Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-17281-x

Check for updates

Enhancing breast cancer diagnosis with deep learning and evolutionary algorithms: A comparison of approaches using different thermographic imaging treatments

Alberto Nogales[1] · **Fernando Pérez-Lara**[1] · **Álvaro J. García-Tejedor**[1]

Received: 8 November 2022 / Revised: 6 August 2023 / Accepted: 22 September 2023

Abstract

The medical field has come a long way in recent years. This fact is directly related to the application of computer science, particularly artificial intelligence. Computer vision is one of its applications with the most significant knowledge transfer to private companies or organizations. Due to the large number of tests based on images, it has multiple benefits in medical diagnosis. These benefits go from health to economics, passing through time savings. Most people know X-rays or scanners, but others have not been applied too much like thermographies. Although they are inexpensive, non-invasive, painless, and easy to implement in remote areas, their scientific evidence is not very extended. In this paper, we evaluate different approaches based on four use cases depending on which treatment we applied to the images. This step leads to various scenarios that could benefit from using advanced hybrid Artificial Intelligence models. Evaluating the solutions will not only provide us with an accurate model. Still, it will also allow us to understand further how the different thermograph information influences the diagnosis. Results show that by separating the thermography by three ranges of temperatures and using a hybrid model of convolutional neural networks and evolutive algorithms, we can achieve accuracy near 94%.

Keywords Medical imaging · Thermography · Breast cancer diagnosis · Deep learning · Convolutional neural networks · Evolutive algorithm

✉ Alberto Nogales
alberto.nogales@ceiec.es

Fernando Pérez-Lara
fernando.perez.lara@outlook.com

Álvaro J. García-Tejedor
a.gtejedor@ceiec.es

[1] CEIEC Research Institute, Universidad Francisco de Vitoria, Ctra. M-515 Pozuelo-Majadahonda Km. 1.800, 28223 Pozuelo de Alarcón, Madrid, Spain

Published online: 13 October 2023

1 Introduction

We consider Artificial Intelligence (AI) one of the most current subjects. It is defined by [44] as a set of methods to analyze and decipher mechanisms of intelligent human behaviours. Then, machines reproduce these mechanisms, not necessarily in the same way. In recent years, AI has made significant progress in several fields. This fact directly links to the creation of massive amounts of data every day and the possibility of acquiring high-performance computing equipment at low prices. Just in the medical data field, the US can reach the yottabyte (10^{24} gigabytes) [41]. The nature of these data is very diverse from images of different body parts, texts from the records of patients, or biological signals like electroencephalograms, [42]. Depending on the type of data, physicians could implement some applications in their routine like medical text summarizing in the field of Natural Language Processing (NLP), [3], epileptic seizure prediction in signal processing, [11] or brain tumor detection in computer vision, [1].

Among the different application fields, we can highlight computer vision as the most developed one in AI. It is defined by [22] as creating autonomous systems that perform human visual system tasks. It also has several subdomains: motion estimation, image restoration, or event detection. The applications of computer vision range from geographical motor sensing to human communication through healthcare, [56]. Medicine is benefiting from computer vision by making more accurate predictions. It saves money in terms of healthy people who are not diagnosed as sick and time dedicated by physicians. It improves the rates of healed people, which impacts the mental health of people or the number of leaves of absence. [4] depicts those synergies between AI techniques and healthcare will improve, making the field more data-dependent. It also states that physician AI applications will support but not replace them unless those unable to handle such techniques.

We named the field of computer vision in medicine medical imaging or medical image analysis. The definition in [2] describes it as diagnosing and studying human diseases using collected images from different body parts. The most usual ones are X-rays, Magnetic Resonance Imaging (MRI), or ultrasound. The problems with these techniques are the high costs that made them unaffordable for some countries, how painful they are, their invasiveness of leaving radiation waste, and their difficulty transporting it due to the size of the machines. This way, thermography/Medical Infrared Thermography (MIR) does not have these problems. In the case of mammographies, they have issues with false positives (healthy people diagnosed with cancer), [20], with consequences in the psychological state of patients and unnecessary expenses for the system, [43]. Previous statements indicate that there is room for improvement.

