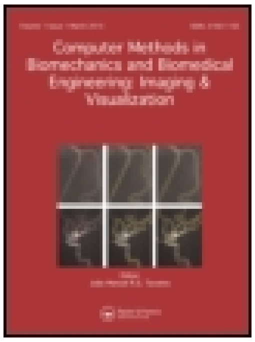

Taylor & Francis
Taylor & Francis Group

Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tciv20

A CNN-based methodology for breast cancer diagnosis using thermal images

J. Zuluaga-Gomez , Z. Al Masry , K. Benaggoune , S. Meraghni & N. Zerhouni

To cite this article: J. Zuluaga-Gomez , Z. Al Masry , K. Benaggoune , S. Meraghni & N. Zerhouni (2020): A CNN-based methodology for breast cancer diagnosis using thermal images, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, DOI: 10.1080/21681163.2020.1824685

To link to this article: https://doi.org/10.1080/21681163.2020.1824685

Published online: 04 Oct 2020.

COMPUTER METHODS IN BIOMECHANICS AND BIOMEDICAL ENGINEERING: IMAGING & VISUALIZATION
https://doi.org/10.1080/21681163.2020.1824685

A CNN-based methodology for breast cancer diagnosis using thermal images

J. Zuluaga-Gomez[a,b,c], Z. Al Masry[a], K. Benaggoune[d], S. Meraghni[e] and N. Zerhouni[a]

[a]FEMTO-ST Institute, Univ. Bourgogne Franche-Comt, CNRS, ENSMM, Besancon, France; [b]Automatic Speech Recognition Research Group, Idiap Research Institute, Martigny, Switzerland; [c]Ecole Polytechnique Federale De Lausanne (EPFL), Switzerland; [d]Laboratory of Automation and Production Engineering, Batna University, Batna, Algeria; [e]LINFI Laboratory, University of Biskra, Biskra, Algeria

ABSTRACT

A recent study from GLOBOCAN disclosed that during 2018 two million women worldwide had been diagnosed with breast cancer. Currently, mammography, magnetic resonance imaging, ultrasound, and biopsies are the main screening techniques, which require either, expensive devices or personal qualified; but some countries still lack access due to economic, social, or cultural issues. As an alternative diagnosis methodology for breast cancer, this study presents a computer-aided diagnosis system based on convolutional neural networks (CNN) using thermal images. We demonstrate that CNNs are faster, reliable and robust when compared with different techniques. We study the influence of data pre-processing, data augmentation and database size on several CAD models. 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. This study exhibits that a CAD system that implements data-augmentation techniques reach identical performance metrics in comparison with a system that uses a bigger database (up to 33%) but without data-augmentation. Finally, this study proposes a computer-aided system for breast cancer diagnosis but also, it stands as baseline research on the influence of data-augmentation and database size for breast cancer diagnosis from thermal images with CNNs

ARTICLE HISTORY
Received 2 April 2020
Accepted 14 September 2020

KEYWORDS
Breast cancer; breast thermography; hyper-parameters optimisation; Convolutional Neural Network; computer aided diagnosis system

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 (All Cancer Globocan 2019). The early detection of breast cancer is imperative to reduce the mortality and morbidity index (Li et al. 2005; What is Cancer? National Cancer Institute 2015; Mambou et al. 2018). 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 (Bray et al. 2012, 2018). Other factors like socioeconomic (Gersten and Wilmoth 2002; Bray et al. 2018), ageing, unhealthy lifestyle (Gersten and Wilmoth 2002; Omran 2001; Bray et al. 2012; Maule and Merletti 2012), environmental issues and growth of the population may perhaps lead to higher risks. In perspective, Li et al. (Li et al. 2005) 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., organisational, psychological, structural, socio-cultural, and religious (Remennick 2006). 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 (Figueiredo 2019). 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 (Zuluaga-Gomez et al. 2019). 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 (Patrício et al. 2018).

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 tumour (Jones 1998). Therefore, thermography measure the temperature based on infrared radiation, e.g. Krawczykm et al. (Krawczyk et al. 2013) summarise that thermography is well suited to detect changes in blood perfusion that are led by inflammation, angiogenesis, benign and malignant tumour. In 1956 the M. D. Lawson (Lawson 1958) recorded for the first time the skins heat energy using a thermocouple (Lawson 1958), resulting in devices such as the Pyroscan (Vogler and Powell 1959). On the one hand, thermography has advantages

CONTACT J. Zuluaga-Gomez ✉ juan.zuluaga@eu4m.eu Prognostics & Health Management Team, Femto-ST Sciences & Technologies, 25000

over other breast techniques, in particular when the tumour is in an early-stage or in dense tissue (Ursin et al. 2005; McCormack 2006). 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. During the last decades, there is an increasing research focused on machine learning techniques (MLT) for breast cancer diagnosis using thermal images; some researchers focused their works on localisation and size of tumours 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 (Al-antari et al. 2018; Kassani et al. 2019) with CNNs. Section 2 exposes that CNNs have not been used widely in the past for breast cancer diagnosis with thermography due to the high computation load. Nonetheless, during the last year CNNs techniques stand as one of the main techniques for pattern recognition in images -or thermal images-.

In this work, a novel CNN-CAD methodology has been developed to target the public breast thermography database called DMR-IR proposed by Marquez (Marques 2012) and Silva et al. (Silva et al. 2014). This CNN-based study has five main contributions listed as follows:

- CNN baseline models have been developed to replicate the results obtained by most of the recent studies regarding the DMR-IR database. This study targets the problem of overfitting which was during the training process over some previous studies.
- In order to compare the proposed CNN's performance, we present a benchmark comparison between state-of-the-art architectures like ResNet, SeResNet, VGG16, Inception, InceptionResNetV2, and Xception. The results demonstrated that smaller and shallow CNN architectures are faster and present better performance in the DMR-IR database.
- Previous survey articles (Yassin et al. 2018; Zuluaga-Gomez et al. 2019) concluded that CAD systems based on texture and statistical features have being used widely than CNN. Thus, we propose a CAD system based on CNNs that avoids 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 further increase the CNNs performance.
- Finally, 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 the minimum number of patients to obtain a given performance goal.

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 tumours in temperature changes over the skin. 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, respectively.

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. (Arena et al. 2003) summarise 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. (Krawczyk et al. 2013) propose an ensemble method for clustering and classification in breast cancer thermal images; additionally, a 5 × 2 K-fold cross-validation was made to reduce the bias and obtain a more robust model. Later, in 2009 Schaefer et al. (Schaefer et al. 2009) 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 centred their studies on the tumours characteristics and behaviour such as Partridge and Wrobel (Partridge and Wrobel 2007), whose designed a method using dual reciprocity joined with genetic algorithms to localise tumours, where they found that smaller and deeply located tumours produce only a limited thermal perturbation making harder their detection. Contrary, it is possible to determine the tumours characteristics when the thermal surface behaviour is known, Das and Mishra (Das and Mishra 2013) affirmed this. Kennedy et al. (Kennedy et al. 2009) 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 (Marques 2012; Silva et al. 2014). The database is composed of 1140 images from 57 patients, where 38 carried anomalies and 19 were healthy women from Brazilian population; 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 grey-scale or coloured image; thus, most of the studies over the DMR-IR database try to identify texture and statistical features from those thermal images and matrices. Rajendra et al. (Acharya et al. 2012) 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 (Araújo et al. 2014) 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. (Mambou et al. 2018) 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. It is important to clarify that 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. Therefore, we tackled this problem during this research.

