual models, supporting that the combined model has the highest accuracy in injury classification. In "Breast cancer detection using an ensemble deep learning method" (Das et al., 2021), they use the effective conversion of one-dimensional data into images and propose a design of a stacked ensemble deep learning model that can increase the classification accuracy performance compared to individual models. In this paper presents a two-stage classification, with three convolutional neural networks as base classifiers in the first stage y then, the first stage results are used to train the second stage classifier "Multilayer Perceptron". It uses a set of breast images to train and validate the proposed model.

4. Matherials and methods

This section describes the dataset constructed and the screening models studied for this work. In addition, the description of concepts, data and evaluation metrics used in the development of the project is presented.

4.1. Imaging data

The dataset was collected from the screening mammography program of Tampere University Hospital, which is responsible for screening the inhabitants of two municipalities in the Pirkanmaa region of Finland. All the imaging data are from the years 2015 to 2017 and corresponds to 607 women, of which 277 belong to patients with breast cancer (cases) and 330 to healthy women (controls). The use of the data has been approved in compliance with local and national laws and regulations.

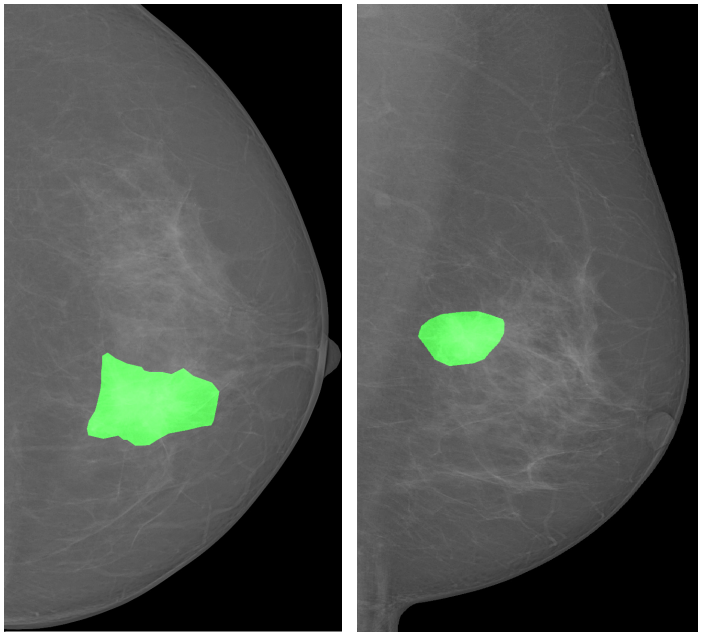
Only data from cases belonging to patients at the time of cancer diagnosis were considered and were matched in case-control relationship by mammography systems, birth years and screening years. Consequently, the final dataset corresponds to 382 women, of which 191 are cases and 191 are controls. In this work, we use bilateral two-view cranio-caudal (CC) and mediolateral oblique (MLO) full-field digital mammography images and all patients included had been diagnosed with unilateral cancers. Therefore, the study dataset includes a total of 1528 mammograms.

For cancer cases in the dataset, there is information about manual segmentations performed by expert radiologists, indicating the location of the malignant lesions, as shown in Figure 3. These are known as: ground-truth segmentations, and they are available for 370 (183 in CC view and 187

in MLO view) out of the 382 images corresponding to the cancer cases.

Figure 3

An instance of ground-truth segmentations for a patient



(a) CC mammography (b) MLO mammography

4.2. Screening models

The consider models are selected from publicly available in the Meta-repository of screening mammography classifiers (Stadnick et al., 2021). Some details of each model are presented below.

- FRCNN: (Faster R-CNN) This model is based on the identification of suspicious areas of the image, which are sent to the classification section, and it is there where it is cataloged as benign, malignant or negative. It was proposed in the paper "Detecting and classifying lesions in mammograms with Deep Learning" in 2018(Ribli et al., 2018), reporting an AUC ROC = 0.95.

- END2END: Its architecture performs classification on small patches of the image and then extends the classifier to the entire image. It was described in the paper "Deep Learning to Improve Breast Cancer Early Detection on Screening Mammography" in 2019 (Shen et al., 2019), reporting an AUC ROC = 0.95.