Thermography is the technique that allows registering and visualizing an object's temperature [19]. It has multiple applications in medicine, such as reconstructive surgery, dermatology, or breast cancer diagnosis, [23]. Although seminal studies of breast cancer diagnosis using thermographies produced a lack of enthusiasm due to the thermal images' subjectivity [27], it has several benefits. Its portability and low cost allow the introduction in countries with limited resources and isolated areas. In terms of body human benefits, it is painless and has no radiation, [29]. Related to the quality of the diagnosis [39] suggest that it could be used in screening stages, obtaining accurate predictions in combination with computer science tools. Roslidar et al. [43], also, the poor results in diagnosis are related to the lack of experience in interpreting the images. Hence, our focus is not solely on developing a diagnostic method with high accuracy, but also on deriving interpretive insights about the use of thermography within the diagnostic process. This dual aim promotes both effective detection and a deeper understanding of the underlying decision mechanisms of the method.

So, if we talk about thermographies and computer vision, we should speak of artificial intelligence techniques. These have demonstrated their good performance in handling large amounts of data to solve human tasks such as breast cancer diagnosis. In particular, deep learning was a significant milestone in computer vision with the rise of AlexNet, [30]. LeCun et al. [32] defines deep learning as models composed of multiple processing layers that learn data representations with multiple levels of abstraction. In the context of diagnosis, deep learning models generate probabilistic outputs, which can be interpreted as the likelihood of belonging to a specific class. In this particular case, the classes correspond to 'healthy individuals' and 'breast cancer patients'. In several of our approaches, we employ multiple deep learning models concurrently for diagnosis. For these scenarios, we incorporate evolutionary algorithms to optimally weigh the contributions of the different deep learning models towards the final diagnosis. This enables us to maximize the synergistic potential of using multiple models and enhance the overall diagnostic performance.

The main motivation for this work lies in three aspects the use of thermographies, the application of deep learning models, and the usage of evolutive algorithms. The benefits of using thermographies are related to costs and the ease of transporting thermographic cameras. Deep learning models have been used because they obtain the best results in computer vision tasks. Regarding evolutive algorithms, they are used in two ways. First, it is used to obtain a formulation to calculate the final diagnosis in the approaches that use multiple deep learning models. Second, it allows us to understand how thermographies can be interpreted which is one of the problems of using this medical test.

The main contribution of this paper consists of evaluating four ways of treating thermographies and how they influence breast cancer diagnosis by using thermal images in deep learning models or deep learning models combined with evolutive algorithms. One of these approaches can be found in the seminal paper by Nogales et al. [37], where a diagnosis was made by separating the thermographies by the red, green, and blue channels of an image into three different ones. Then, these images are fed to three other convolutional neural networks whose results are processed by an evolutive algorithm to obtain how much each model contributes to the final diagnosis. The usage of the evolutive algorithm allows us to leverage the influence of the three-channel colors in the final depending on their performance when diagnosing a patient with breast cancer or not. This decision was taken to avoid calculating the percentage of belonging to one class or another (final diagnosis) just averaging the contribution given by the three CNN models. In this way, the more accurate models will increase their percentage of influence on the final diagnosis. In two of these approaches, the followed workflow is the same, but we are adding a dynamic separation of the low, medium, and high-temperature ranges in the present work. This dynamic separation is based on the temperature differences that exist in the patients' bodies. By using this dynamic separation, we obtain two new approaches which are a similar hybrid model to the seminal paper but only using the temperature ranges: in one case with high, medium, and low temperatures and another with the two most influential ranges of temperature. Apart from that, we implemented one case that uses the thermography images as a whole by feeding them directly to a convolutional neural network.

The innovation in this paper stems from the dynamic preprocessing techniques utilized to divide images according to temperature ranges, the combined approach that fuses deep learning models with evolutionary algorithms, and the outcomes after their training. These results offer not only high accuracy in breast cancer diagnosis but also provide interpretive insights, shedding light on the influence of various temperature ranges on the diagnostic process.

