

---

# A CNN-BASED METHODOLOGY FOR BREAST CANCER DIAGNOSIS USING THERMAL IMAGES

---

A PREPRINT

J. Zuluaga-Gomez<sup>a, d, e, \*</sup>, Z. Al Masry<sup>a</sup>, K. Benagoune<sup>b</sup>, S. Meraghni<sup>c</sup>, N. Zerhouni<sup>a</sup>

<sup>a</sup>FEMTO-ST institute, Univ. Bourgogne Franche-Comté, CNRS, ENSMM, Besançon, France

<sup>b</sup>Laboratory of Automation and Production engineering, Batna University, Batna, Algeria

<sup>c</sup>LINFI Laboratory, University of Biskra, Biskra, Algeria

<sup>d</sup>Electrical Engineering Department, University of Oviedo, Gijon, Spain

<sup>e</sup>Universidad Autonoma del Caribe, Barranquilla, Colombia

October 31, 2019

## Highlights:

- Efficiency and reliability for breast cancer diagnosis through thermography
- CNNs performance enhancement with data augmentation techniques
- Smaller and simpler CNNs architectures perform better than state-of-the-art CNNs
- Trade-off measurement between data augmentation and the database size

## ABSTRACT

**MicroAbstract:** A recent study from GLOBOCAN disclosed that during 2018 two million women worldwide had been diagnosed from breast cancer. This study presents a computer-aided diagnosis system based on convolutional neural networks as an alternative diagnosis methodology for breast cancer diagnosis with thermal images. Experimental results showed that lower false-positives and false-negatives classification rates are obtained when data pre-processing and data augmentation techniques are implemented in these thermal images.

**Background:** There are many types of breast cancer screening techniques such as, mammography, magnetic resonance imaging, ultrasound and blood sample tests, which require either, expensive devices or personal qualified. Currently, some countries still lack access to these main screening techniques due to economic, social or cultural issues. The objective of this study is to demonstrate that computer-aided diagnosis(CAD) systems based on convolutional neural networks (CNN) are faster, reliable and robust than other techniques. **Methods:** Despite the disadvantages of the traditional machine learning techniques with spatial information, CNNs stand as main techniques for pattern recognition on images -or thermal images-. We performed a study of the influence of data pre-processing, data augmentation and database size versus a proposed set of CNN models. Furthermore, we developed a CNN hyper-parameters fine-tuning optimization algorithm using a tree parzen estimator. **Results:** Among the 57 patients database, our CNN models obtained a higher accuracy (92%) and F1-score (92%) that outperforms several state-of-the-art architectures such as ResNet50, SeResNet50 and Inception. Also, we demonstrated that a CNN model that implements data-augmentation techniques reach identical performance metrics in comparison with a CNN that uses a database up to 50% bigger. **Conclusion:** This study highlights the benefits of data augmentation and CNNs in thermal breast images. Also, it measures the influence of the database size in the performance of CNNs. *MSC 2010:* 62P10 & 68T10

---

\*Corresponding author: Juan Pablo Zuluaga, ORCID: 0000-0002-6947-2706. Prognostics & Health Management Team, Femto-ST Sciences & Technologies, Besançon Cedex, France, 25000. Email: juan.zuluaga@eu4m.eu

**Keywords** Breast Cancer · Breast Thermography · Hyper-parameters Optimization · Convolutional Neural Network · Computer Aided Diagnosis Systems

## 1 Introduction

Breast cancer is the most commonly diagnosed cancer in women worldwide; then, it has become significant public health. It was the first leading cause of cancer-linked death among women in 2018, reaching approximately 15% of the total number of registered cases [1]. The early detection of breast cancer is imperative to reduce the mortality and morbidity index [2, 3, 4]. Some studies suggest that emerging economies have almost a double risk of cancer, where the mortality-to incidence ratio in developed countries is 0.20, but in less developed countries is almost twice, thus 0.37 [5, 6]. Other factors like socioeconomic [5, 7], aging, unhealthy lifestyle [7, 8, 6, 9], environmental issues, and growth of the population may perhaps lead to higher risks. In perspective, Li et al. [10] prove the correlation between body weight, parity, number of births, and menopausal status concerning breast cancer. On the other hand, some countries keep multiple barriers for developing an effective breast cancer screening system, e.g., organizational, psychological, structural, socio-cultural, and religious [11]. Physicians, self-examination, and imaging techniques can perform detection of abnormalities in the breast, but a biopsy is the only way to confirm whether there is cancer [12]. Imaging techniques like mammography, ultrasound, and magnetic resonance imaging currently stand as the main techniques for early breast cancer screening. However, limitations such as x-rays, expensiveness, dense tissue during young age, false positives (FP), and false-negative (FN) rates encouraged researchers and institutions to research alternative techniques like thermography deeply. Contrary to other modalities, thermography is a non-invasive, non-inclusive, radiation-free, and low-cost technique [13]. Thus, thermography could be used to diagnose early-stage breast cancer e.g., young women and in dense breast patients. Frequently, these novel techniques, such as thermography, are coupled with computer-aided diagnosis (CAD) systems. A CAD system is a computational tool or algorithm capable of identifying patterns in many types of data e.g., clinical 2D and 3D clinical databases; consequently, several research teams are measuring the impact of CAD systems in the diagnosis of breast cancer patients [14].

Thermography measure the temperature based on infrared radiation. In medicine, the skin’s surface temperature gives health insights because, the radiance from human skin is an exponential function of the surface temperature, in other words, it is influenced by the level of blood perfusion in the tumor [15], e.g. Krawczyk et al. [16] summarize that thermography is well suited to detect changes in blood perfusion that are led by inflammation, angiogenesis, benign and malignant tumor. In 1956 the M. D. Lawson [17] recorded for the first time the skin’s heat energy using a “thermocouple” [17], then authors presented similar devices e.g. the Pyroscan [18]. On the one hand, thermography has advantages over other techniques, in particular when the tumor is in an early-stage or in dense tissue [19, 20]. On the other hand, the thermography stands as a technique capable of overcoming the limitations of mammography such as x-rays, painfulness during the test, and not-permissible cost in some underdevelopment countries. Consequently, during the last decades, there is an increasing research focused on machine learning techniques (MLT) for breast cancer diagnosis using thermal images; some researchers focus their works on the localization and size of tumors in phantoms and simulated models; but others scientists have been focused on characteristics like breast quadrants, menstrual cycle and acquisition protocols.

During the last years, promising results have been achieved in various medical imaging applications for breast cancer diagnosis [21, 22] with CNNs. As mentioned in Section 2, we concluded that CNNs have not been used widely in the past for breast cancer diagnosis with thermography, maybe because the CNNs were not efficient as texture or statistical features, or perhaps because the computing load was too high. Nonetheless, during the last years, CNNs techniques stand as one of the main techniques for pattern recognition in images -or thermal images-. In this work, we develop a novel CNN-CAD methodology that targets the public breast thermography database called “DMR-IR” proposed by Marquez [23] and Silva et al. [24]. This CNN-based study has five main contributions listed as follows:

- We created baseline CNN models to replicate the results obtained by most of the recent studies regarding the DMR-IR database. This allowed us to find a weak spot during the training process that other studies have not to tackle previously. Therefore, we propose a new unbiased methodology to reduce further the likely training overfitting.
- In order to compare our CNN performance, here, we present a benchmark comparison between state-of-the-art architectures like ResNet, SeResNet, VGG16, Inception, InceptionResNetV2, and Xception. We demonstrated that smaller and less-complex CNN architectures are much better for the DMR-IR database.
- Following some survey articles [13, 25], we concluded that just a couple of authors had employed CNN techniques instead of texture and statistical features for this database. Thus, we propose a better CNN architecture than the state-of-the-art ones, but also limiting the overfitting during the training process.

- Besides the comparison between state-of-the-art and our proposed CNN architectures, we also developed a hyper-parameters optimization algorithm based on a tree parzen estimator to increase the performance of the models further.
- Finally, knowing that normally the databases in the biomedical environment are limited, expensive, hard to acquire and changes depending on the acquisition protocol, we measured the influence of data augmentation and database size during the training, with the intention to suggest a minimum number of patients to obtain an effective CAD system.

Although the CAD system has been trained with the DMR-IR database, this approach is useful for other databases of thermal breast images. The outline of this article is as follows. Section 2 explains the related work and the main ideas behind thermography and the influence of breast tumors in temperature changes. Section 3 describes the acquisition protocol and the methodology for data pre-processing and data augmentation. In order to illustrate the novelty and advantage of our methodology regarding other studies, we compared the results of four sets of experiments in Section 4. Lastly, discussion and conclusions are presented in Section 5 and 6.

## 2 Current techniques for breast cancer diagnosis from thermal images

The rapid growth of virtual collaboration, programming tools and computing performance have raised the interest of many researchers over CAD systems in the biomedical area. Arena et al. [26] summarize the benefits of thermography over the classical methods for breast cancer diagnosis. They tested a weighted algorithm in 109 tissue proven cases of breast cancer, that generates positive or negative result based on six features (threshold, nipple, areola, global, asymmetry and hot spot). Krawczyk et al. [16] propose an ensemble method for clustering and classification in breast cancer thermal images; additionally, a 5x2 K-fold cross-validation was made to reduce the bias and obtain a more robust model. Later, in 2009 Schaefer et al. [27] performed a fuzzy logic classification algorithm on 150 cases having an accuracy of 80%; they explain that statistical feature analysis is a key source of information for achieving high accuracy, i.e. symmetry between left and right breast, standard temperature deviation, max-min temperatures, among others. In addition, some researchers centered their studies on the tumor's characteristics and behavior such as Partridge and Wrobel [28], whose designed a method using dual reciprocity joined with genetic algorithms to localize tumors, where they found that smaller and deeply located tumors produce only a limited thermal perturbation making harder their detection. Contrary, it is possible to determine the tumor's characteristics when the thermal surface behavior is known, Das and Mishra [29] affirmed this. Kennedy et al. [30] make a comparison between breast cancer screening techniques such as thermography, mammography and ultrasound.