Table 1 shows a summary of the last studies using the DMR-IR database, which can be classified as (i) texture and statistical features based achieving 95% accuracy (Silva et al. 2016; Abdel-Nasser et al. 2019) or (ii) CNN based achieving more than 90% accuracy (MdFO and Lattari 2018; Fernández-Ovies et al. 2019). 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. 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 patients image-sequence is assigned to either, train or test set as presented in Figure 1. The main contributions are done under Approach 2 (further explained in Section 3) in the red-delimited are of Figure 1; also, this figure defines the pre-processing, training and database split frameworks of some Table 1 studies.

During the last 6 years, several reviews concerning infrared technologies have emerged, delimiting the status, main protocols and mew directions of breast cancer diagnosis with imaging techniques (Borchartt et al. 2013; Kandlikar et al. 2017; Yassin et al. 2018; Zuluaga-Gomez et al. 2019). 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 full breast thermal image has unnecessary areas such as chess, background and other parts of the body, but this data is not useful during the CNN training or during the features identification process. Hence, the process that provides a clean breast image and without non-necessary areas to a CAD system is executed by a region of interest (ROI) algorithm. Regarding the DMR-IR database (Marques 2012; Silva et al. 2014), several authors have based their research on ROI algorithms rather than identifying patterns in thermal images, e.g. (Mahmoudzadeh et al. 2015) use Extended Hidden Markov Models (EHMM), BayesNet and random forest in a 160-individuals for optimisation of breast segmentation algorithms. Sathish et al. (Sathish et al. 2019) extracted the breasts ROI and uses asymmetry and local energy features of wavelet sub-bands to determine whether the patient has cancer. They also concluded that the normalisation of each thermal image could improve the general efficiency of the segmentation algorithm. In addition, extreme learning machines (MAd et al. 2018) and efficient coding (Bhowmik et al. 2018) have been used for ROI segmentation.

To summarise, 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,

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.)
(Acharya et al. 2012)	2012	Support vector machines (SVM) for texture features and statistical analysis	Static	50 (25/25)
(Silva et al. 2014)	2014	K-nearest neighbours (KNN) algorithm to classify Affine Scale-Invariant Feature Transform (database owners)	Static & Dynamic	149
(Araújo et al. 2014)	2014	They obtained the nterval data in the symbolic data analysis & statistical analysis	Static	50 (31/19)
(Mahmoudzadeh et al. 2015)	2015	Extended hidden Markov models for breast segmentation	Static	160
(Silva et al. 2015)	2015	K-means and clustering from silhouette, Davies-Bouldin and Calinski-Harabasz indexes	Dynamic	22 (11/11)
(Silva et al. 2016)	2016	BayesNet, KNN & Radom Forest (RF) models for pixel intensity and time series analysis & Static	Dynamic	80 (40/40)
(Sathish et al. 2019)	2017	SVM & Genetic Algorithm (GA) for classification of normalised breast thermograms using local energy features	Static	100 (47/53)
(Mambou et al. 2018)	2018	SVM, Artificial Neural Networks (ANN), Deep ANN, Recurrent ANN	Static	64 (32/32)
(Karim et al. 2018)	2018	SVM, KNN & ANN for texture features and statistical analysis	Static	80 (30/50)
(Bhowmik et al. 2018)	2018	Bilateral asymmetry and statistical analysis for annotation of thermograms	Static	100 (49/51)
(MAd et al. 2018)	2018	Multi-Layer Perceptron (MLP), DT & RF using Zernike and Haralick moments as features	Static	100 (30/70)
(MdFO and Lattari 2018)	2018	CNN models for static & dynamical analysis	Static & Dynamic	137 (42/95)
(Fernández-Ovies et al. 2019)	2019	State-of-the-art benchmark of several CNN architectures	Static	216 (41/175)
(Abdel-Nasser et al. 2019)	2019	Learning-to-rank (LTR) and texture analysis methods like histogram of oriented gradients	Dynamic	56 (37/19)

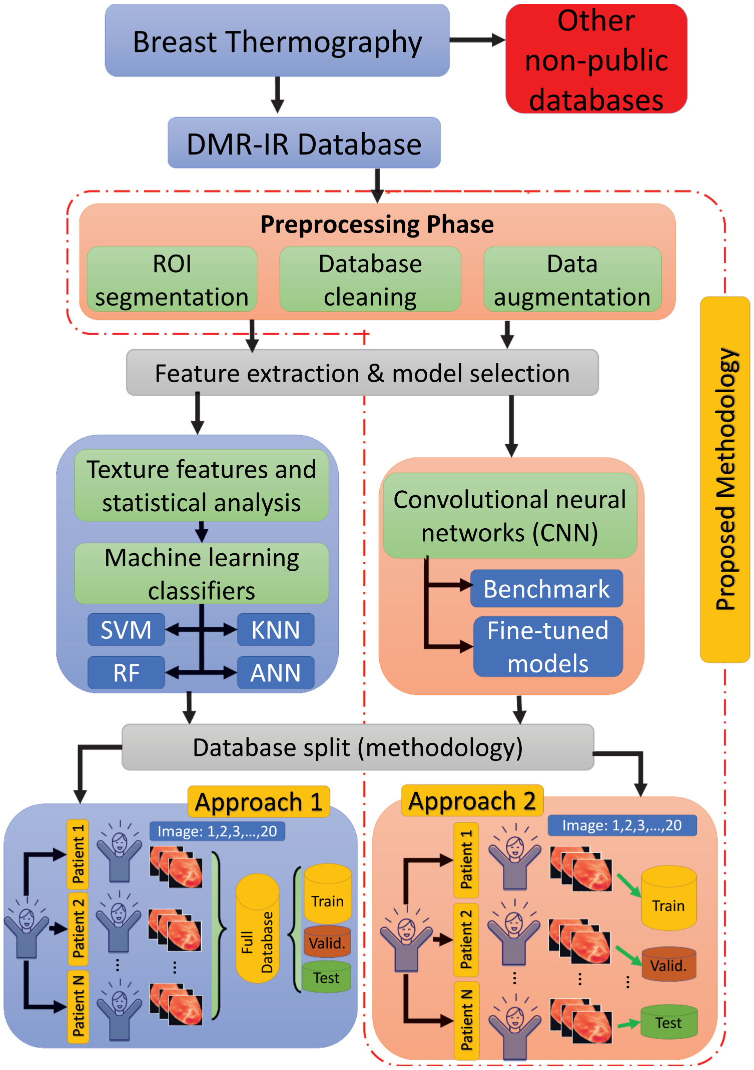

Figure 1. Current approaches for feature extraction and database splitting in DMR-IR database.