- DMV-CNN: (Deep Multi-View Convolutional Neural Network) This model is used for breast cancer classification as described in the paper "Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening" published in 2019 (Wu et al., 2019), reporting an AUC ROC = 0.89.

- GMIC: (Globally-Aware Multiple Instance Classifier) This model first locates the regions of interest in the whole image through a global network, then these regions are processed in a local network and, finally, merges the information from both branches to perform the prediction. It was described in the paper "An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization" in 2021 (Shen et al., 2021), reporting an AUC ROC = 0.93.

- GLAM: (Global-Local Activation Maps) Its architecture is similar to that of GMIC except that its global network is different, and it was designed to produce high-resolution segmentations. It was proposed in the paper "Weakly-supervised High-resolution Segmentation of Mammography Images for Breast Cancer Diagnosis" in 2021 (Liu et al., 2021), reporting an AUC ROC = 0.88.

The selected models are open source so their execution is made from the repositories avail-

able by the authors, their implementation is in python using in the case of FRCNN the caffe library, in End2end the keras library and in DMV-CNN, GMIC and GLAM the pythorch library.

To obtain the detection predictions of the deep learning models, the mammographic images in .png format and a pickle file with a list of dictionaries containing the information of the screening mammograms are sent to them. In addition, in the case of End2end a declaration of the average intensity of the pixel was made, because this is a necessary parameter for its execution.

4.3. Saliency methods

In image processing, saliency is a tool that allows describing the salient or important pixel regions of an image for a given job. In deep learning, heatmaps are an important tool, as they are a visual description, showing the parts of an image that significantly influenced the prediction of the model, thus allowing to interpret the process and to see the strengths and weaknesses when making a decision.

Previous related research is 'Benchmarking saliency methods for chest X-ray interpretation' (Saporta et al., 2022), in this paper seven saliency methods are evaluated quantitatively, on multiple neural network architectures using two evaluation metrics. They first establish a human benchmark for chest X-ray segmentation and examine under which clinical conditions saliency maps might be more likely to fail to localize important pathologies compared to a human expert benchmark. The results obtained were: 1. although one of the evaluated methods (Grad-CAM) generally localized pathologies better with respect to the other salience methods, all seven have a notably worse performance compared to human annotation. 2. The difference in localization performance between the human reference and Grad-CAM was greatest for pathologies of smaller size and more complex

shapes. 3. Model confidence was positively correlated with Grad-CAM localization performance. Based on the results, this work demonstrates that several important limitations of saliency methods should be considered before relying on them for deep learning explainability in medical imaging.

4.3.1. Heatmaps. As mentioned above, heatmaps make it possible to identify the parts of an input image that most influence the final prediction. The implementation of the DMV-CNN, GMIC and GLAM models allows the direct generation of pixel-level heatmaps from their own architecture. In the case of END2END, the GRAD-CAM algorithm is used, which from the gradients of the classification identifies the places where the final score depends more on the data, i.e. the places where the gradient is higher (Selvaraju et al., 2017).

Before generating heat maps, for each model, a sweep of the total set of images is performed to obtain the maximum pixel value obtained, and thus standardize the map in a range of values from 0 to 1.

4.3.2. Thresholding. For each image belonging to a breast with cancer, heatmaps are generated, one for each screening model, and then thresholding is applied to generate binary segmentations. The two thresholding techniques employed are shown below:

- Otsu's method: This algorithm works by searching iteratively for the threshold value that optimally separates the image into two classes, the foreground and the background, it does it by maximizing a metric called the interclass pixel intensity variance (Bangare et al., 2015).

- Iterative thresholding: Thresholding method in which a threshold value that maximizes the MIoU with respect to the ground-truth segmentation provided by the radiologists is itera-

tively searched for.

4.4. Integration of models

This section describes two methods proposed to obtain a global prediction, which integrates the predictions generated by the deep learning-based models.