The majority of studies related to CAD systems and infrared imaging techniques for breast cancer diagnosis employ the public and web-available database from [23, 24]. The database is composed of 1140 images from 57 patients, where 38 carried anomalies and 19 were healthy women from Brazil; additionally, each patient has a sequence-set of 20 images. The interpretation of a breast thermography test could be either, temperature matrices or heat map images, as it is the proposed database (also called DMR-IR database). Thermal images share similarities with a standard gray-scale or colored image; thus, mostly of the studies over the DMR-IR database try to identify texture and statistical features from those thermal images and matrices. Rajendra et al. [31] built an algorithm using support vector machines for automatic classification of normal and malignant breast cancer. They extracted from the DMR-IR database texture features from the co-occurrence matrix and statistical features from the temperature matrix, achieving an accuracy, sensitivity and specificity of 88.1%, 85.71% and 90.48%, respectively. Araujo [32] presented a symbolic data analysis on 50 patients' thermograms obtaining four variables from the thermal matrices; also, he applied leave-one-out cross validation framework during the training process obtaining 85.7% of sensitivity for the malignant class and accuracy of 84%. Mambou et al. [3] describe a method to use deep neural networks and support vector machines for breast cancer diagnoses, but also, they call attention to camera sensitivity and physical model of the breast. Nevertheless, in these above-mentioned studies are not stated how it is split the 1140-image database, neither if during the training process the whole twenty-image sequences from each patient belongs to either, train or test set or both datasets simultaneously.

Table 1 shows a summary of the last studies using the DMR-IR database. Their approaches have two big branches: on the one hand, a big portion of these studies have used texture and statistical features, where they have achieved 95% accuracy [33, 34]. On the other hand, other authors rather have chosen CNNs, achieving more than 90% accuracy [35, 36]. A main concerning with Table 1 studies is that each one presents a variable number of patients, which allows us to infer that the database has suffered changes over the last years such as, the inclusion of new patients. Besides that, it is important to recall that most of these works do not mention sufficient information regarding the database split methodology during the training framework. Then, there are two possible approaches. On the one hand, it is possible to stack all twenty-image sequences from each patient in one database and then split it in train/test datasets. On the other hand, each patient's image-sequence is assigned to either, train or test set as presented in Figure 1. Indeed,

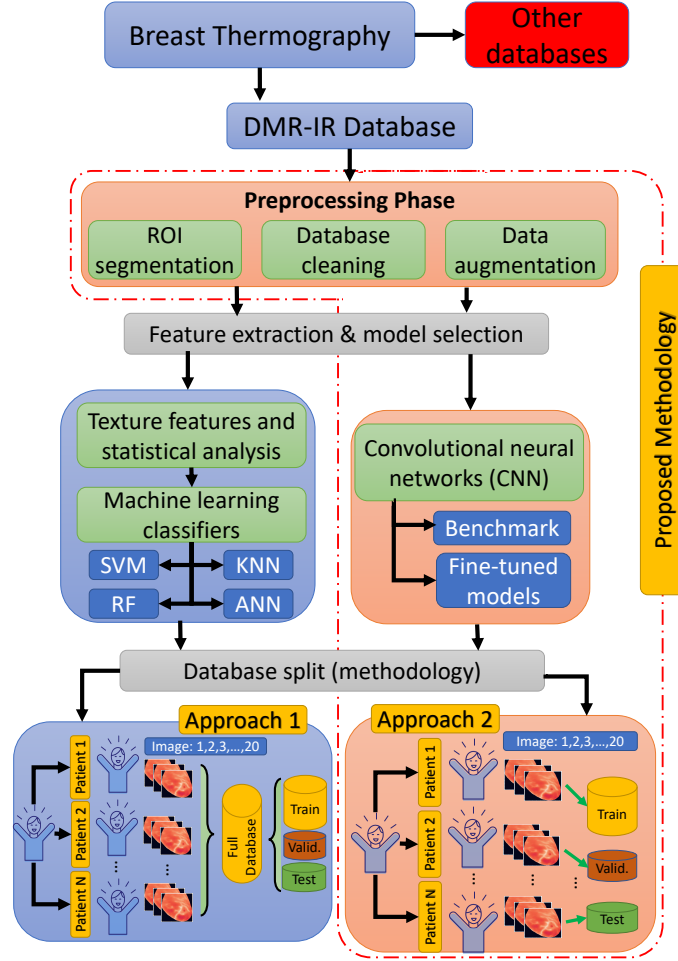


Figure 1: Current approaches for feature extraction and database splitting in DMR-IR database.

our main contributions are done under Approach 2 in the red-delimited are of Figure 1; also, this figure defines the pre-processing, training and database split frameworks of Table 1 studies.

During the last six years, several reviews concerning infrared technologies have emerged, well delimiting the status, main protocols and new directions of breast cancer imaging diagnosis techniques [43, 25, 44, 13]. One significant fact mentioned in those reviews is that CAD thermography systems need to reduce the utmost non-relevant information in the thermal images. A thermal image typically has unnecessary areas such as chest, background and other parts of the body, but this data is not useful and acts as noise during the training in CNN models or during the features identification process. Hence, the process provides a clean breast image without a non-necessary area to a CAD system is executed by a region of interest (ROI) algorithm. Regarding the DMR-IR database [23, 24], several authors have based their research on ROI algorithms rather than identifying patterns in thermal images e.g. [37] use extended hidden Markov models (EHMM), BayesNet and random forest in a 160-individuals for optimization of breast segmentation algorithms. Sathish et al. [39] extracted the breast's ROI and uses asymmetry and local energy features of wavelet sub-bands to determine whether the patient has cancer. They also concluded that the normalization of each thermal image could improve the general efficiency of the segmentation algorithm. In addition, extreme learning machines [42] and efficient coding [41] have been used for ROI segmentation.

To summarize, it is essential to recall that several aspects influence the overall performance and complexity of a given system such as pre-processing techniques, features extraction, statistical analysis, a region of interest selection, CAD technique (machine learning approach), training framework, database splitting and post-processing. Nevertheless, algorithm's complexity is not directly proportional to the algorithm's performance. This paper differs from previously published studies (see Table 1) using the DMR-IR database [23, 24], since our main goal is to demonstrate that a CNN-based CAD system outperforms the texture features based algorithms from Table 1, but at the same time, it

Table 1: Summary of algorithms based on machine learning techniques, statistical and texture features. These studies are based on the DMR-IR database for breast cancer diagnosis

Ref.	Year	Machine learning technique / Extracted Features	Acquisition protocol	Numb. of patients (Malig./Heal.)
[31]	2012	Support vector machines (SVM) for texture features and statistical analysis	Static	50 (25/25)
[24]	2014	K-nearest neighbors (KNN) algorithm to classify Affine Scale-Invariant Feature Transform (database owners)	Static & Dynamic	149
[32]	2014	They obtained the interval data in the symbolic data analysis & statistical analysis	Static	50 (31/19)
[37]	2015	Extended hidden Markov models for breast segmentation	Static	160
[38]	2015	K-means and clustering from silhouette, Davies-Bouldin and Calinski-Harabasz indexes	Dynamic	22 (11/11)
[33]	2016	BayesNet, KNN & Radom Forest (RF) models for pixel intensity and time series analysis & Static	Dynamic	80 (40/40)
[39]	2017	SVM & Genetic Algorithm (GA) for classification of normalized breast thermograms using local energy features	Static	100 (47/53)
[3]	2018	SVM, Artificial Neural Networks (ANN), Deep ANN, Recurrent ANN	Static	64 (32/32)
[40]	2018	SVM, KNN & ANN for texture features and statistical analysis	Static	80 (30/50)
[41]	2018	Bilateral asymmetry and statistical analysis for annotation of thermograms	Static	100 (49/51)
[42]	2018	Multi-Layer Perceptron (MLP), DT & RF using Zernike and Haralick moments as features	Static	100 (30/70)
[35]	2018	CNN models for static & dynamical analysis	Static & Dynamic	137 (42/95)
[36]	2019	State-of-the-art benchmark of several CNN architectures	Static	216 (41/175)
[34]	2019	Learning-to-rank (LTR) and texture analysis methods like histogram of oriented gradients	Dynamic	56 (37/19)

is less complex, easier to train and capable of generalizing more when new patients come. This paper (i) presents a new methodology (see Figure 1) which has a greater performance outperforming some studies from Table 1; (ii) compares the performance of several state-of-the-art CNN architectures (benchmark); (iii) proposes a methodology for increase the CNN performance when hyper-parameters optimization is used and; (iv) determines the impact of data augmentation, data pre-processing and database size during and after the training process.

### 3 Database description and proposed methodology

In this research, we propose several CNN-based experiments for the diagnosis of breast cancer using thermal images using the popular, free, and public available DMR-IR database, which is accessible through a user-friendly online interface (<http://visual.ic.uff.br/dmi>). In the first step, we applied data pre-processing and data augmentation for each thermal image e.g. crop, resizing, and breast normalization. In the second step, we defined several sets of interconnected experiments that tested different CNNs architectures under different training frameworks based on the database split methodologies from Figure 1 (Section 2) and following the Figure 2 workflow. As our study is based on CNN, the overview of a training process is as follows: firstly, each thermal breast image is forwarded through a given number of hidden layers until a loss function is computed; secondly, the loss function is back-propagated into these layers, modifying the weights in accordance with an optimizer e.g. Adam. Finally, this procedure is looped for given N numbers of epochs until it reaches the desired performance metric value.

The pipeline is delimited over three phases. Firstly, it is uploaded all the 1140 thermal matrices and images into Python. Then, the algorithm divides and matches the information for each patient with their respective diagnose (healthy or breast cancer). We used OpenCV python's library for ROI extraction. Phase 1 supports the data pre-processing and augmentation for each of our proposed CAD systems. Phase 2 conveys the core of our scientific contribution, which is four sets of experiments further explained in the three following sections. This phase behaves depending on two auxiliary input functions: (i) a conventional training strategy and, (ii) a Bayesian optimization + conventional training.

Lastly, Phase 3 evaluates the performance of our model using several types of metrics. Figure 2 summarizes the pipeline of our methodology.

### 3.1 Workflow description

As mentioned before, our methodology is governed by Figure 2 workflow. As the proposed methodology is interconnected, some experimental outputs become experimental inputs for other phases. Therefore, our experimental results are conducted by the following steps.