statistical analysis, a region of interest selection, CAD technique (machine learning approach), training framework, database splitting and post-processing. Nevertheless, algorithms complexity is not directly proportional to the algorithms performance. This paper differs from previously published studies (see Table 1) using the DMR-IR database (Marques 2012; Silva et al. 2014), since our main goal is to develop CNN-based CAD system that outperforms recent studies employing texture features (such as the ones presented in Table 1), but at the same time with less complexity and easier to train. This study (i) presents a new unbiased methodology (see Figure 1) for breast cancer diagnosis using thermal images and based in CNNs; (ii) compares the performance of several state-of-the-art CNN architectures (benchmark); (iii) demonstrates the advantages of hyper-parameters optimisation process over traditional training methodologies 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 only free and public-available DMR-IR database, which is accessible through a user-friendly online interface (http://visual.ic.uff.br/dmi [last accessed 21 January 2020]). We divide the overall process in the following steps: (i) in the first step, we implemented a ROI segmentation algorithm followed by a data cleaning process (or database pre-processing) for each thermal image, e.g. crop, resizing, and breast normalisation. (ii) secondly, 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 Figure 2 workflow. As our study is based on CNN, the training process is in this way: 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

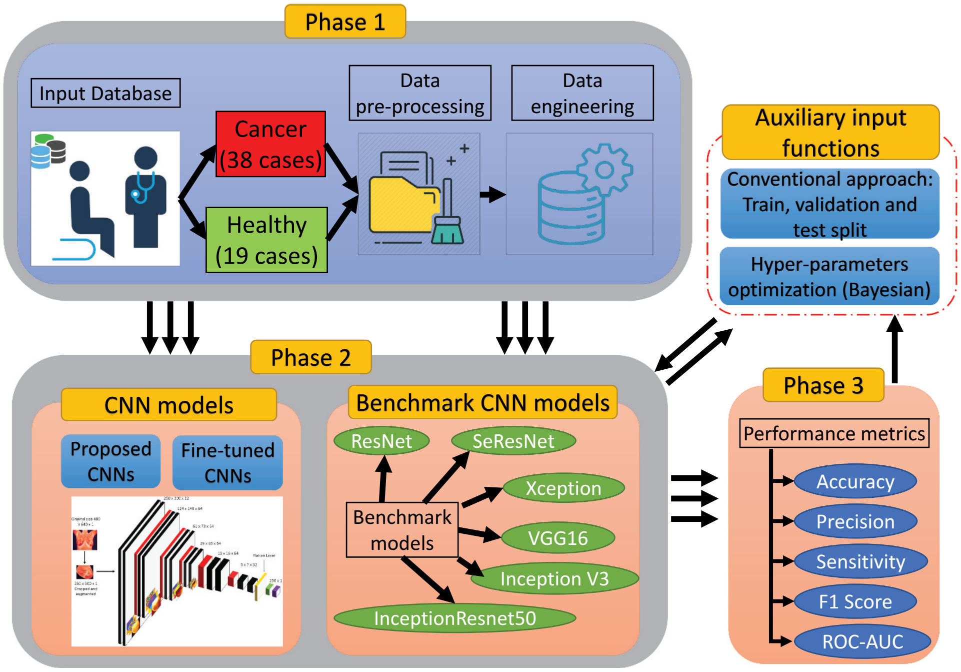

Figure 2. The detailed workflow of fine-tuned and benchmark models. There are three phases: (1) describes the data acquisition, pre-processing and data augmentation; (2) shows the three core activities (baseline, benchmark and fine-tuned CNN models) and; (3) displays the performance metrics used for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

layers, modifying the weights in accordance with an optimiser e.g. Adam. Finally, this procedure is looped for a given N numbers of epochs until it reaches the desired performance metric value (e.g. F1 score, ROC-AUC).

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 pythons library for ROI segmentation. 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 (also in Figure 2). This phase behaves depending on two auxiliary input functions: (i) a conventional training strategy and, (ii) a Bayesian optimisation + conventional training. Lastly, Phase 3 evaluates the performance of our model using several types of metrics. Figure 2 summarises 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 Algorithm 1 steps.

Algorithm 1 Data pre-processing & data augmentation

```
procedure DATA ENGINEERING
    if Augmentation = True then select one or more:
    end if
end procedure
procedure MAIN ALGORITHM
main:
    if Scale Database → True then
    end if
    if Crop Database → True then
    end if
    if Resize Database → True then
    end if
    goto Training Algorithm.
    END
end procedure
```

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 38 present a malignant breast. The diagnostic has been prior confirmed via mammography, ultrasound and biopsies by specialised physicians. The thermal images are captured with a FLIR thermal camera model SC620, which has a sensitivity of less than 0.04 C and captures standard −40 C to 500 C. Each infrared image has a dimension of 640 × 480 pixels; the software creates two types of files: (i) a heat-map file; (ii) a matrix with 640 × 480 points, e.g. 307,200 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 (Ng 2009) and Satish (Kandlikar et al. 2017) acquisition protocol, which has been gathered jointly with physicians to ensure the databases 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 100mK at 30 C; the camera at least should have 120 × 120 thermal points. Third, the recommended rooms temperature are between 18 and 25 C, humidity between 40% and 75%, carpeted floor, avoiding any source of heat such as, personal computers, devices that generate heat and strong lights.

Step 2: Pre-processing and data augmentation

We uploaded all the temperature matrices and mask images into Python 3.7. After the data acquisition step, each breast ROI is segmented from the original grey-scale mask image, but depending on the patients health status, one (sick) or both (healthy) breasts are taken into consideration. In the database cleaning phase, we followed referenced methodologies (Ali et al. 2015; Sathish et al. 2019; Bhowmik et al. 2018; MAd et al. 2018), such as cropping, resizing and normalisation of each thermal breast image. The product of this process is a thermal image with a size of 250 × 300 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 and 45 degrees; (iii) 20% zoom and; (iv) normalised 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 CNNs performance. It is important to mention that we assumed that the databases acquisition protocol has been done rigorously (Ng 2009; Kandlikar et al. 2017); thus, minimising the bias and obtaining a high-quality dataset.

Step 3: Baseline and benchmark CNN models

During the second phase, the CAD system has two types of 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 Acharya et al. 2012; Araújo et al. 2014, 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 the database splitting (Acharya et al. 2012; Araújo et al. 2014; Mambou et al. 2018) during the training process; thus, there are doubts about the algorithms reliability and robustness when new cases will come (possibly biasing the models). Contrary, we provide a detailed methodology starting from data preparation until the train/test phase, which guarantees the bias and overfitting minimisation. Under that proposed training framework, we tested several state-of-the-art CNN architectures: ResNet (Yu et al. 2018), SeResNet (Hu et al. 2018), VGG16, Inception, InceptionResNetV2 (Szegedy et al. 2017) and Xception (Chollet 2017). Afterwards, based on the baseline and the state-of-the-art CNN models performance, we proposed a new CNN-based CAD system.