4.4.1. Logistic regression. Logistic regression is a statistical method based on one or more input variables from which a efficient binary result is predicted in the output. It is a generalized linear model, in which the independent variables x_i together with a set of parameters w_i build a model (1), which for a new data can succeed in estimating the probability of outcome. The parameters w_i in the model indicate the weight and type of relationship between these independent or input variables and the dependent or output variable. A noteworthy feature of this model is that it makes use of the sigmoid function (2) that models binary output probability y, where a 0 represents one class and a 1 another class.

$$z = w_1x_1 + w_2x_2 + w_3x_3 + ... + w_Nx_N + b \tag{1}$$

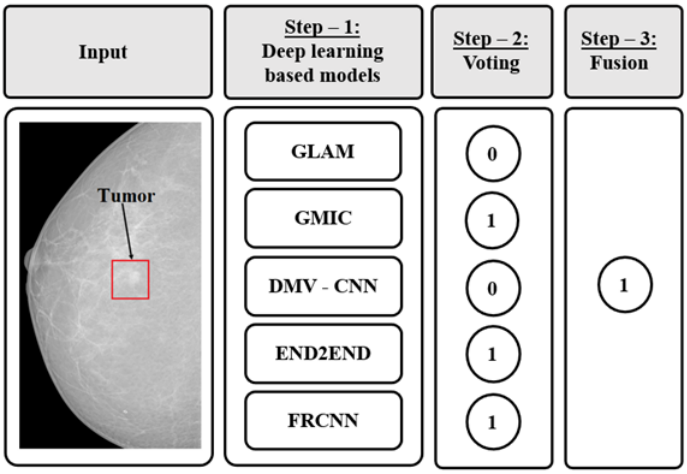
$$y = \frac{1}{1+e^{-z}} \tag{2}$$

Logistic regression was used in this work because it allows the calculation of a total prediction (dependent variable) from the predictions previously obtained by each of the classifiers (independent variables).

4.4.2. bootstrap aggregating. Bootstrap aggregation, also known as bagging, is a statistical method widely used in model ensemble. It allows combining several models by averaging in cases of regression-related problems, or voting for most of their predictions in classification cases, in order to improve the overall prediction. This algorithm helps to reduce the variance of models when trained on different data sets.

Figure 4

Bootstrap aggregating example



Bootstrap aggregating was used as a technique to obtain an overall prediction from voting classifiers based on deep learning, an example of which is shown in Figure 4.

4.5. Performance measurements

To measure detection performance, the Receiver Operating Characteristic (ROC) curve was used. This is a graphical representation showing the performance of a binary classification system as the decision threshold varies. This graph is constructed from the true positive rate (TPR) also known as sensitivity versus the false positive rate (FPR) also known as false alarm probability.

In this work we report the area under the ROC curve (AUC ROC) as it measures the ability to discriminate between two groups, this measure can vary between 0 and 1, where 1 would represent a perfect experiment, i.e. the higher the AUC, the higher the correct classification in the test for the two groups.

The 95% confidence interval (CI) is also reported, with which we can measure the uncertainty associated with the estimation of the parameter, since it gives us the range of values that probably contains the real parameter with a given confidence level.

K-fold cross-validation was used to achieve a reliable evaluation of the performance of the logistic regression's model. In this statistical method, the data are divided into k subsets or folds of the same size; its main idea is to train the model with K-1 folds and test with the remaining fold. This process is repeated k number of times, taking into account that the test fold is different in each test, when the k experiments are performed, the average of the results is calculated to obtain the final result.