#### Step 1: Database acquisition protocol

The DMR-IR database has a population of 57 patients, with an age between 21 and 80 years old; 19 patients are healthy and 37 present a malignant breast. The diagnostic has been prior confirmed via mammography, ultrasound and biopsies. The thermal images are captured with a FLIR thermal camera model SC620, which has a sensitivity of less than  $0.04^{\circ}\text{C}$  and captures standard  $-40^{\circ}\text{C}$  to  $500^{\circ}\text{C}$ . Each infrared image has a dimension of  $640 \times 480$  pixels; the software creates two types of files: (i) a heat-map file; (ii) a matrix with  $640 \times 480$  points e.g. 307200 thermal points. Firstly, each patient undergoes thermal stress for decreasing the breast surface temperature and then twenty-image sequences are captured per 5 minutes. As a thermography test may be considerably affected when guidelines are not followed, the DMR-IR database followed the Ng [45] and Satish [43] acquisition protocol, which has been gathered jointly with physicians to ensure the database's quality. Here, it is mentioned several standards that lead to high quality and unbiased thermal images. Firstly, each patient should avoid tea, coffee, large meals, alcohol and smoking before the test. Secondly, the camera needs to run at least 15 min prior to the evaluation, having a resolution of  $100\text{mK}$  at  $30^{\circ}\text{C}$ ; the camera at least should have  $120 \times 120$  thermal points. Third, the recommended room's temperature is between  $18$  and  $25^{\circ}\text{C}$ , humidity between  $40\%$  and  $75\%$ , carpeted floor, avoiding any source of heat such as, personal computers, devices that generate heat and strong lights.

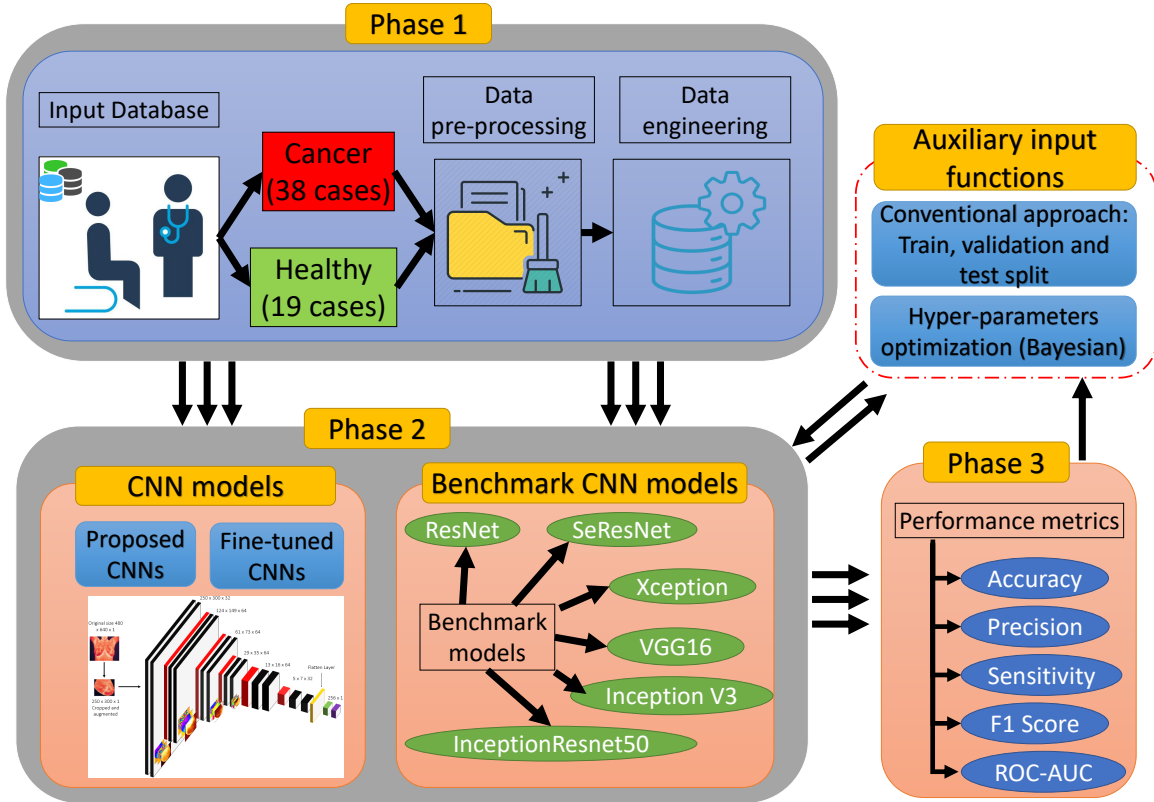


Figure 2: The detailed workflow of fine-tuned and benchmark models. There are three phases: (1) describes the data acquisition, pre-processing and data augmentation; (2) shows the three core activities (baseline, benchmark and fine-tuned CNN models) and; (3) displays the performance metrics used for evaluating all the CNN architectures.

## Step 2: Pre-processing and data augmentation

In the proposed algorithm at first, we have uploaded all the temperature matrices and mask to Python 3.7. After the data acquisition step, each breast ROI is segmented from the original gray-scale mask image, but depending on the patient’s health status, one (sick) or both (healthy) breasts are taken into consideration. In the pre-processing phase, we followed referenced methodologies [39, 42, 41, 46], such as cropping, resizing and normalization of each thermal breast image. The product of this process is a thermal image with a size of 250x300 temperature points; consequently, we reduced by a quarter the computational cost. The data augmentation step conveys four types of image data generation: (i) horizontal and vertical flip; (ii) rotation between 0-45 degrees; (iii) 20% zoom and; (iv) normalized noises, e.g. Gaussian. Algorithm 1 presents a pseudo-code of the above-mentioned techniques, where the pre-processing and data augmentation methodologies are the same for all the performed experiments (excluding the fourth experiment). The fourth experiment measures the influence of data augmentation and database (DMR-IR) size on the CNN’s performance. It is important to mention that we assumed that the database’s acquisition protocol has been done rigorously [43, 45], thus, minimizing the bias and obtaining a high-quality dataset.

---

### Algorithm 1 Data pre-processing & data augmentation

---

```

procedure DATA ENGINEERING
  DBase  $\leftarrow$  Input pre-processed database
  if Augmentation = True then select one or more:
    DBase  $\leftarrow$  DBase + horizontal or vertical flip
    DBase  $\leftarrow$  DBase + 0-45° image rotation
    DBase  $\leftarrow$  DBase + 20% zoom
    DBase  $\leftarrow$  DBase + normalized noises
  YIELD (DBase)
end if
end procedure
procedure MAIN ALGORITHM
main:
  (select one or more enhancement techniques)
  if Scale Database  $\rightarrow$  True then
    Dbase  $\leftarrow$  Scale (Dbase)
  end if
  if Crop Database  $\rightarrow$  True then
    Dbase  $\leftarrow$  Crop (Dbase)
  end if
  if Resize Database  $\rightarrow$  True then
    Dbase  $\leftarrow$  Resize (Dbase)to250x300
  end if
  DBase augmented  $\leftarrow$  YIELD (Database pre-processed)
  goto Training Algorithm.
END
end procedure

```

---

## Step 3: Baseline and benchmark CNN models

During the second phase, the CAD system has two types of the auxiliary input function, as depicted in Figure 2. Firstly, we propose a database train, validation, and test split of 50/20/30, respectively. In order to match the methodologies done by other authors, we propose a CAD system that tests several baseline CNN architectures (proposed in Figure 3) under this training framework. In fact, some authors have obtained promising results using different methodologies and pre-processing techniques; nonetheless, other authors do not mention explicitly how is the database split [3, 31, 32] during the training process; thus, there are doubts about the algorithms’ reliability and robustness when new cases will come. Contrary, we provide a detailed methodology starting from data preparation until the train/test phase, which guarantees the bias and overfitting minimization, even when new cases will come. Under that proposed training framework, we have tested several state-of-the-art CNN architectures: ResNet [47], SeResNet [48], VGG16, Inception, InceptionResNetV2 [49]and Xception [50]. Afterward, we made a comparison between the baseline and the state-of-the-art models, finding that simpler CNN models performed better than big CNN architectures; therefore, we applied optimization techniques to design an optimal CNN architecture that should perform better than experiment 1 and benchmark ones.

#### Step 4: CNN fine-tuning and hyper-parameters optimization

After comparing the baseline models with the two proposed approaches and the benchmark models, we decided to explore methodologies further to raise the performance of our CNN models; therefore, we propose a hyper-parameters Bayesian optimization based on a tree parzen estimator. Bayesian optimization is a probabilistic model-based approach for finding the global minimum of any function that returns a real-value metric (in our case F1-score). This methodology is also called a sequential model-based optimization because it builds a probabilistic model of an objective function that is based on past results. Each time the model receives new evidence, it updates the probability model (also called "surrogate model"), creating a new one with the last examples. The longer the algorithm runs, the closer the surrogate function comes to resembling the actual objective function. We implemented on Python 3.7 the tree parzen estimator (TPE) using the HyperOPT library [51]. In fact, there are four key phases in order to build a TPE pipeline:

- Firstly, it is necessary to define a domain space that will change depending on the model's evolution. The domain space may vary depending on the past results or the type of optimizer,
- The optimization algorithm. In this case, a TPE bayesian model,
- The objective function receives a set of hyper-parameters then, create a machine-learning model (here, a CNN),
- An evaluation metrics function receives a set of predicted values and real labels (cancer or healthy). Next, it returns metrics like accuracy, precision, sensitivity, F1-score and ROC-AUC.