Step 4: CNN fine-tuning and hyper-parameters optimisation

Step 4 conveys a hyper-parameters bayesian optimisation based on a tree parzen estimator. Bayesian optimisation 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 optimisation 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

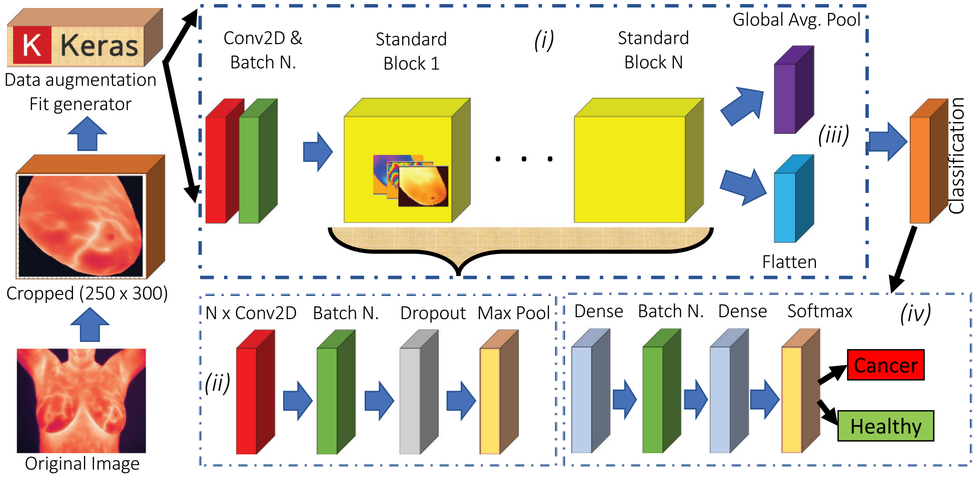

Figure 3. Convolutional neural network architecture for the baseline and hyper-parameters optimisation experiments. The (i) global overview of the model is composed of (ii) standard blocks (iii) coupled with two possible top layers (flatten or global average pooling). Then a classification block (iv)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

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 (Bergstra et al. 2013). 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 models evolution. The domain space may vary depending on the past results or the type of optimizer,
- The optimisation 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.

The CAD system's performance tells how close the system rightly diagnose whether a patient is having the disease and the ones who does not. In our experimentation, the patients carrying a malignant breast are the true class and the ones healthy is the false class. Therefore, the easiest way to summarise a CAD systems 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. F1-score deals with the imbalanced class distribution problem where accuracy does not. Thus, in the biomedical area and specifically in breast cancer diagnosis, F1-score is much more convenient than the accuracy. 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.

Remark 1: Thence, knowing that the early diagnosis of breast cancer is crucial for the patients 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. The F1-score is also recognised 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 similar CNN models, we have chosen the one with greater sensitivity, as it takes into account the FN. The below equations depict the proposed metrics.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$F1score = 2.\frac{Precision * Sensitivity}{Precision + Sensitivity} \tag{3}$$

4. Experimental setup up and results

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

4.1. Experimental setup

The experimental setup is composed of four consecutive experiments, explained as follows: firstly, we have tested several baseline CNN models following Figure 3 architecture and presented in Table 2. Our algorithm was trained with a Tesla K80 GPU unit, free and available from**'Google Colab' (product from Google Research)**. The training framework splits the database as follows, 50% train, 20% validation and 30% test sets, following both approach methodologies from Figure 2. Each CNNs model implemented batch normalisation layers, ReLU activation function and we tested several optimisers such as Adam, RMSprop and SGD. Regarding the CNN architectures, it has been tested different number of Conv2D layers, dropout rates and number of units in the last dense connected layers. We defined the input image size of 250 × 300 temperature pixels. The training process has been done under the mini-augmented training batches (32 augmented images per step), with 50 steps per epoch (aggregating 50 evaluation of 32 instances, per epoch), and 40 epochs in total. Finally, we summarised 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 only tested the database split

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)	Optimiser	Top Layer	Accuracy	F1 score	Precision	Sensitivity	ROC-AUC	Time per epoch (s)
CNN 1	biased	(5,3)	SGD	Flatten	**0.99**	**0.99**	**0.99**	**0.98**	**0.99**	**26**
	unbiased	(5,3)	SGD	GAP	**0.86**	**0.87**	**0.84**	**0.90**	**0.85**	**30**
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

methodology Approach 2 from Figure 1. Instead of splitting the whole database immediately, we have done a balanced splitting by patient; 39 patients for the train (780 images) and 17 for the test set (340 images). Again, we summarised the results and we selected the best model.

The second set of experiments compares state-of-the-art CNN architectures with our previous results from the first set of experiments. For each state-of-the-art CNN, we keep the original architecture, but changing the top layer by a Flatten or global average pooling (GAP) layers, followed by two dense layers of 1024 units and then 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 units 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 third set of experiments aims to find the optimal CNN architecture using bayesian optimisation. As explained in Section 3, the optimisation algorithm needs a search space to identify the optimal set of 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 optimiser; (vi) kernel size; (vii) pooling layer size; (viii) batch normalisation; (ix) dropout rate; (x) number of dense units in the last two layers; (xi) top layer type. We set the optimisation algorithm to 50 iterations; therefore, building, training and testing different CNN architectures. It was set to 50 iterations due to processing load and time, in fact, during the last 10 simulations were not a significant F1-score change (less than 2%).

The fourth set of experiments takes as input the top model from the bayesian optimisation phase and measure the performance in different training frameworks. 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. The training pipeline is as follows: we (1) selected randomly 10 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. Afterwards, 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 demonstrate the advantages of neural networks 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 normalisation and dropout rate; (iii) optimiser. In total, the algorithm runs a forty epochs simulation per each model and per each database split methodology (Approach 1 and 2 from Figure 1). Table 2 summarises all the CNN performance metrics with each proposed splitting approach, where CNN*i*, with $i = 1, 2, 3$. We implemented a dropout rate to increase the models 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.

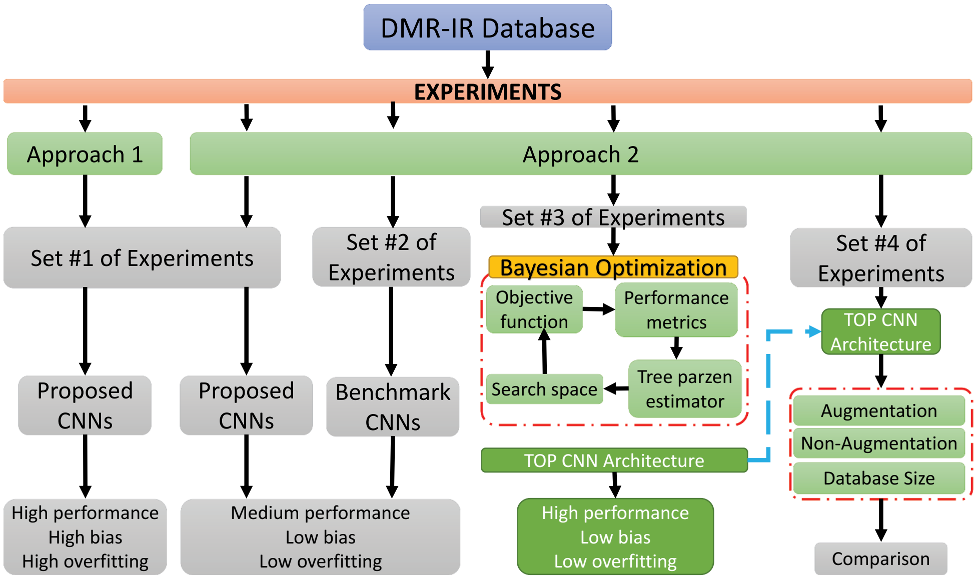

Figure 4. Experiment proposed workflows. Experiment 1 compares the CNN performance with Approach 1 and Approach 2 from Figure 1. Experiment 2 presents a CNN state-of-the-art benchmark. Experiment 3 applies a Bayesian optimisation to determine the optimal CNN architecture. Experiment 4 defines the influence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

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 on state-of-the-art CNN architectures.