The rest of the paper has the following sections. Section 2 makes a state of art related to artificial intelligence and its application with thermographies and breast cancer. Section 3 describes the used dataset and the different methods that have been used. Section 4 describes in-depth the solution proposed for the three use cases and its results, the performance of the best model after a human evaluation. Finally, Section 5 gets some conclusions and exposes future works.

2 Background

Thermographies have applications in several fields. [7] uses thermography images of buildings to evaluate Building Energy Model (BEM) simulations. Infrared images study cracks and propagation patterns in pillars, [49]. They are also used by [12] to detect pregnancy in horses. In the case of breast cancer, we have the following papers. Morales-Cervantes et al. [36] makes a classifier based on thermal asymmetries using the Sobel operator. Early breast cancer detection with temperature-based analysis (TBA), Intensity-based analysis (IBA), and Tumor Location Matching (TLM) is described in [18]. Finally, [21] segment breast cancer thermographies by improving the Chimp Optimization Algorithm.

AI techniques were applied to breast cancer with different medical tests except for thermography. [16] use textural features extracted from mammographies to estimate the diagnostic accuracy using classical AI algorithms (linear regression, decision trees, or k-nearest neighbour). Similar algorithms like logistic regression and tree-based ones with mammographies are applied by [40] to discriminate between benign and malign tumours and detect the hormone receptor. The use of Magnetic Resonance Images (MRI) has been used by [25, 24]. The former applies the QuantX score and the latter Support Vector Machine (SVM) both for breast cancer diagnosis. Finally, there are examples of using ultrasound tests to diagnose breast cancer. In [38], a Computer-Aided Diagnosis (CAD) system that uses feature extraction and SVM can differentiate between malign and benign tumors in ultrasound images. Wu et al. [57] also uses ultrasound images for breast cancer diagnosis but uses linear regression.

The previous papers use thermographies from several fields apart from breast cancer or breast cancer with AI techniques but using other medical tests. None of them applies both artificial intelligence techniques and thermographies. In the case of thermography for the diagnosis of breast cancer using artificial intelligence, [17] compiles a set of works. As can be seen, most works are classical ones and no deep learning models are applied in any of the works that appear in the review. For sure, the year of publication has some relation but in 2016 deep learning models were being applied for breast cancer diagnosis. This way, we cannot find many papers due to the lack of open datasets. The best known is [50], which was compiled at Hospital Universitário Antônio Pedro (HUAP) in Brazil and is freely available on its website.[1] This dataset is used in [35], where the training set has a size of 1,520 images after applying data augmentation strategies. The paper describes a method using a k-nearest neighbour algorithm that can diagnose breast cancer using a mobile phone with an accuracy of 99.21%. de Santana et al. [46] makes a comparison of different classifiers like Multilayer Perceptron (MLP), Random Forest (RF), or SVM, obtaining an accuracy of 73.38% in the best case with 1,052 images. A hybrid model of an SVM with a Radial basis function kernel can diagnose breast cancer with 83.22% accuracy after being trained with images of 25 healthy people, 23 patients with benign tumors, and 12 with malignant, [18].

[1] http://visual.ic.uff.br/dmi/projeto.php

Another hybrid model of random subset feature selection (RSFS) and genetic algorithm is proposed by [47]. In this case, the training dataset contains 100 images, and the model obtains 91% accuracy. In [13], the classification task involved employing an SVM)for accurate categorization, while feature extraction was performed using linear discriminant analysis (LDA). The dataset in this work consists of two databases containing some features from different patients. A comparison of multiple ML techniques for predicting and diagnosing breast cancer is presented in [26]. The training dataset consists of 30 image features such as compactness, smoothness, and concavity. Finally, [15] combine the K-nearest neighbours classifier (KNN) with neighbourhood components analysis (NCA) to create a method that discriminates between breast cancer subtypes using Fourier transform infrared (FTIR) spectroscopic imaging.