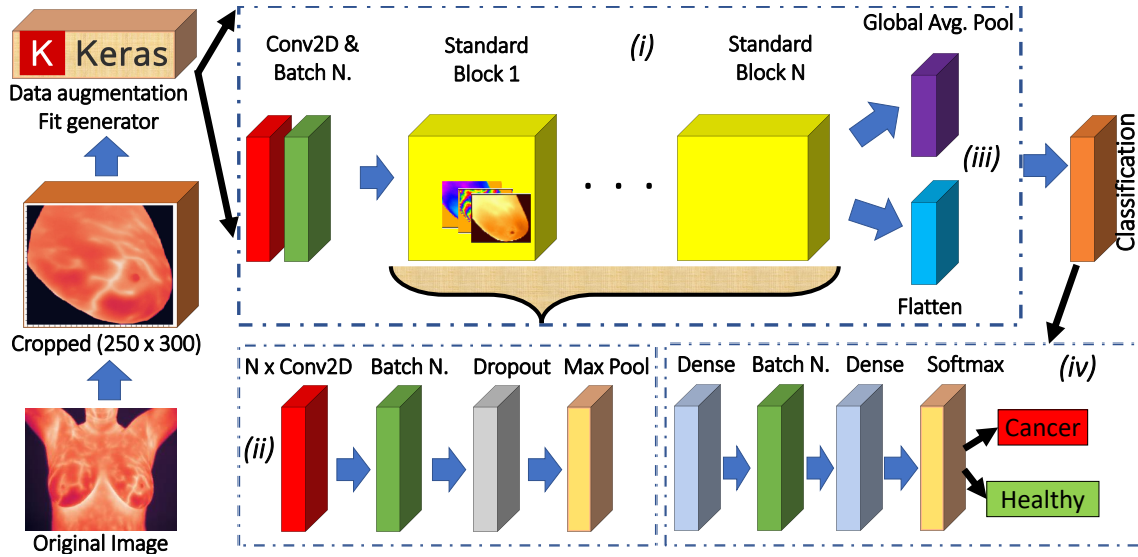


Figure 3: Convolutional neural network architecture for the baseline and hyper-parameters optimization experiments. The (i) global overview of the model is composed of (ii) standard blocks (iii) coupled with two possible top layers (flatten or global average pooling). Then a classification block (iv) provides the breast cancer diagnostic.

The CAD system's performance tells how close is the system to correctly diagnose whether a patient is having the disease and the ones who do not. In our experimentation, the people carrying a malignant breast are the true class and the ones healthy are the false class. Therefore, the easiest way to summarize a CAD system's performance is with evaluation metrics. We demonstrate the performance of our CNN models with several metrics such as accuracy, precision, sensitivity, F1-score and ROC-AUC. However, we have chosen F1-score rather than accuracy. On the one hand, the F1-score takes both false positives and false negatives into account; on the other hand, accuracy takes true positives and true negatives. The false-negative (FN) is a result that indicates a person does not have breast cancer when the person actually does have it. The false positive (FP) is a result that indicates a person does have breast cancer when the person actually does not have it. The below equations depict the proposed metrics.

$$precision = \frac{TP}{TP + FP} \quad (1)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

$$F1 - score = 2 \cdot \frac{Precision * Sensitivity}{Precision + Sensitivity} \quad (3)$$



**Remark 1:** Thence, knowing that the early diagnosis of breast cancer is crucial for the patient’s survival, the FN and FP are much more crucial parameters for a CAD system, in order to diagnose the disease and reduce the mortality index. Additionally, F1-score deals with the imbalanced class distribution problem where accuracy does not. Thus, in the biomedical area and specifically in breast cancer diagnosis, it is much more convenient than the F1-score. This metric is also recognized as the harmonic mean between precision (Equation 1) and recall (Equation 2) as depicted in Equation 3. Finally, in the case of an equal F1-score in two CNN models, we have chosen the one with greater sensitivity, as it takes into account the FN.

## 4 Experimental setup up and results

This section delimits our study but also conveys the main finding of our studies and the top CNN models obtained from empirically experimentation and hyper-parameters optimization. Additionally, we study the influence of data augmentation and database size on the CNNs performance.

### 4.1 Experimental setup

The experimental set up is composed of four consecutive experiments, explained as follows: firstly, we have developed a baseline CNN models following Figure 3 architecture. Our algorithm was trained with a Tesla K80 GPU unit, free and available in Google Colab. The training framework splits the database as follows, 50% train, 20% validation and 30% test sets, following the Approach methodology from Figure 2. All the experiments had batch normalization layers, ReLU activation function and we tested several optimizers such as Adam, RMSprop and SGD. Throughout the experiments, it has been tested several architectures varying the number of Conv2D layers, the dropout rate and the number of units in the last dense connected layers. The input image size is 250x300 temperature “pixels”. The training process has been done under the mini-augmented training batches (32 augmented images per step), with 50 steps per epoch (50 evaluation of 32 instances, per epoch), and 40 epochs in total. Finally, we summarized the results and we selected the best model based on performance metrics and execution time. As a second part of the first set of experiments, we have done a crucial change in the database splitting from here and now on (Approach 2 from Figure 1). Instead of splitting the whole database immediately, we have done a balanced splitting by patient thus, 39 patients for the train (780 images) and 17 for the test set (340 images). Again, we summarized the results and we selected the best model.

The significant change in the performance metrics found during the first set of experiments motivated us to search alternative CNNs architectures to further improve our proposed CAD system. Consequently, the second set of experiments compares state-of-the-art CNN architectures with our previous results from the first experimentation. We rearranged several novel and up-to-date CNNs architectures. We keep each original CNN architecture, but we changed the top layer by a Flatten or global average pooling (GAP) layer, followed by two dense layers of 1024 units and a two-unit dense layer with Softmax activation function. It is important to recall that all the performance metrics obtained from here and now on are based on blind test samples, i.e. samples that have not been seen by the models during the training. Similarly, we have applied three callback functions in Keras, those are: (i) model checkpoint to save the weights of our top model; (ii) learning rate scheduler to apply a decay learning rate after each epoch and; (iii) early stopping monitor to reduce the overfitting and stop the training process when the model has stopped to learn. The second set of experiments brought a main conclusion: the simpler the CNN model, the higher the performance metrics. Thus, as discussed in Section 3, we decided to apply optimization techniques to find an optimal CNN architecture.

The third set of experiments aims to find the optimal CNN architecture; particularly we have developed a Bayesian optimization. As explained in Section 3, the optimization algorithm needs a search space where to examine and chose the optimal hyper-parameters. Consequently, we defined as hyper-parameters (based on Figure 3): (i) minimum and maximum number of blocks; (ii) number of Conv2D layers per each block; (iii) number of filters per Conv2D layer; (iv); (v) type of optimizer; (vi) kernel size; (vii) pooling layer size; (viii) batch normalization; (ix) dropout rate; (x) number of dense units in the last two layers; (xi) top layer type. In the final step, the optimization algorithm chooses from fifty CNN models, the one with highest F1-score.

The fourth set of experiments takes as input the top model from the Bayesian optimization phase and applies different training frameworks. Our main goal is to measure the CNN’s performance under different training scenarios. On the one hand, we compared the results with and without data augmentation; on the other hand, we varied the database split ratio between train, validation and test datasets. As a result of this experimentation, we are able to advise an optimized population size needed to replicate each result obtained throughout this study. The training framework is as follows, we (1) selected randomly ten patients, five healthy and five with the disease; (2) trained five CNN models with 10, 20, 30, 40 and 47 patients with an 80/20 train and validation sets split; (3) evaluated the performance of each of the five models.

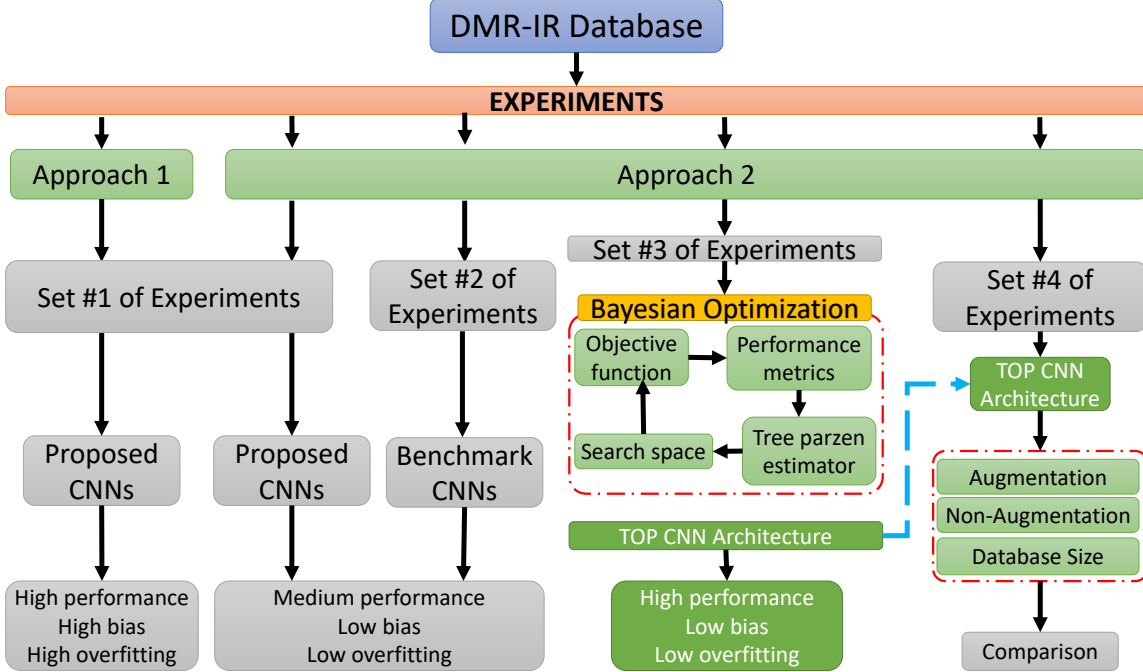


Figure 4: Experiment proposed workflows. Experiment 1 compares the CNN performance with Approach 1 and Approach 2 from Figure 1. Experiment 2 presents a CNN state-of-the-art benchmark. Experiment 3 applies a Bayesian optimization to determine the optimal CNN architecture. Experiment 4 defines the influence of data pre-processing and data augmentation on the DMR-IR database.

Afterward, we repeated the process from (1) to (3) four times, such as a “k-fold” cross-validation and we obtained the mean of each performance metric. A general overview of these proposed set of experiments could be found in Figure 4.