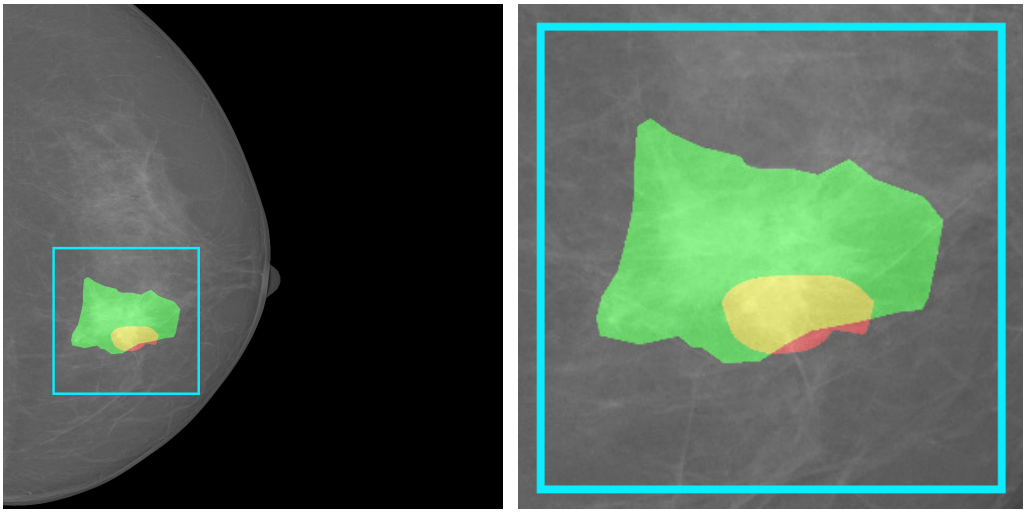
The Mean Intersection over Union (MIoU) is an evaluation metric that measures how much, on average, the segmentations obtained from the saliency methods overlapped with the ground-truth segmentations. The intersection over union (IoU) is obtained from the following equation:

$$IoU(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

In Figure 5 you can see graphically the regions that are taken into account for this evaluation metric. In green is the manual segmentation made by the expert radiologist, in red the area obtained by one

of the deep learning-based models after thresholding and in yellow the intersection between them.

Figure 5
Segmentation overlapping



(a) Breast with malignant lesion (b) Region of focus in segmentations

Finally, the MIoU is reduced to calculating the average IoU for the entire data set.

4.6. Design restrictions

A number of restrictions are taken into account in this research work and are described below. First, the processing of clinical data such as mammographic images, which are subject to the applicable personal data policies. Second, the use of models based on deep learning requires that the original models be made available in open source or that permissions for their use be granted prior to their use. Third, the compatibility of the algorithms with full-field digital mammography images.

5. Experiments and results

In this chapter we show the results obtained in each of the stages of this research. In section 5.1, we present the AUC ROC obtained by each of the models based on deep learning in the classification of the imaging data stated in 4.1, in section 5.2, we show the results obtained by analyzing the relationship between the predictions and the tumor location obtained from the four screening models considered and in section 5.3, the performance measures achieved in the integration of the predictions of the five learning models taken into account are shown.

5.1. A benchmark of the models

The implementation of the models was performed to subsequently evaluate their performance with the imaging data described in section 5.1 and the AUC ROC measure was used for the evaluation. The results obtained in this stage of detection are shown below:

Table 1
AUC ROC evaluation

Model	Image level		Breast level		Patient level	
	AUC ROC	95% CI	AUC ROC	95% CI	AUC ROC	95% CI
GLAM	0.525	0.491-0.558	0.528	0.480-0.576	0.529	0.471-0.587
GMIC	0.569	0.535-0.602	0.585	0.538-0.633	0.585	0.528-0.642
DMV - CNN	0.572	0.538-0.605	0.581	0.533-0.629	0.586	0.529-0.643
END2END	**0.694**	0.662-0.726	**0.750**	0.707-0.793	**0.747**	0.698-0.796
FRCNN	**0.796**	0.768-0.825	**0.832**	0.794-0.869	**0.778**	0.732-0.824

Table 1 shows the results on three sets of predictions, the first set corresponds to the individual image level predictions, the second set is constructed from the average of the predictions

of the two views of each breast (CC and MLO) and the third set corresponds to the average of the four views of patients' mammograms.

In general terms, it is analyzed from the results obtained that all the models present drawbacks in the generalization of their performance when a data source external to the data distribution with which they were trained is used. However, it is observed that in detection the best results were obtained by the oldest models (FRCNN and END2END), and the worst result was obtained by the most recent model (GLAM), whose value is not statistically significant since it does not exceed 0.5 in the lower limit of the confidence intervals.

5.2. Saliency analysis

For this analysis, we selected images pertaining to breasts with cancer and for which ground-truth segmentations were available. These are described in section 4.1.