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 optimised in architecture, but they 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, training epochs and the previous mentioned callbacks.

Table 3 shows the performance metrics for each CNN model during the forty-epoch training, testing 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) optimiser yielded much better results; thus, we implemented this set of hyper-parameters for all the proposed state-of-the-art CNN models. The SeResNet18 model yielded the best results with a 90% accuracy, 91% precision, 90% sensitivity, 91% F1-score and 90% ROU-AUC.

Table 3 suggests that simpler CNN models are better for the DMR-IR database; as the complex the CNN architecture, the worse the models performance, Figure 5 plots the accuracy and loss results over the training process of three SeResNet CNN models (presented in Table 2). 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 regularisation after detecting overfitting in the SeResNet models. From a general 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.

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)
SeResNet18	**0.90**	**0.91**	**0.91**	**0.90**	**0.90**	**30**
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
Top CNN (App 2.)	**0.86**	**0.87**	**0.84**	**0.90**	**0.85**	**30**

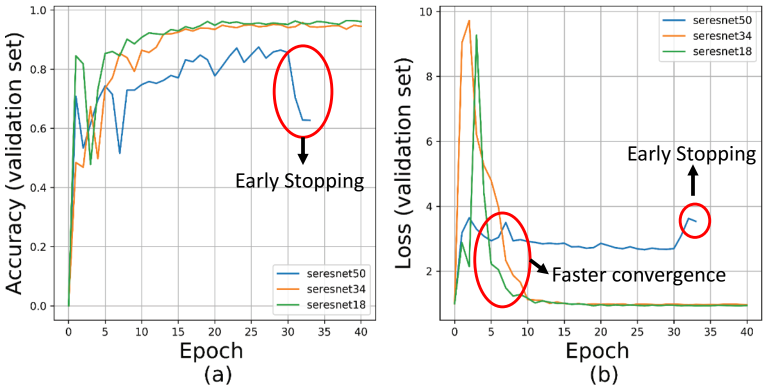

Figure 5. Validation datasets performance of SeResNet18, SeResNet34 and SeResNet50 during training for 40 epochs. Models (a) accuracy and (b) losses in the validation dataset.

The best results during the bayesian optimisation experiments were obtained when we associated GAP or flatten layers with a dropout rate between 0–0.3 and with 6 or 7 CNN blocks. To demonstrate the success of the hyper-parameters optimisation, 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. The mean accuracy score raised by a 6% and 8% compared with the first and second set experiments, respectively. The hyper-parameters optimisation problem was targeted as a minimisation problem; despite an overall of 207.360 possible combinations of hyper-parameters, we tested 50 different sets following Table 4 search space. Table 5 presents a summary of the top models obtained during the Bayesian optimisation and during experiments 1 and 2. We listed the top CNN models, where CNN-Hyp*i*, with *i* = 1, 2, 3.

Table 4. Search space for Bayesian optimisation 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 regulariser	0	0.2
Qualitative	Optimiser type	Adam, SGD, RMSProp	
	Droupout	Yes, No	
	Batch Normalisation	Yes, No	
	Type of activation function	Elu, ReLU	
	Type of top layer	Flatten or GAP	

To get a general overview of the results of our proposed experiments, Figure 6 summarises 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 models reliability. Finally, we measured the performance evolution in the metrics from the empirically given hyper-parameters (CNN with Approach 2) and the optimised set of hyper-parameters (Bayesian Optimisation) for obtaining the topmost CNN architecture. It is important to note that the Bayesian optimisation 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.

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

Model	Architecture (Num. of blocks, num of layers)	Optimiser	Top Layer	Accuracy	F1 score	Precision	Sensitivity	ROC-AUC	Time per epoch (s)
CNN-Hyp 1	(6,3)	RMSProp	Flatten	**0.94**	**0.91**	**0.92**	**0.92**	**0.92**	**40**
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
CNN 1 (App 2.)	(5,3)	SGD	GAP	0.84	0.90	0.87	0.86	0.85	30

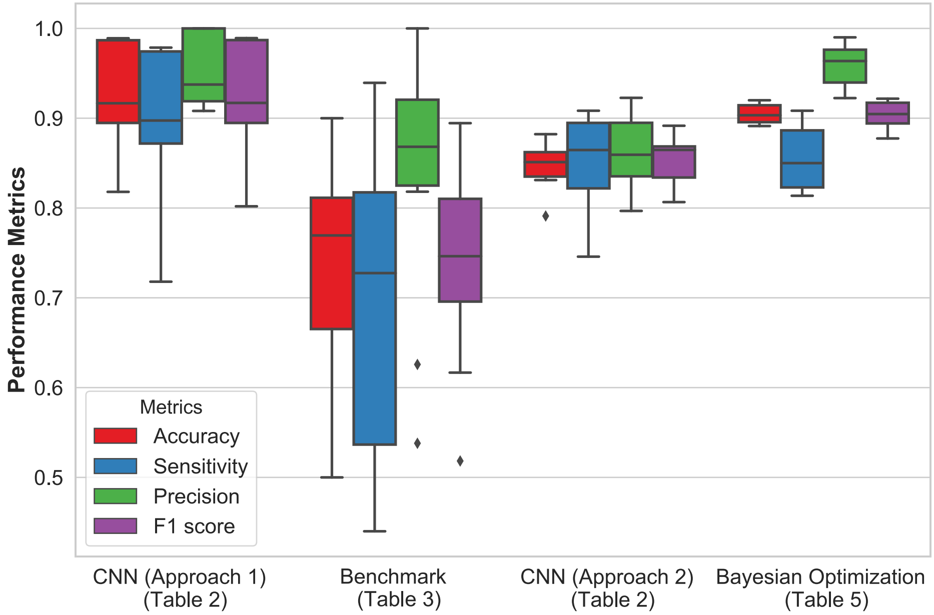

Figure 6. Summarised box-plot of performance metrics for all sets of experiments. Experiment 1 uses the database split approach 1 and 2 from Figure 1, respectively. Experiment 2 corresponds to CNN benchmarking models. Experiment 3 shows the top results obtained throughout the Bayesian optimised CNN models.