## 4.2 Experimental results

The purpose of the first set of experiments was to establish baseline CNN models, which tells the advantages of MLT over texture and statistical features in breast thermography. As mentioned in Section 3, a CNN has various hyper-parameters that influence the learning during the training framework, resulting in satisfactory or unacceptable results; thus, it is imperative to find the best combination of these parameters to ensure the CAD system reliability and robustness. In the first set of experiments, the leading parameters were: (i) number of CNN layers and filters; (ii) batch normalization and dropout rate; (iii) optimizer. Despite the training methodology is not a cataloged as hyper-parameter, it was important to separate the results depending on it. As discussed in the experimental setup, during the first set of experiments, we tested two different database split approaches (see Figure 1) in four CNN architectures changing the set of hyper-parameters randomly. In total, the algorithm runs a forty epochs simulation per model, which summed 200 epochs per database split methodology. Table 2 summarizes all the CNN performance metrics with each proposed splitting approach, where  $CNN_i$ , but  $i = 1, 2, 3, 4$ . We implemented a dropout rate to increase the model’s robustness by dropping out inputs from one layer to the next one. Additionally, it is set a ten-epoch early stopping callback to reduce the overfitting during the training process.

The model CNN 1 yielded the best performance in both cases, approaches 1 and 2. It is important to recall that we selected the best model based on **Remark 1** from Section 3. In the first instance, CNN1 yielded 99% accuracy, 99% precision, 98% sensitivity, 99% F1 Score and 99% ROU-AUC. Nevertheless, the second instance showed a lower performance with an 88% accuracy, 88% precision, 91% sensitivity, 89% F1-score and 88% ROU-AUC. Indeed, each CNN model has better result when using Approach 1, because there is a high probability that the CNN models under this database split methodology had images from the same patient in both datasets, train and test. In other words, images from the twenty-image sequences pertaining to a given patient could be belonging to both, the train and test set (or validation set) simultaneously. The baseline results obtained from the first set of experiments suggested that more experimentation was needed in order to reach a CAD system with high performance, low bias and low overfitting. Therefore, the idea of searching for better CNN architectures concluded in a new set of experiments based

Table 2: Comparison of five performance metrics on four CNN architectures. The hyper-parameters were given empirically and it has been tested Approach 1 (biased) and Approach 2 (unbiased) database split methodology from Figure 1 for each CNN.

Model	Class	Architecture (Num. of blocks, num of layers)	Optimizer	Top Layer	Accuracy	F1 score	Precision	Sensitivity	ROC-AUC	Time per epoch (s)
CNN 1	Biased	(5,3)	SGD	Flatten	<b>0.99</b>	<b>0.99</b>	<b>0.99</b>	<b>0.98</b>	<b>0.99</b>	<b>26</b>
	Unbiased	(5,3)	SGD	GAP	<b>0.86</b>	<b>0.87</b>	<b>0.84</b>	<b>0.90</b>	<b>0.85</b>	<b>30</b>
CNN 2	biased	(6,4)	SGD	Flatten	0.99	0.99	0.99	0.97	0.99	26
	unbiased	(6,4)	Adam	GAP	0.83	0.82	0.92	0.75	0.84	29
CNN 3	biased	(7,4)	Adam	GAP	0.92	0.92	0.94	0.90	0.92	23
	unbiased	(7,4)	Adam	GAP	0.85	0.86	0.83	0.89	0.84	21
CNN 4	biased	(4,3)	Adam	Flatten	0.89	0.89	0.92	0.87	0.90	21
	unbiased	(4,3)	Adam	GAP	0.86	0.87	0.90	0.84	0.87	25

on state-of-the-art CNN architectures, which might be capable of overcoming the weaknesses encountered during the first set of experiments.

The second set of experiments involves the benchmark of state-of-the-art CNN architectures such as ResNet, SeResNet, Inception version 3, VGG16, InceptionResNet V2 and Xception. Table 3 exhibits the performance metrics for all the proposed models. Generally, these cutting-edge CNN models are well optimized in architecture, but come at a cost of high number of parameters; indeed, higher than the models from experiment 1. We kept the database split methodology (Approach 2), datasets proportion, the number of training epochs and the early stopping callback all along the second set of experiments.

Table 3: Summary of performance metrics of each CNN model from the second set of experiments (benchmark) and the top model from the first set of experiment (Approach 1).

Model	Accuracy	F1-score	Precision	Sensitivity	ROC AUC	Time per epoch (s)
<b>SeResNet18</b>	<b>0.90</b>	<b>0.91</b>	<b>0.91</b>	<b>0.90</b>	<b>0.90</b>	<b>30</b>
SeResNet34	0.86	0.86	0.91	0.81	0.86	35
SeResNet50	0.82	0.81	0.85	0.78	0.82	42
ResNet50	0.79	0.77	0.90	0.68	0.80	30
VGG16	0.90	0.89	0.85	0.94	0.90	22
InceptionV3	0.80	0.80	0.82	0.78	0.80	21
InceptionResNetV2	0.65	0.72	0.93	0.59	0.72	44
Xception	0.90	0.89	0.89	0.90	0.90	30
<b>Top CNN (App 2.)</b>	<b>0.86</b>	<b>0.87</b>	<b>0.84</b>	<b>0.90</b>	<b>0.85</b>	<b>30</b>

Table 3 shows the performance metrics for each individual model during the forty-epoch training. We tested for each CNN model both top layers, flatten and GAP layers. In all the cases, the GAP predominated with higher performance e.g. the Inception V3 CNN model had a 30% improvement of F1-score when using GAP (not shown in Table 3) rather than flatten layer. During the experimentation, GAP layers and Adam (rather than RMSProp or SGD) optimizer yielded much better results; thus, we implemented this set of hyper-parameters for all the proposed state-of-the-art CNN models. The preeminent CNN model was the SeResNet18 with a 90% accuracy, 91% precision, 90% sensitivity, 91% F1-score and 90% ROU-AUC. We reach a breakpoint during this experiment, which suggests that simpler CNN models are better for the DMR-IR database.

As the complex the CNN architecture, the worse the model’s performance, the Figure 5 plots the accuracy and loss results over the training process of three SeResNet CNN models (presented in Table 3). Each plot represents one CNN architecture with a GAP layer followed by two fully connected layers of 1024 units each. Likewise, some CNN did not reach the forty-epoch goal due to the early stopping callback, which allows us to stop training when the model has stopped to learn. We applied L2 regularization after detecting overfitting in some CNN models. From a general

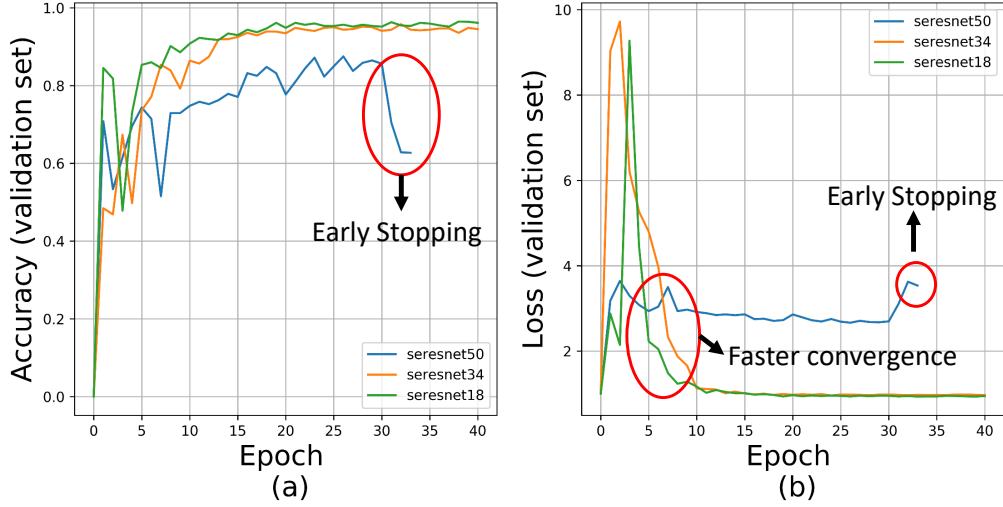


Figure 5: Validation datasets performance of SeResNet18, SeResNet34 and SeResNet50 during training for 40 epochs. Model’s (a) accuracy and (b) losses in the validation dataset.

point of view, most of the state-of-the-art CNN architectures were not as regular as the ones presented in experiment 1 (suggested CNN architecture by the authors); we believe that these models are better for small datasets such as DMR-IR database, which it is a binary classification problem (healthy or malignant breast). Contrary, Table 3 CNN benchmark models are for multi-class classification on huge databases like ImageNet. In conclusion, these benchmark results motivated us to pursuit optimization techniques to obtain the best CNN architecture.

A hyper-parameter is a parameter that controls the learning process of a given algorithm. The acquisition process of biomedical databases in most of the cases is expensive and should follow strict guidelines in order to obtain high-quality data; thus, these databases usually are both, small and unbalanced. We implemented a Bayesian optimization (explained in Section 3) of CNN hyper-parameters to deal with these problems. Firstly, our algorithm chooses a set of hyper-parameters from the proposed search space of Table 4. Secondly, the objective function creates a CNN model based on these learning and architectural parameters. The TPE algorithm provides a result based on the current and past performance metrics of the previous models.

Table 4: Search space for Bayesian optimization of CNN hyper-parameters with a tree parzen estimator. Figure 3 delimits the CNN architecture.

	Hyper-parameter	Min	Max
Quantitative	Number of blocks	2	4
	2D Conv. layers per block	2	5
	Number of filters per layer	64	512
	Kernel size (n x n)	2	4
	Pooling layer size (n x n)	2	3
	Dense Layers (Num. units)	256	1024
	L2 regularizer	0	0.2
Qualitative	Optimizer type	Adam, SGD, RMSProp	
	Droupout	Yes, No	
	Batch Normalization	Yes, No	
	Type of activation function	Elu, ReLU	
	Type of top layer	Flatten or GAP	

The best result was obtained when we associated GAP or flatten layers with a dropout rate between 0 and 0.3, but also with 6 or 7 CNN blocks. To demonstrate the success of the hyper-parameters optimization, our CNN top model yielded a 92% accuracy, 98% precision, 87% sensitivity, 92% F1-score and 92% ROU-AUC as classification metrics in the DMR-IR database. Non-conventional techniques such as our optimization algorithm are able to increase the CAD system’s performance; specifically, the mean accuracy score raised by a 6% and 8% compared with experiments 1 and

2, respectively. The hyper-parameters optimization problem was targeted as a minimization problem; despite an overall of 207.360 possible combinations of hyper-parameters, we tested fifty different sets. Table 5 presents a summary of the top models obtained during the Bayesian optimization and during experiments 1 and 2. We proposed three CNN models, where CNN-Hyp $i$ , but  $i = 1, 2, 3$ .