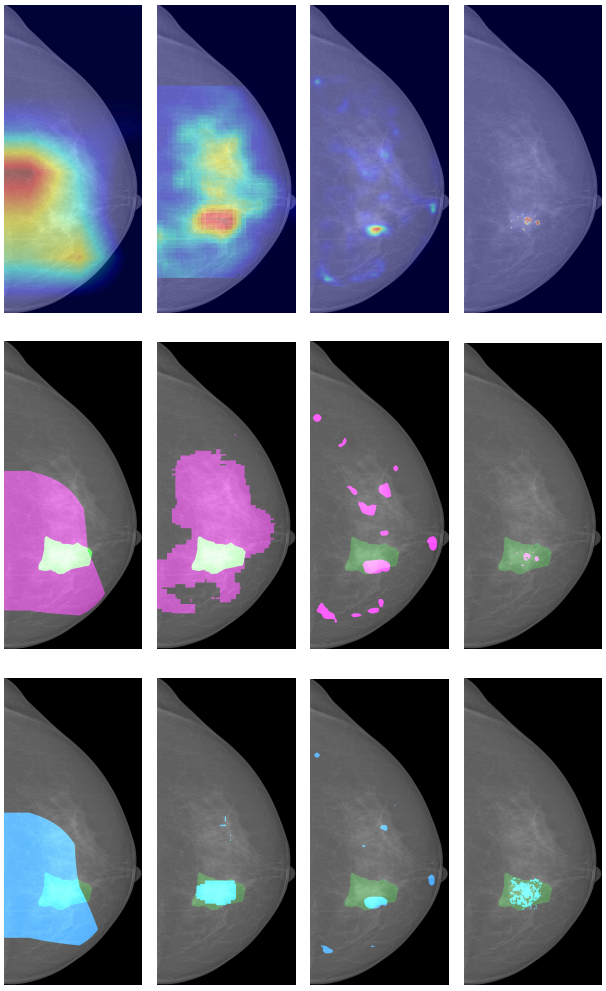
We ran 4 of the 5 models (It was not possible to extract the characteristics of the internal layers of FRCNN to generate heatmaps) with these images and obtained the heatmap for each (see first row Figure 6), then, from these we performed two independent sets of segmentations for each screening model: Otsu segmentations (see fuchsia regions second row Figure 6) and iterative thresholding segmentations (see blue regions third row Figure 6). These sets were compared with the ground-truth (see green region second and third row Figure 6) set to analyze the saliency localization performance for each model. The results obtained are shown in table 2, in the MIoU all regions section.

The values obtained in this evaluation were generally low. In the case of Otsu thresholding, GLAM (the most recent model) presented the highest MIoU with 0.0817 (8.17%), while with

iterative thresholding, the highest value was obtained by DMV-CNN with 0.2054 (20.54%), with a quite significant difference between the two thresholding methods. On the other hand, END2END, the model with the highest predictive evaluation, had the lowest MIoU values.

Figure 6

An instance of saliency analysis. First column: END2END; second column: DMV-CNN; third column: GMIC; fourth column: GLAM. First row: heatmaps; second row: overlapping of ground-truth (green) and otsu (fuchsia) segmentations; third row: overlapping of ground-truth (green) and iterative (blue) segmentations



In an attempt to improve on the results obtained previously, the MIoU evaluation was per-

formed by selecting only intersection regions. This step is applied in both thresholding methods, Otsu segmentations and iterative thresholding segmentations. First, only the images that after thresholding at some point intersect with the annotation given by the expert radiologist are taken into account, then, from these images are eliminated the areas that are highlighted but at no point intersect with the radiologist's annotation, i.e. the false positives, and finally, these are compared with the set of ground-truth segmentations, to analyze the saliency localization performance of each model.

The results obtained are shown in table 2, in the MIoU intersection regions section, where is observed that in localization the best result for both types of thresholding was obtained by GLAM with 0.2283 (22.83%) in Otsu thresholding and with 0.3588 (35.88%) in iterative thresholding. The lowest result for Otsu thresholding was obtained by DMV-CNN with 0.0413 (4.13%) and for iterative thresholding it was END2END with 0.0919 (9.19%).