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 discusses the main insights about the results presented throughout Section 4 and some recommendations towards future works.

5. Discussion and conclusion

The results presented throughout Table 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 unbiased methodology that highly decrease the overfitting during the training process of CNNs; (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 optimisation for fine-tuning CNN architecture and, (v) measured the influence of data augmentation techniques and database size in the DMR-IR database. Figure 8 provides a general overview of our approach of a CAD CNN-based system for breast cancer diagnosis. It is important to remark that the DMR-IR database is the only available public database for researchers. Hence, this is a baseline study for future works in breast thermography.

During the last years, there have been a demand for high quality, cheap, and reliable CAD systems for breast cancer diagnosis; but specifically, in early detection tasks. CNN-based CAD systems for thermography stands as one methodology that could satisfy those requirements. However, the lack of

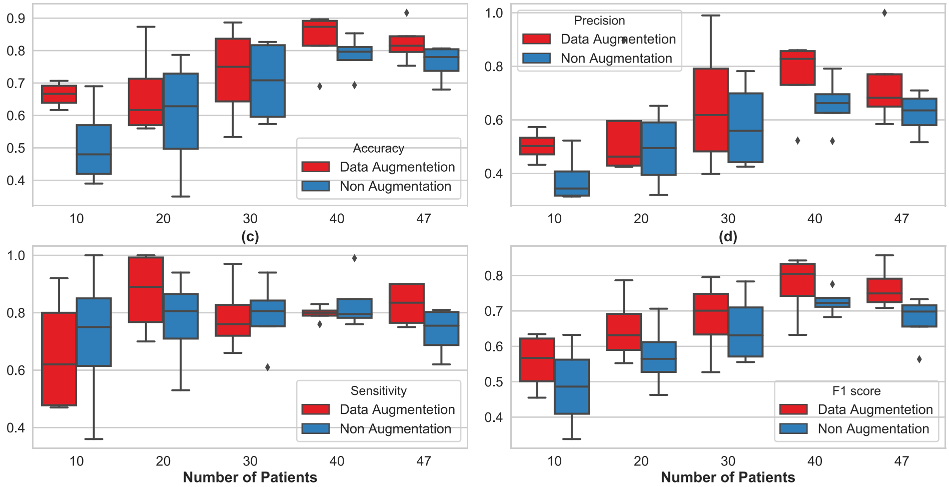

Figure 7. CNN averaged performance metrics over number of patients taken from the DMR-IR database, with and without data augmentation. The test set has randomly chosen ten patients for all cases (4 folds). (a) Accuracy, (b) Precision, (c) Sensitivity and (d) F1-score.

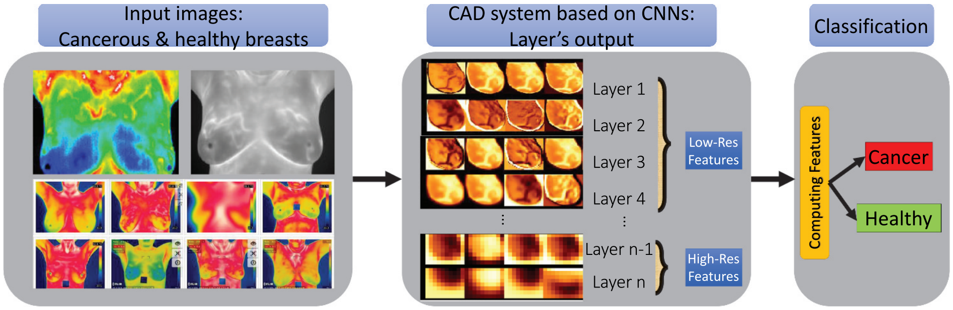

Figure 8. Global approach of the proposed CAD system. Input images are healthy and cancerous thermal images for training and testing. The convolutional neural network compute the features following the Figure 3 architecture; finally, it gives a result based on the previous computations.

public databases has limited the studies towards thermography. In fact, the only public and free database is the DMR-IR (Marques 2012; Silva et al. 2014). We assumed that DMR-IR is one of the main databases in thermography due to both, its high quality (fulfiling the standard acquisition protocols (Ng 2009; Kandlikar et al. 2017)) and its 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), normalisation techniques and types of CAD system (texture and statistical features based or CNN-based).

Despite past studies have diverged in the database sizes, we have seen two experimental methodologies. On the one hand, some authors used texture and silhouette features coupled with machine learning techniques (Acharya et al. 2012; Silva et al. 2014; Araújo et al. 2014; Karim et al. 2018; Bhowmik et al. 2018; MAd et al. 2018; Abdel-Nasser et al. 2019) to detect whether a patient does have cancer. On the other hand, a pair of studies used machine-learning techniques straightforward with the DMR-IR thermal images (MdFO and Lattari 2018; Fernández-Ovies et al. 2019). As far as it is known, gathering texture, silhouette and statistical features demand much more time than applying MLT to the raw – but pre-processed – thermal images. In addition, algorithms based on these features required more computing time and resources to reach the required reliability and robustness, this due to the large intra-class appearance variations in the thermal images triggered by changes in illumination, rotation, scale, blur, noise, occlusion, etc. Likewise, the main idea of CNN-based CAD system is to minimise 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. The first set of experiments measured the impact of database splitting methodology over the training process. On the one hand, the first set of Table 2 CNN models follow the Approach 1 for database split methodology, where some authors have presented results using a small (Araújo et al. 2014; Sathish et al. 2019), medium (Bhowmik et al. 2018) and full (MAd et al. 2018; MdFO and Lattari 2018) 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, where all images/sequences pertaining to a given patient either, all belong to the training or the testing set (or validation set); thence, minimising 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 optimiser showed better results when training under Approach 1; contrary, mixing GAP layer and Adam optimiser yielded higher performance under Approach 2. Fernndez-Ovies et al. (Fernández-Ovies et al. 2019) perform a benchmark comparison of several state-of-the-art CNN architectures such as ResNet and VGG, employing the DMR-IR database. Likewise, as a second general contribution, we intended with Table 3 and Figure 5 to provide a benchmark comparison of several state-of-the-art CNN architectures.

In previous studies (MdFO and Lattari 2018; Fernández-Ovies et al. 2019), 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 due, to the high quantity of weights and parameters to tune (sometimes ¿25 million), so we arrive to a breakthrough conclusion: the patterns in the DMR-IR thermal images are not too complex to be generalised by a CNN. In consequence, the complex the CNN (width, depth and number of filters), the hard to generalise the thermal images. In order to verify these conclusions, we developed a specific experimentation using several SeResNet (Hu et al. 2018) but changing the number of residual layers. In the first case, we obtained an 81% accuracy 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). Figure 5 shows the validation accuracy and losses versus epoch during the training period for all the SeResNet tested models.