Table 5: Performance metrics comparison between the top CNN models from the third set of experiments (Bayesian optimization) and experiment 1 and 2. The database splitting follows the Approach 2 from Figure Figure 1.

Model	Architecture (Num. of blocks, num of layers)	Optimizer	Top Layer	Accuracy	F1 score	Precision	Sensitivity	ROC-AUC	Time per epoch
<b>CNN-Hyp 1</b>	(6,3)	RMSProp	Flatten	<b>0.94</b>	<b>0.91</b>	<b>0.92</b>	<b>0.92</b>	<b>0.92</b>	<b>40</b>
CNN-Hyp 2	(6,4)	SGD	GAP	0.99	0.83	0.90	0.90	0.91	64
CNN-Hyp 3	(7,2)	Adam	flatten	0.98	0.87	0.92	0.92	0.92	24
SeResNet18	-	Adam	30	0.91	0.90	0.91	0.90	0.90	30
Top CNN (App 2.)	(5,3)	SGD	GAP	0.84	0.90	0.87	0.86	0.85	30

To get a general overview of the results of our proposed experiments, Figure 6 summarizes the averaged performance metrics from experiments 1 to 3. We decided to show in this figure the experiments in ascending order: firstly, despite experiment 1 successful performance, we concluded that it was biased and over-fitted due to the training framework (database split Approach 1), weakening the models’ robustness. Secondly, the benchmark experimentation had higher dispersion in comparison with the other set of experiments, diminishing the model’s reliability. Finally, we measured the evolution in the performance metrics from the empirically given hyper-parameters (CNN with Approach 2) and the optimized set of hyper-parameters (Bayesian Optimization) for obtaining the topmost CNN architecture. It is important to note that the Bayesian optimization experiment displays an average increase of 7% in F1-score compared with experiment 1 (App. 2) and the benchmark experiments.

### 4.3 Influence of data augmentation and database size

The reliability and availability of databases for breast cancer diagnosis using thermography are major challenges nowadays. Consequently, this section brings guidance for new researchers in breast thermography, dealing with databases’ size issues and the role of data augmentation. In the fourth and final set of experiments, we measured the influence of data augmentation techniques and database size in the models’ performance for the DMR-IR database. In addition, each performed experiment has been tested with and without data augmentation techniques.

Figure 7 plots the averaged performance metrics varying the train/validation dataset size from 10 to 47 patients, with and without data augmentation techniques. We decided to choose an averaged performance (4 fold of metrics) rather than one set metrics because, the averaged performance decreases the excessively high bias and variance of CNNs in unseen data. We followed the k-fold cross-validation methodology, but instead changing the train/validation set, we tested four different “test sets” i.e. four test folds. The main objective during this set of experiments is to prove the advantage of data augmentation rather than no data augmentation. Section 5 discuss the main insights about the results presented throughout Section 4 and some recommendations towards future works.

## 5 Discussion and conclusion

The results presented throughout Tables 2 to 5 provide a general overview of our contribution in (i) comparing the CNNs’ performance over different database split methodologies in the DMR-IR database; (ii) providing a new methodology that highly decrease the overfitting and biasing during the training process of CNNs for this database; (iii) a benchmark comparison of state-of-the-art CNN models for the DMR-IR database. In addition, we (iv) demonstrated the benefits of hyper-parameters optimization for fine-tuning CNN architecture and; (v) measured the influence of data augmentation techniques and datasets sizes in the DMR-IR database. Hence, it is becoming as baseline information for future works in either breast thermography databases or conception of new ones.

During the last years, there have been a demand for high quality, cheap, and reliable CAD systems for a breast cancer diagnosis; but specifically, early detection. CNN-based CAD systems for thermography stands as one methodology that could satisfy those requirements. However, the lack of public databases has limited the studies towards thermography. In fact, the only public and free database is the DMR-IR [23, 24]. We assumed that DMR-IR is one of the main

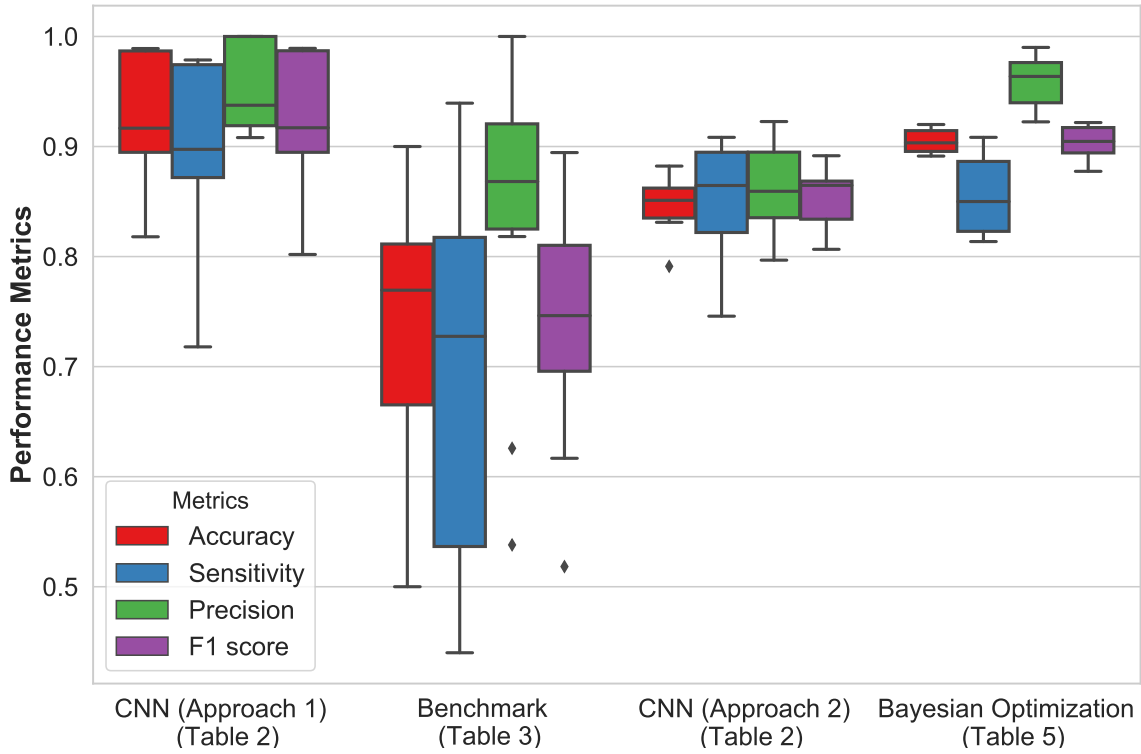


Figure 6: Summarized box-plot of performance metrics for all sets of experiments. Experiment 1 uses the database split approach 1 and 2 from Figure 1, respectively. Experiment 2 corresponds to CNN benchmarking models. Experiment 3 shows the top results obtained throughout the Bayesian optimized CNN models.

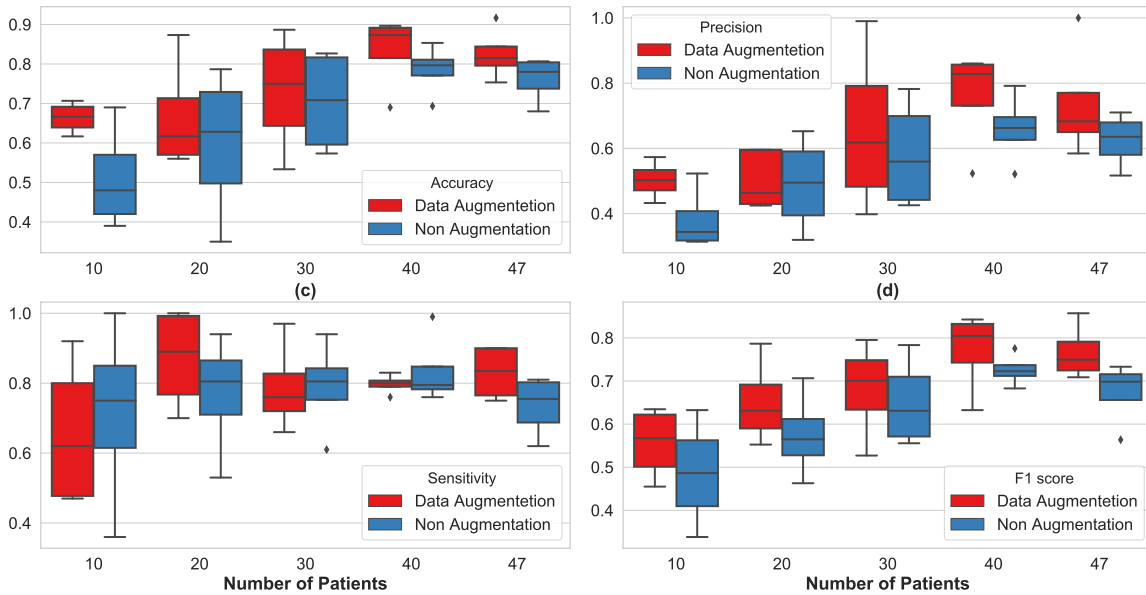


Figure 7: CNN averaged performance metrics over number of patients taken from the DMR-IR database, with and without data augmentation. The test set has randomly chosen ten patients for all cases (4 folds). (a) Accuracy, (b) Precision, (c) Sensitivity and (d) F1-score.

databases in thermography due to both, his high quality (fulfilling the standard acquisition protocols [43, 45]) and his acceptance in the research community (see Table 1). Nonetheless, when referring to past studies, it has been becoming almost impossible to compare the results impartially from study to study due to difference in the training framework, database size, datasets split ratio (between train/validation/test sets), normalization techniques and types of CAD system (texture and statistical features or CNNs).