Table 2
MIoU evaluation

Model	MIoU all regions		MIoU intersection regions	
	Otsu	Iterative	Otsu	Iterative
GLAM	**0.0817**	0.1285	**0.2283**	**0.3588**
GMIC	0.0786	0.1102	0.1977	0.2109
DMV - CNN	0.0323	**0.2054**	0.0413	0.3335
END2END	0.0218	0.0551	0.0552	0.0919

5.3. Integration of models

In this section we show the experiments performed from the methods mentioned in section 4.4 and their respective results in table 3, for which the AUC ROC and the confidence interval 95% were used as performance measures. In these experiments we used the predictions obtained from the deep learning-based models for the set of images described in section 4.1, and it was performed on image, breast and patient level predictions. Table 3 shows that the highest value obtained for the AUC ROC performance measure is achieved when the predictions are grouped by breasts for both methods tested.

Table 3
Results of integration methods

Set of predictions	Bootstrap aggregating		Logistic regression	
	AUC ROC	95% CI	AUC ROC	95% CI
Image	0.683	0.650-0.716	0.800	0.771-0.828
Breast	**0.704**	0.658-0.749	**0.841**	0.803-0.875
Patient	0.695	0.642-0.747	0.792	0.747-0.837

6. Discussion and conclusion

In this work we propose a methodology for the integration of deep learning-based models for the detection of lesions in screening mammography, this methodology includes three fundamental steps to obtain a global prediction:

- Build dataset. The creation of a mammography dataset is required to perform logistic regression and evaluate the integration results.
- Implement and run the models. At this stage it is necessary to have access to the algorithms and codes of the screening models. In this way, it will be possible to perform their correct implementation and subsequently obtain the cancer predictions for the built dataset.
- Perform logistic regression. Finally, from the predictions obtained and the labels of each mammogram (e.g. 1 for cases and 0 for controls) it will be possible to perform a logistic regression to adjust the group of predictions to a single prediction.

In the construction of this methodology, the importance of linking the predictions at the breast level is highlighted, since it increases the classification performance with respect to its individual management. Also, the comparison between the highest value obtained in the AUC ROC of the individual models (FRCNN) and the one achieved with the integration methodology shows that there are no statistically significant differences between these results.

In reference to lesion localization, the results show that a high value in the model detection performance measure does not guarantee that the model performs an effective salience analysis.

This behavior is observed when the models obtain a significantly lower performance in MIoU (table 2) than that obtained in the AUC ROC (table 1). In addition, it is worth mentioning that the results obtained in detection for all models differ notably, decreasing in all cases with respect to those reported in the papers published for each model, which shows the problem of reproducibility and generalization.

References

Bangare, S. L., Dubal, A., Bangare, P. S., and Patil, S. (2015). Reviewing otsu's method for image thresholding. *International Journal of Applied Engineering Research*, 10(9):21777–21783.

Byrne, C., Smart, C. R., Chu, K. C., and Hartmann, W. H. (1994). Survival advantage differences by age: evaluation of the extended follow-up of the breast cancer detection demonstration project. *Cancer*, 74(S1):301–310.

Cai, T., He, H., and Zhang, W. (2018). Breast cancer diagnosis using imbalanced learning and ensemble method. *Applied and Computational Mathematics*, 7(3):146–154.

Das, A., Mohanty, M. N., Mallick, P. K., Tiwari, P., Muhammad, K., and Zhu, H. (2021). Breast cancer detection using an ensemble deep learning method. *Biomedical Signal Processing and Control*, 70:103009.

FDA, Approval document for Lunit INSIGHT MMG (Published November 17, 2021). $\texttt{www.accessdata.fda.gov/cdrh}_{d}ocs/pdf21/K211678.pdf$.

FDA, Approval document for MammoScreen (Published March 25, 2020). $\texttt{www.accessdata.fda.gov/cdrh}_{d}ocs/pdf19/K192854.pdf$.

Freeman, K., Geppert, J., Stinton, C., Todkill, D., Johnson, S., Clarke, A., and Taylor-Phillips, S. (2021). Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. *bmj*, 374.

Gastounioti, A., Conant, E. F., and Kontos, D. (2016). Beyond breast density: a review on the advancing role of parenchymal texture analysis in breast cancer risk assessment. *Breast cancer research*, 18(1):1–12.

Hák, T., Janoušková, S., and Moldan, B. (2016). Sustainable development goals: A need for relevant indicators. *Ecological indicators*, 60:565–573.