Similar to our work, the use of deep learning models to diagnose breast cancer using thermographies is low, as can be verified in [43]. In this review, only eight papers until 2020 accomplish the characteristics. CNNs preprocess, make data augmentation, and modify dataset size, [58]. In this case, the model works with 57 patients obtaining an accuracy of 92%. In [55], CNNs are also applied to classify thermography. It gets an accuracy of over 92% by segmenting infrared images by combining the curvature function k and the gradient vector flow. Another work using CNNs with the thermographies of the Brazilian dataset alongside clinical data from the patients can diagnose breast cancer with 97% accuracy, [45]. Lahane et al. [31] uses CNNs with segmented infrared images to classify between normal, benign, and malign. Then, [5] modified a deep learning model called V4 (MV4) for the early detection of breast cancer. In [14] different combinations of frontal, lateral 45-right, and lateral 45-left thermographic images are used with well-known CNN models such as DenseNet, ResNet, or VGG to diagnose breast cancer. Another work [6] made a comparison of three methods that combine attention mechanisms with VGG16 to diagnose breast cancer in thermal images. Lastly, a paper with some similarities to ours is [10] whose study involves deep learning-based segmentation of regions of interest, extraction of statistical and textural features, vascular network analysis, and a CNN classifier to distinguish between normal and abnormal breasts.

Our work differs from the previous ones as it not only builds a classifier to diagnose breast cancer with thermographies and tries to explain how the thermographies behave by understanding how the different ranges of temperature influence the diagnosis. The main objective is to evaluate four different breast cancer classifiers using other methods to preprocess infrared images. Nogales et al. [37] presented one of the proposed solutions by segmenting them into three colour channels: red, green, and blue. For each segmented image, a CNN was used to diagnose breast cancer. Then, an evolutive algorithm was applied to make the final diagnosis of the model. In the present paper, we have added three other use cases. The best one uses the previous model modified by segmenting the images with a dynamic process of the low, medium, and high-temperature ranges. Then, we implement two more use cases: one trains a CNN model directly with the infrared image and another similar to the best model presented in this paper but only with the two most influential temperature ranges.

3 Materials and methods

In this section, there is a description of the dataset used to train the models. Then, a formal definition of the methods used during the four use cases. Finally, there is a description of the workflows in these use cases.