Despite past studies that have diverged in the database sizes, we have seen two experimental methodologies. On the one hand, some authors have used texture and silhouette features coupled with machine learning techniques [31, 24, 32, 40, 41, 42, 34] to detect whether a patient does have cancer. On the other hand, a pair of studies use machine-learning techniques straightforward with the DMR-IR thermal images [35, 36]. As far as it is known, gathering texture, silhouette and statistical features demand much more time than applying MLT e.g. CNN. In addition, algorithms based on these features need more time and resources to reach the required reliability and robustness, this due to the large intra-class appearance variations triggered by changes in illumination, rotation, scale, blur, noise, occlusion, etc. Likewise, the main idea of CNN-based CAD system is to minimize the rate of pre-processing and data management needed prior to conceiving a robust machine-learning system, focusing further on the CNN architecture itself. In other words, the developing time of a fully operational CAD system based on CNNs is fewer compared to one based on texture and statistical features.

As this study is the employing of parallel MLT rather than algorithms based on texture and statistical features from thermal images, we have chosen CNN due to their performance in spatial-databases. High-level APIs such as Keras from Tensorflow [52] allows the rapid development of robust CNN architectures. We focused our first set of experiments on demonstrating the impact and consequences of the database splitting methodology over the training. On the one hand, the first set of Table 2 CNN models follows the Approach 1 for database split methodology, where some authors have presented results using a small [32, 39], medium [41] and full [42, 35] part of the DMR-IR database. The main concern with this methodology is the high performance achieved during training e.g. our top model has an accuracy, F1-score and precision of more than 98%. On the other hand, the second set of Table 2 CNN models use a more robust training framework, which all images/sequences are pertaining to a given patient either, all belong to the training or the testing set (or validation set); thence, minimizing the bias and over-fitting. Although these models yielded an average accuracy and F1-score of 84% and 85%, a thermography-based CAD system requires higher performance to overcome techniques like mammography. Generally, in the first set of experiments, CNN models with flatten layer and SGD optimizer had better results when training under Approach 1; contrary, mixing GAP layer and Adam optimizer yielded higher performance under Approach 2. No one before Fernández-Ovies et al. [36] has made a benchmark of state-of-the-art CNN architectures such as ResNet or VGG, employing the DMR-IR database. Likewise, as a second general contribution, Table 3 and Figure 5 depict a benchmark study of several state-of-the-art CNN architectures. In previous studies [35, 36], the essential contribution was not a CNN benchmark study but rather the employment CNNs as core MLT for their CAD systems. We noticed that CNNs models with Inception modules (e.g. Inception V3 and InceptionResnet) had a lower performance because these architectures have many weights and parameters to tune, so we arrive to a breakthrough conclusion: the patterns in the DMR-IR thermal images are not too complex to be generalized by a CNN. In consequence, the complex the CNN (width, depth and number of filters), the hard to generalize the DMR-IR thermal images. In order to verify these conclusions, we developed specific experimentation using several SeResNet [48] but changing the number of residual layers. In the first case, we obtained an 81% accuracy, 85% precision, 78% sensitivity and 81% F1-score with a SeResNet50, but following our assumption that the simpler the model, the better; we tested a SeResNet34 and SeResNet18. Consequently, we obtained a 9% accuracy, sensitivity and F1-score improvement when using the simpler model –SeResNet18-. To further prove our hypothesis, Figure 5 shows the validation accuracy and losses versus epoch during the training period, where the simpler the model, the faster the model converged.

The idea of simpler a CNN model yields better results that motivated to seek non-conventional techniques to improve our CNN models. Specifically, we implemented a Bayesian Optimization based on a TPE to obtain the optimal CNN architecture (see Figure 3) from the search space suggested in Table 4. As mentioned before, the top models obtained throughout the optimization performed much better than the experiment 1 and 2. In general, the flatten layer achieved better results than GAP, the SGD needed more processing time than RMSprop and Adam optimizer, but in all the cases, the results were comparable. To summarize, we plotted in Figure 6 the averaged results per experiment and per metric, from experiments 1 to 3. We deduced from Figure 6 that experiment 1 App. 1 obtained the best performance metrics but at the cost of high bias and over-fitting during the training; contrary, the App. 2 yielded high performance, but the CNN architecture was given empirically. The average of experiment 2 produced a high variance in the box-plots, because some benchmark CNN models achieved high performance, but other who does not. Finally, experiment 3 collects all the positive things such as low variance, low bias and low overfitting on the averaged performance metrics on three CNN models; moreover, rather than gives an empirical architectures to this models, we opted to apply a Bayesian optimization that yielded the optimal architecture, which overcomes all the previous CNN models.

Despite the main advantages of CNNs, one of the main known drawbacks in MLT-based CAD systems is the quantity of available data, specifically in our case, breast thermal images. In most of the circumstances, gather more data demands expensive and rigorous protocols, which should ensure the databases' high quality and reliability. Consequently, we targeted this problem inversely following the performance evolution of several CNN models when was applied data augmentation and when was altered the database size i.e. the number of patients. The Figure 7 summarizes the accuracy, sensitivity, precision and F1-score of the proposed before-mentioned comparison and also explained in Section 4. Therefore, the fourth and last set of experiments suggested –as expected- that the larger the database size (i.e. from 10 to 47 patients), the more the CNNs generalize the data and the more the performance increase. When the performance increased, the CNN models were more regular, having therefore less variance, as can be seen in Figure 7 (d). Overall, the data augmentation techniques during all the simulations performed much better than no data augmentation; for instance, the mean F1-score in all cases was at least 10% higher. If we compare the F1-score (Figure 7 (d)) of the experiments with databases sizes of 10, 20 and 30 patients, we conclude that a CNN model which uses data augmentation techniques requires 50% less number of patients to reach the same performance that a model which does not use it. Specifically, the performance of an experiment with 20-patients database and data augmentation is comparable with a one with 30-patients database and no data augmentation.

In addition, we saw an incremental evolution of the performance metrics when the database increased as well, but between 40 and 47 patients, it was seen stabilization and decreasing in the variance. To put in context, the variance between data augmentation and no data augmentation when 10 patients were 7% and 16%, respectively; similarly, for 20 (10,1% in both), 30 (11% and 13%), 40 (9% and 1%) and 47 patients (5% and 4%) there was a constant decrease in variance, therefore showing the models' robustness improving. In conclusion, the CNNs performance is a trade-off between data augmentation versus database size, the higher the database' volume, the higher the performance. Likewise, the more data augmentation the better. This is a far-reaching conclusion, which gives helpful insights for further experimentation with the DMR-IR database or for researchers, which seek to conceive new breast thermography databases. Therefore, this pioneering study could clarify upcoming experimentation with breast thermograms, where there is no information on how big the database should be in order to obtain acceptable performances.

Finally, we note that the application of this work is centered on demonstrating that CNN-based CAD systems are more viable than the ones based on texture and statistical features because of both robustness and easy implementation. We have reviewed several studies, their techniques and methodologies towards databases of thermal image for a breast cancer diagnosis; nevertheless, it is important to mention some limitations. Firstly, the lack of information (thermal images) limits the generalization that an MLT could reach (the more data the better). Secondly, the physicians and researchers expect to know what the algorithm is computing, but normally the CNN models are recognized as black box MLT; thus, innovative techniques are measuring the CNN's inside behavior throughout the training process. Further research in this area could clarify some still unanswered questions. Thirdly, the physicians prefer systems that give an image and a probability rather than a merely probability of having cancer; therefore, future work should develop CAD systems that deal with these issues.

To finish up, this article proposes a novel CNN-based method for breast cancer diagnosis using thermal images. We showed that a well-delimited database split technique is needed in order to reduce the bias and overfitting during the training process. The paper presents the last studies on the DMR-IR database. Experimental results confirm that our database split methodology minimizes the overfitting and bias during training. In addition, this paper conveys the first state-of-the-art benchmark of CNN architectures such as ResNet, SeResNet, VGG16, Inception, InceptionResNetV2 and Xception for the DMR-IR database. Likewise, this study establishes the first CNN hyper-parameters optimization in a thermography database for breast cancer, where the top CNN model achieved a 92% accuracy, 94% precision, 91% sensitivity and 92% F1-score. We demonstrated that the trade-off between database size and data augmentation techniques are crucial in classification tasks lacking sufficient data such as the one presented in the present study. We have demonstrated that CAD systems for breast cancer diagnosis with thermal images can be valuable and reliable additional tools for physicians, but further research is needed on bigger databases and in multi-class classification problems.

## Disclosure statement

The authors have stated that they have no conflicts of interest.

## Funding

This work has been supported by the INTERREG (France - Switzerland) under the SBRA project.



## Acknowledgement

JZ, ZA, KB, SM and NZ contributed to conception and design. JZ led the data pre-processing and algorithm conception. JZ, ZA, KB and SM contributed to analysis and discussion of the results. JZ contributed with the preparation of the manuscript. JZ, ZA, KB, SM and NZ contributed with to the reviews of the manuscript. All authors read and approved the final manuscript. This work has been supported by the EIPHI Graduate school (contract "ANR-17-EURE-0002").