Hekal, A. A., Moustafa, H. E.-D., and Elnakib, A. (2022). Ensemble deep learning system for early breast cancer detection. *Evolutionary Intelligence*, pages 1–10.

Hosni, M., Abnane, I., Idri, A., de Gea, J. M. C., and Alemán, J. L. F. (2019). Reviewing ensemble classification methods in breast cancer. *Computer methods and programs in biomedicine*, 177:89–112.

Hsieh, S.-L., Hsieh, S.-H., Cheng, P.-H., Chen, C.-H., Hsu, K.-P., Lee, I.-S., Wang, Z., and Lai, F. (2012). Design ensemble machine learning model for breast cancer diagnosis. *Journal of medical systems*, 36:2841–2847.

Humphrey, L. L., Helfand, M., Chan, B. K., and Woolf, S. H. (2002). Breast cancer screening: a summary of the evidence for the us preventive services task force. *Annals of internal medicine*, 137(5_Part_1):347–360.

Lakhani, P., Prater, A. B., Hutson, R. K., Andriole, K. P., Dreyer, K. J., Morey, J., Prevedello, L. M., Clark, T. J., Geis, J. R., Itri, J. N., et al. (2018). Machine learning in radiology: applications beyond image interpretation. *Journal of the American College of Radiology*, 15(2):350–359.

Liu, K., Shen, Y., Wu, N., Chłędowski, J., Fernandez-Granda, C., and Geras, K. J. (2021). Weakly-supervised high-resolution segmentation of mammography images for breast cancer diagnosis. *arXiv preprint arXiv:2106.07049*.

Mendelson, E. B. (2019). Artificial intelligence in breast imaging: potentials and limitations. *American Journal of Roentgenology*, 212(2):293–299.

Mongan, J., Moy, L., and Kahn Jr, C. E. (2020). Checklist for artificial intelligence in medical imaging (claim): a guide for authors and reviewers. *Radiology. Artificial Intelligence*, 2(2).

Pesapane, F., Codari, M., and Sardanelli, F. (2018). Artificial intelligence in medical imaging: threat or opportunity? radiologists again at the forefront of innovation in medicine. *European radiology experimental*, 2(1):1–10.

Ribli, D., Horváth, A., Unger, Z., Pollner, P., and Csabai, I. (2018). Detecting and classifying lesions in mammograms with deep learning. *Scientific reports*, 8(1):1–7.

Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., et al. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3):199–217.

Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S. Q., Nguyen, C. D., Ngo, V.-D., Seekins, J., Blankenberg, F. G., Ng, A. Y., et al. (2022). Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(10):867–878.

SDG, U. (2019). Sustainable development goals. *The energy progress report. Tracking SDG*, 7.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., and Sieh, W. (2019). Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, 9(1):1–12.

Shen, Y., Wu, N., Phang, J., Park, J., Liu, K., Tyagi, S., Heacock, L., Kim, S. G., Moy, L., Cho, K., et al. (2021). An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization. *Medical image analysis*, 68:101908.

Stadnick, B., Witowski, J., Rajiv, V., Chłędowski, J., Shamout, F. E., Cho, K., and Geras, K. J. (2021). Meta-repository of screening mammography classifiers. *arXiv preprint arXiv:2108.04800*.

Varoquaux, G. and Cheplygina, V. (2022). Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, 5(1):1–8.

Wei, J., Chan, H.-P., Wu, Y.-T., Zhou, C., Helvie, M. A., Tsodikov, A., Hadjiiski, L. M., and Sahiner, B. (2011). Association of computerized mammographic parenchymal pattern measure with breast cancer risk: a pilot case-control study. *Radiology*, 260(1):42.

WHO, W. H. O. (2021). Breast cancer. https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer.

WHO, W. H. O. (2022). Cancer. https://www.who.int/news-room/fact-sheets/detail/cancer.

Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzębski, S., Févry, T., Katsnelson, J., Kim, E., et al. (2019). Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE transactions on medical imaging*, 39(4):1184–1194.

Zhang, K., Geng, W., and Zhang, S. (2018). Network-based logistic regression integration method for biomarker identification. *BMC systems biology*, 12(9):113–122.

Zhu, C., Idemudia, C. U., and Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating pca and k-means techniques. *Informatics in Medicine Unlocked*, 17:100179.