Following the previous assumptions, we implemented a Bayesian Optimisation based on a TPE to obtain the optimal CNN architecture (see Figure 3) from the search space suggested in Table 4. The main idea was to suggest an optimal CNN architecture that maximises the performance rather than proposing a new CNN architecture. To summarise, we plotted in Figure 6 the averaged results per experiment and per metric, from experiments 1 to 3. Therefore, it can be concluded that experiment 1 App. 1 yielded 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, due to some simple versus complex CNN architectures. 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 giving an empirical architectures to this models, we opted to apply a Bayesian optimisation that gives the optimal number of layers.

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, increasing the database size demands expensive and rigorous protocols assuring the databases high quality and reliability. Consequently, we targeted this problem inversely following the performance evolution of several CNN models when altering both, data augmentation and database size, i.e. the number of patients. Figure 7 summarises the accuracy, sensitivity, precision and F1-score of these tests. Therefore, the last set of experiments suggested as expected – that the larger the database size (i.e. from 10 to 47 patients), the more the CNNs generalise 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). The data augmentation techniques during the training process rise the CNN performance; in fact, 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 33% 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, the performance increased when the database size increased as well, but between 40 and 47 patients. 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 and database size where higher the database volume, the higher the performance. Likewise, the more data augmentation the better. Therefore, this pioneering study could clarify upcoming experimentation with breast thermograms, where there is no initial information about how big the database should be in order to obtain acceptable performances.

Finally, we note that the application of this work is centred on demonstrating that CNN-based CAD systems are more viable than the ones based on texture and statistical features because of 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 generalisation of our CNN-CAD systems could reach. Secondly, the physicians and researchers expect to know what the algorithm is computing, but normally the CNN models are recognised as black box MLT; thus, innovative techniques are measuring the CNNs inside behaviour throughout the training process. Further research in this area could clarify some still unanswered questions. Thirdly, the physicians prefer systems that provide an image or tangible information such as Saliency maps, rather than a merely probability of having cancer.

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 assures the bias and overfitting decreasing during the training process. The paper presents the last studies on the DMR-IR database. Experimental results confirm that our database split methodology minimises 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 hyperparameters optimisation 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 here. 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 and investment are needed in order to build new public databases with multi-class classification problems.

Acknowledgments

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.

Article 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 complex CNNs
- Trade-off measurement between data augmentation and database size

Disclosure statement

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

Funding

This work has been supported by the EIPHI Graduate school (contract 'ANR-17-EURE-0002').

Notes on contributors

Juan Pablo Zuluaga is a PhD student in Automatic Speech Recognition and Artificial Intelligence at The École Polytechnique Fédérale de Lausanne and the Idiap Research Institute in Switzerland. His research is focused on state-of-the-art techniques for speech recognition in technical areas such as air-traffic communications. He is also a member of two students' associations, the driverless sub-team of the EPFL Racing Team and Asclepios.

Zeina Al Masry is an Associate Professor at the École Nationale Supérieure de Mécanique et des Microtechniques in Besançon, France. She is doing her research activities at the FEMTO-ST institute in the Prognostics and Health Management (PHM) research group. Her research works concern stochastic processes and applied statistics for PHM applications in medical and industrial fields.

Khaled Benaggoune is a Ph.D. student in industrial computing at Batna University, Algeria. He had a Master's degree in industrial engineering from the industrial computing department, Batna 2 University. His works focus on artificial intelligence applications in the industrial and medical fields.

Safa MERAGHNI is a PhD student in Artificial Intelligence field at LINFI laboratory in Biskra University , Algeria. She had an engineering degree in computer science in 2011 from Ecole Supérieure d'Informatique (ESI), Algiers and a Master degree in Artificial intelligence in 2015 at university of Biskra. She is working on the use of artificial intelligence and Information technologies in industry and medical fields.

Noureddine Zerhouni is a Full Professor at the École Nationale Supérieure de Mécanique et des Microtechniques in Besançon, France. He is a member of the Department of Automatic Control and Micro- Mechatronic Systems at the FEMTO-ST Institute. His current research interests include intelligent maintenance, and prognostics and health management.

ORCID

J. Zuluaga-Gomez http://orcid.org/0000-0002-6947-2706
Z. Al Masry http://orcid.org/0000-0002-6673-0140

References

Abdel-Nasser M, Moreno A, Puig D. 2019. Breast cancer detection in thermal infrared images using representation learning and texture analysis methods. Electronics. 8(1):100. doi:10.3390/electronics8010100.

Acharya UR, Ng EYK, Tan J-H, Sree SV. 2012. Thermography based breast cancer detection using texture features and support vector machine. J Med Syst. 36(3):1503–1510. doi:10.1007/s10916-010-9611-z.

Al-antari MA, Al-masni MA, Choi M-T, Han S-M, Kim T-S. 2018. A fully integrated computer-aided diagnosis system for digital x-ray mammograms via deep learning detection, segmentation, and classification. Int J Med Inform. 117:44–54. doi:10.1016/j.ijmedinf.2018.06.003.

Ali MA, Sayed GI, Gaber T, et al. Detection of breast abnormalities of thermograms based on a new segmentation method. In: 2015 Federated Conference on Computer Science and Information Systems (FedCSIS); Lodz, Poland: IEEE; 2015. p. 255–261.

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]

Araújo MC, Lima RC, De Souza RM. 2014. Interval symbolic feature extraction for thermography breast cancer detection. Expert Syst Appl. 41 (15):6728–6737. doi:10.1016/j.eswa.2014.04.027.

Arena F, Barone C, DiCicco T 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); Vol. 2; Cancun, Mexico: IEEE; 2003. p. 1129–1132.

Bergstra J, Yamins D, Cox DD. 2013. Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In International conference on machine learning, p. 115–123.

Bhowmik MK, Gogoi UR, Majumdar G, Bhattacharjee D, Datta D, Ghosh AK. 2018. Designing of ground-truth-annotated dbt-tu-ju breast thermogram database toward early abnormality prediction. IEEE J Biomedical and Health Info. 22(4):1238–1249. doi:10.1109/JBHI.2017.2740500.

Borchartt TB, Conci A, Lima RC, Resmini R, Sanchez A. 2013. Breast thermography from an image processing viewpoint: A survey. Signal Processing. 93(10):2785–2803. doi:10.1016/j.sigpro.2012.08.012.

Bray F, Ferlay J, Soerjomataram I, et al. 2018. Global cancer statistics 2018: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 68(6):394–424.

Bray F, Jemal A, Grey N, Ferlay J, Forman D. 2012. Global cancer transitions according to the human development index (2008–2030): a population-based study. Lancet Oncol. 13(8):790–801. doi:10.1016/S1470-2045(12)70211-5.

Chollet F. Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI; 2017. p. 1251–1258.

Das K, Mishra SC. 2013. Estimation of tumor characteristics in a breast tissue with known skin surface temperature. J Therm Biol. 38(6):311–317. doi:10.1016/j.jtherbio.2013.04.001.

Fernández-Ovies FJ, Alférez-Baquero ES, de Andrés-galiana EJ, et al. Detection of breast cancer using infrared thermography and deep neural networks. In: International Work-Conference on Bioinformatics and Biomedical Engineering; Granada, Spain: Springer; 2019. p. 514–523.