## References

- [1] All Cancer Globocan 2018 - International Agency for Research on Cancer WHO. <http://gco.iarc.fr/today/data/factsheets/cancers/39-All-cancers-fact-sheet.pdf>, 2019. [Online; accessed 03-March-2019].
- [2] What is Cancer? National Cancer Institute. <https://www.cancer.gov/about-cancer/understanding/what-is-cancer>, 2015. [Online; accessed 03-March-2019].
- [3] Sebastien Mambou, Petra Maresova, Ondrej Krejcar, Ali Selamat, and Kamil Kuca. Breast cancer detection using infrared thermal imaging and a deep learning model. *Sensors*, 18(9):2799, 2018.
- [4] Xu Li, Essex J Bond, Barry D Van Veen, and Susan C Hagness. An overview of ultra-wideband microwave imaging via space-time beamforming for early-stage breast-cancer detection. *IEEE Antennas and Propagation Magazine*, 47(1):19–34, 2005.
- [5] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.
- [6] Freddie Bray, Ahmedin Jemal, Nathan Grey, Jacques Ferlay, and David Forman. Global cancer transitions according to the human development index (2008–2030): a population-based study. *The lancet oncology*, 13(8):790–801, 2012.
- [7] Omer Gersten and John R Wilmoth. The cancer transition in japan since 1951. *Demographic Research*, 7:271–306, 2002.
- [8] AR Omran. The epidemiologic transition: a theory of the epidemiology of population change-milbank mem. *Fund. Quart.-49-1971*, pages 509–538, 2005.
- [9] Milena Maule and Franco Merletti. Cancer transition and priorities for cancer control. *The lancet oncology*, 13(8):745–746, 2012.
- [10] Tong Li, Limei Sun, Naomi Miller, Trudey Nicklee, Jennifer Woo, Lee Hulse-Smith, Ming-Sound Tsao, Rama Khokha, Lisa Martin, and Norman Boyd. The association of measured breast tissue characteristics with mammographic density and other risk factors for breast cancer. *Cancer Epidemiology and Prevention Biomarkers*, 14(2):343–349, 2005.
- [11] Larissa Remennick. The challenge of early breast cancer detection among immigrant and minority women in multicultural societies. *The breast journal*, 12:S103–S110, 2006.
- [12] Alisson Augusto Azevedo Figueiredo, Jefferson Gomes do Nascimento, Fernando Costa Malheiros, Luis Henrique da Silva Ignacio, Henrique Coelho Fernandes, and Gilmar Guimaraes. Breast tumor localization using skin surface temperatures from a 2d anatomic model without knowledge of the thermophysical properties. *Computer methods and programs in biomedicine*, 172:65–77, 2019.
- [13] J Zuluaga-Gomez, N Zerhouni, Z Al Masry, C Devalland, and C Varnier. A survey of breast cancer screening techniques: thermography and electrical impedance tomography. *Journal of medical engineering & technology*, pages 1–18, 2019.
- [14] Miguel Patrício, José Pereira, Joana Crisóstomo, Paulo Matafome, Manuel Gomes, Raquel Seiça, and Francisco Caramelo. Using resistin, glucose, age and bmi to predict the presence of breast cancer. *BMC cancer*, 18(1):29, 2018.
- [15] Bryan F Jones. A reappraisal of the use of infrared thermal image analysis in medicine. *IEEE transactions on medical imaging*, 17(6):1019–1027, 1998.
- [16] Bartosz Krawczyk, Gerald Schaefer, and Shao Ying Zhu. Breast cancer identification based on thermal analysis and a clustering and selection classification ensemble. In *International Conference on Brain and Health Informatics*, pages 256–265. Springer, 2013.
- [17] RN Lawson. A new infrared imaging device. *Canadian Medical Association Journal*, 79(5):402, 1958.

- [18] William R Vogler and Ralph Waldo Powell. A clinical evaluation of thermography and heptyl aldehyde in breast cancer detection. *Cancer research*, 19(2):207–209, 1959.
- [19] Valerie A McCormack and Isabel dos Santos Silva. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. *Cancer Epidemiology and Prevention Biomarkers*, 15(6):1159–1169, 2006.
- [20] Giske Ursin, Linda Hovanessian-Larsen, Yuri R Parisky, Malcolm C Pike, and Anna H Wu. Greatly increased occurrence of breast cancers in areas of mammographically dense tissue. *Breast Cancer Research*, 7(5):R605, 2005.
- [21] Mugahed A Al-antari, Mohammed A Al-masni, Mun-Taek Choi, Seung-Moo Han, and Tae-Seong Kim. A fully integrated computer-aided diagnosis system for digital x-ray mammograms via deep learning detection, segmentation, and classification. *International journal of medical informatics*, 117:44–54, 2018.
- [22] Sara Hosseinzadeh Kassani, Peyman Hosseinzadeh Kassani, Michal J Wesolowski, Kevin A Schneider, and Ralph Deters. Breast cancer diagnosis with transfer learning and global pooling. *arXiv preprint arXiv:1909.11839*, 2019.
- [23] R de S Marques. [automatic segmentation of thermal mammogram images, dissertation]. In *Instituto de Computação Universidade Federal Fluminense*. Instituto de Computação Universidade Federal Fluminense, 2012. Portuguese.
- [24] LF Silva, DCM Saade, GO Sequeiros, AC Silva, AC Paiva, RS Bravo, and A Conci. A new database for breast research with infrared image. *Journal of Medical Imaging and Health Informatics*, 4(1):92–100, 2014.
- [25] Nisreen IR Yassin, Shaimaa Omran, Enas MF El Houby, and Hemat Allam. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer methods and programs in biomedicine*, 156:25–45, 2018.
- [26] Francis Arena, Clement Barone, and Thomas DiCicco. Use of digital infrared imaging in enhanced breast cancer detection and monitoring of the clinical response to treatment. In *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No. 03CH37439)*, volume 2, pages 1129–1132. IEEE, 2003.
- [27] Gerald Schaefer, Michal Závisek, and Tomoharu Nakashima. Thermography based breast cancer analysis using statistical features and fuzzy classification. *Pattern Recognition*, 42(6):1133–1137, 2009.
- [28] PW Partridge and LC Wrobel. An inverse geometry problem for the localisation of skin tumours by thermal analysis. *Engineering Analysis with Boundary Elements*, 31(10):803–811, 2007.
- [29] Koushik Das and Subhash C Mishra. Estimation of tumor characteristics in a breast tissue with known skin surface temperature. *Journal of Thermal Biology*, 38(6):311–317, 2013.
- [30] Deborah A Kennedy, Tanya Lee, and Dugald Seely. A comparative review of thermography as a breast cancer screening technique. *Integrative cancer therapies*, 8(1):9–16, 2009.
- [31] U Rajendra Acharya, Eddie Yin-Kwee Ng, Jen-Hong Tan, and S Vinitha Sree. Thermography based breast cancer detection using texture features and support vector machine. *Journal of medical systems*, 36(3):1503–1510, 2012.
- [32] Marcus C Araújo, Rita CF Lima, and Renata MCR De Souza. Interval symbolic feature extraction for thermography breast cancer detection. *Expert Systems with Applications*, 41(15):6728–6737, 2014.
- [33] Lincoln F Silva, Alair Augusto SMD Santos, Renato S Bravo, Aristófanos C Silva, Débora C Muchaluat-Saade, and Aura Conci. Hybrid analysis for indicating patients with breast cancer using temperature time series. *Computer methods and programs in biomedicine*, 130:142–153, 2016.
- [34] Mohamed Abdel-Nasser, Antonio Moreno, and Domenc Puig. Breast cancer detection in thermal infrared images using representation learning and texture analysis methods. *Electronics*, 8(1):100, 2019.
- [35] Matheus de Freitas Oliveira Baffa and Lucas Grassano Lattari. Convolutional neural networks for static and dynamic breast infrared imaging classification. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 174–181. IEEE, 2018.
- [36] Francisco Javier Fernández-Ovies, Edwin Santiago Alférez-Baquero, Enrique Juan de Andrés-Galiana, Ana Cernea, Zulima Fernández-Muñiz, and Juan Luis Fernández-Martínez. Detection of breast cancer using infrared thermography and deep neural networks. In *International Work-Conference on Bioinformatics and Biomedical Engineering*, pages 514–523. Springer, 2019.
- [37] E Mahmoudzadeh, MA Montazeri, M Zekri, and S Sadri. Extended hidden markov model for optimized segmentation of breast thermography images. *Infrared Physics & Technology*, 72:19–28, 2015.
- [38] Lincoln F Silva, Giomar O Sequeiros, Maria Lúcia O Santos, Cristina AP Fontes, Débora C Muchaluat-Saade, and Aura Conci. Thermal signal analysis for breast cancer risk verification. In *MedInfo*, pages 746–750, 2015.

- [39] Dayakshini Sathish, Surekha Kamath, Keerthana Prasad, and Rajagopal Kadavigere. Role of normalization of breast thermogram images and automatic classification of breast cancer. *The Visual Computer*, pages 1–14, 2017.
- [40] Chebbah Nabil Karim, Ouslim Mohamed, and Temmar Ryad. A new approach for breast abnormality detection based on thermography. *Medical Technologies Journal*, 2(3):245–254, 2018.
- [41] Mrinal Kanti Bhowmik, Usha Rani Gogoi, Gautam Majumdar, Debotosh Bhattacharjee, Dhritiman Datta, and Anjan Kumar Ghosh. Designing of ground-truth-annotated dbt-tu-ju breast thermogram database toward early abnormality prediction. *IEEE journal of biomedical and health informatics*, 22(4):1238–1249, 2017.
- [42] Maíra Araújo de Santana, Jessiane Mônica Silva Pereira, Fabrício Lucimar da Silva, Nigel Mendes de Lima, Felipe Nunes de Sousa, Guilherme Max Silva de Arruda, Rita de Cássia Fernandes de Lima, Washington Wagner Azevedo da Silva, and Wellington Pinheiro dos Santos. Breast cancer diagnosis based on mammary thermography and extreme learning machines. *Research on Biomedical Engineering*, 1(AHEAD):0–0, 2018.
- [43] Satish G Kandlikar, Isaac Perez-Raya, Pruthvik A Raghupathi, Jose-Luis Gonzalez-Hernandez, Donnette Dabydeen, Lori Medeiros, and Pradyumna Phatak. Infrared imaging technology for breast cancer detection—current status, protocols and new directions. *International Journal of Heat and Mass Transfer*, 108:2303–2320, 2017.
- [44] Tiago B Borchardt, Aura Conci, Rita CF Lima, Roger Resmini, and Angel Sanchez. Breast thermography from an image processing viewpoint: A survey. *Signal Processing*, 93(10):2785–2803, 2013.
- [45] EY-K Ng. A review of thermography as promising non-invasive detection modality for breast tumor. *International Journal of Thermal Sciences*, 48(5):849–859, 2009.
- [46] Mona AS Ali, Gehad Ismail Sayed, Tarek Gaber, Aboul Ella Hassanien, Vaclav Snasel, and Lincoln F Silva. Detection of breast abnormalities of thermograms based on a new segmentation method. In *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 255–261. IEEE, 2015.
- [47] Xin Yu, Zhiding Yu, and Srikumar Ramalingam. Learning strict identity mappings in deep residual networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.
- [48] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.
- [49] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [50] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.
- [51] James Bergstra, Daniel Yamins, and David Daniel Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. *Jmlr*, 2013.
- [52] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, and Craig Citro and. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.