Figueiredo AAA. 2019. do Nascimento JG, Malheiros FC, et al. 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.

Gersten O, Wilmoth JR. 2002. The cancer transition in japan since 1951. Demogr Res. 7:271–306. doi:10.4054/DemRes.2002.7.5.

Hu J, Shen L, Sun G Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT; 2018. p. 7132–7141.

Jones BF. 1998. A reappraisal of the use of infrared thermal image analysis in medicine. IEEE Trans Med Imaging. 17(6):1019–1027. doi:10.1109/42.746635.

Kandlikar SG, Perez-Raya I, Raghupathi PA, Gonzalez-Hernandez J-L, Dabydeen D, Medeiros L, Phatak P. 2017. Infrared imaging technology for breast cancer detection – current status, protocols and new directions. Int J Heat Mass Transf. 108:2303–2320. doi:10.1016/j.ijheatmasstransfer.2017.01.086.

Karim CN, Mohamed O, Ryad T. 2018. A new approach for breast abnormality detection based on thermography. Med Tech J. 2(3):245–254. doi:10.26415/2572-004X-vol2iss3p245-254.

Kassani SH, Kassani PH, Wesolowski MJ, et al. Breast cancer diagnosis with transfer learning and global pooling. arXiv preprint arXiv:190911839. 2019.

Kennedy DA, Lee T, Seely D. 2009. A comparative review of thermography as a breast cancer screening technique. Integr Cancer Ther. 8(1):9–16. doi:10.1177/1534735408326171.

Krawczyk B, Schaefer G, Zhu SY Breast cancer identification based on thermal analysis and a clustering and selection classification ensemble. In: International Conference on Brain and Health Informatics; Maebashi, Gunma: Springer; 2013. p. 256–265.

Lawson R. 1958. A new infrared imaging device. Can Med Assoc J. 79(5):402.

Li T, et al. 2005. The association of measured breast tissue characteristics with mammographic density and other risk factors for breast cancer. Cancer Epidemiology Biomarkers & Prevention. 14(2):343–349. doi:10.1158/1055-9965.EPI-04-0490

Li X, Bond EJ, Van Veen BD, Hagness SC. 2005. An overview of ultra-wideband microwave imaging via space-time beamforming for early-stage breast-cancer detection. IEEE Antennas Propag Mag. 47 (1):19–34. doi:10.1109/MAP.2005.1436217.

MAd S, Pereira JMS, FLd S, et al. 2018. Breast cancer diagnosis based on mammary thermography and extreme learning machines. Res Biomedical Eng. 34(1):45–53.

Mahmoudzadeh E, Montazeri M, Zekri M, Sadri S. 2015. Extended hidden Markov model for optimized segmentation of breast thermography images. Infrared Physics & Technology. 72:19–28. doi:10.1016/j.infrared.2015.06.012.

Mambou S, Maresova P, Krejcar O, Selamat A, Kuca K. 2018. Breast-cancer detection using infrared thermal imaging and a deep learning model. Sensors. 18(9):2799. doi:10.3390/s18092799.

Marques R. 2012. [automatic segmentation of thermal mammogram images, dissertation]. In: Instituto de computação universidade federal fluminense. Research on Biomedical Engineering, (AHEAD). Portuguese: Instituto de Computação Universidade Federal Fluminense.

Maule M, Merletti F. 2012. Cancer transition and priorities for cancer control. Lancet Oncol. 13(8):745–746. doi:10.1016/S1470-2045(12)70268-1.

McCormack VA. 2006. dos Santos Silva I. Breast Density and Parenchymal Patterns as Markers of Breast Cancer Risk: A Meta-analysis Cancer Epidemiology and Prevention Biomarkers. 15(6):1159–1169.

MdFO B, Lattari LG Convolutional neural networks for static and dynamic breast infrared imaging classification. In: 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI); Parana, Brazil: IEEE; 2018. p. 174–181.

Ng EYK. 2009. A review of thermography as promising non-invasive detection modality for breast tumor. Int J Thermal Sci. 48(5):849–859. doi:10.1016/j.ijthermalsci.2008.06.015.

Omran AR. 2001. The epidemiologic transition: a theory of the epidemiology of population change. Bulletin of the World Health Organization. 79:161–70.

Partridge P, Wrobel L. 2007. An inverse geometry problem for the localisation of skin tumours by thermal analysis. Eng Anal Bound Elem. 31 (10):803–811. doi:10.1016/j.enganabound.2007.02.002.

Patrício M, Pereira J, Crisóstomo J, Matafome P, Gomes M, Seiça R, Caramelo F. 2018. Using resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer. 18(1):29. doi:10.1186/s12885-017-3877-1.

Remennick L. 2006. The challenge of early breast cancer detection among immigrant and minority women in multicultural societies. Breast J. 12 (s1):S103–S110. doi:10.1111/j.1075-122X.2006.00204.x.

Sathish D, Kamath S, Prasad K, et al. 2019. Role of normalization of breast thermogram images and automatic classification of breast cancer. The Visual Computer, 35(1): 57–70.

Schaefer G, Závišek M, Nakashima T. 2009. Thermography based breast cancer analysis using statistical features and fuzzy classification. Pattern Recognit. 42(6):1133–1137. doi:10.1016/j.patcog.2008.08.007.

Silva L, Saade D, Sequeiros G, Silva AC, Paiva AC, Bravo RS, Conci A. 2014. A new database for breast research with infrared image. J Med Imaging and Health Info. 4(1):92–100. doi:10.1166/jmihi.2014.1226.

Silva LF, Santos AAS, Bravo RS, Silva AC, Muchaluat-Saade DC, Conci A. 2016. Hybrid analysis for indicating patients with breast cancer using temperature time series. Comput Methods Programs Biomed. 130:142–153. doi:10.1016/j.cmpb.2016.03.002.

Silva LF, Sequeiros GO, Santos MLO, et al. 2015. Thermal signal analysis for breast cancer risk verification. In: MedInfo. p. 746–750.

Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA; 2017.

Ursin G, Hovanessian-Larsen L, Parisky YR, Pike MC, Wu AH. 2005. Greatly increased occurrence of breast cancers in areas of mammographically dense tissue. Breast Cancer Res. 7(5):R605. doi:10.1186/bcr1260.

Vogler WR, Powell RW. 1959. A clinical evaluation of thermography and heptyl aldehyde in breast cancer detection. Cancer Res. 19(2):207–209.

What is Cancer? National Cancer Institute [https://www.cancer.gov/about-cancer/understanding/what-is-cancer]; 2015. [Online; accessed 03-March-2019]

Yassin NI, Omran S, El Houby EM, Allam H. 2018. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. Comput Methods Programs Biomed. 156:25–45. doi:10.1016/j.cmpb.2017.12.012.

Yu X, Yu Z, Ramalingam S Learning strict identity mappings in deep residual networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT. 2018.

Zuluaga-Gomez J, Zerhouni N, Al Masry Z, et al. 2019. A survey of breast cancer screening techniques: thermography and electrical impedance tomography. J Med Eng Technol. 43(5):305–322.