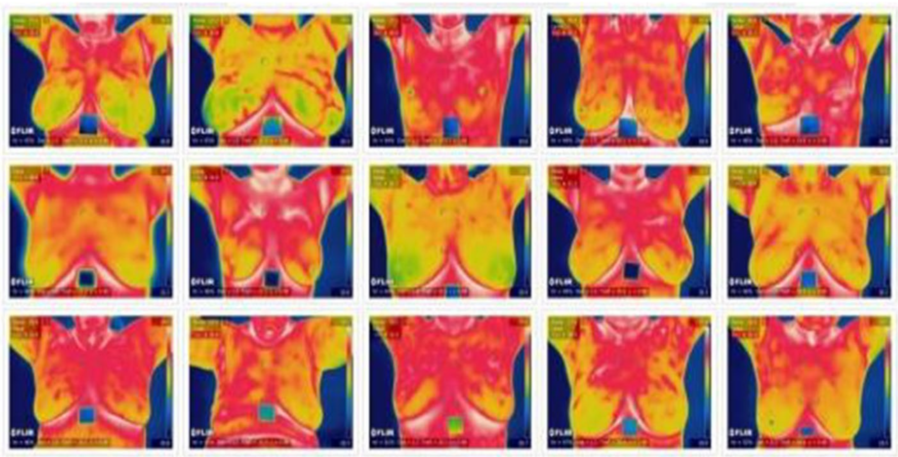

Fig. 1 Examples of the dataset used in the research

3.1 Breast cancer thermographies

During this work, we have used Database for Mastology Research (DMR) 5, which contains IR images for early detection of breast cancer. It contains thermographies, thermal matrices, and personal and clinical features from 287 University Hospital Antônio Pedro (HUAP) patients of the Federal Fluminense University of Brazil. The images were collected with a FLIR SC620 thermal camera with a sensitivity of less than 0.04ºC range and captured standard -40ºC to 500ºC. Compiled images have a size of 640 × 4,680 pixels and were obtained with a dynamic and a static protocol depending on the body's behaviour concerning heat transfer. During the former, the temperature of the body is stable. In the latter, the patient's body is cooled by a fan, so the temperature is monitored during recovery. Images of the static protocol were five (frontal, two laterals of the left at 45º and 90º, plus the same for the right lateral). For the dynamic protocol, twenty images were taken for five minutes plus two lateral images of 90º. Thermographies correspond to 287 patients with a total amount of 3,534 images that, suppose an average of 27 images per patient. The primary label of the image indicates whether the patient is healthy or sick. The images have been built based on the temperature matrices; the information extracted from the camera. Examples of the dataset are compiled in Fig. 1.

Apart from the images, some personal and clinical data were collected. Personal data includes age, race, or marital status. Clinical data comprises eating habits, use of prosthesis, use of the hormone, the inclusion of mammography, or application of radiotherapy. Another data is related to the moment of taking the images, which refers to body temperature and consumption of coffee or alcohol before the test.

The dataset acquisition was approved by the Ethical Committee of the HUAP and registered at the Brazilian Ministry of Health under number CAAE: 01042812.0.0000.5243 and is freely available under registration on the website.[2]

[2] http://visual.ic.uff.br/dmi/

3.2 An evaluation of breast cancer diagnosis based on thermographies image treatments.

This work aims to evaluate how we can improve the accuracy of different models depending on the treatment strategies applied to breast cancer IR images. By doing this, we are not only obtaining a model to diagnose whether a patient is sick or healthy based on thermography but also analyzing if the information represented in this medical test can improve the results. For this purpose, we proposed four other methods for evaluating four different ways of treating the image.

The simplest one consists of just using the complete image in a convolutional neural network to obtain good accuracy in the diagnosis.

The second use case is the one presented in Nogales et al. [37] which separates the thermography by color channels (red, green, and blue), uses these images in three different deep learning models, and leverages the contribution of each model to the final diagnosis by using an evolutive algorithm.

The third one is a two-step method based on the one presented in Nogales et al. [37], but in this case, it uses a dynamical separation of the low, medium, and high ranges of temperatures instead of the red, green, and blue channels. This method works as follows. First, we divide the thermography into three others that correspond to the low, medium, and high temperatures registered in the body. Given that average body temperature varies from individual to individual, these temperature ranges are dynamically established based on the maximum and minimum values obtained from the patient's thermal matrix. This approach ensures a customized evaluation reflective of the patient's unique thermal profile. The three images are then used to train three different CNN models to obtain a diagnosis for each temperature range. Each CNN has an output that represents the probability of being a patient with cancer or a healthy one, being the sum of them 1. In the second stage of the model, we merge the previous outputs to obtain an averaged final diagnosis of the patient. Depending on the temperature range, the model has more facility to get an accurate diagnosis; we are weighing the contribution of each of them. We have implemented an evolutionary algorithm to find the consensus-weighted contributions for all test cases to solve the previous problem. Figure 2 describes the workflow of this model. The IR image is separated into three images representing the low, medium, and high-temperature ranges. Then, a convolutional model is used to make a diagnosis with each of the segmented thermographies. The formulation of the final diagnosis accuracy is obtained with the evolutionary algorithm by making predictions with a test set.

The final use case works the same as the previous one but only uses the two most influential temperature ranges.

Convolutional neural networks Convolutional neural networks are the most significant milestone in deep learning, particularly computer vision [33]. However, its real impact occurred in 2012, when [30] presented AlexNet, which could find the main characteristics of an image regardless of the location. These models could find parts to analyze an image and recognize elements previously found in others.

As in other deep learning models, CNNs have an input layer, different hidden layers, and an output layer. The second one contains convolutional blocks formed by a convolutional layer and a pooling layer. These blocks are stacked so they can extract the main visual characteristics of the images used to train the model. By applying these blocks, the dimensionality of the images is reduced until a small piece of information contains its main

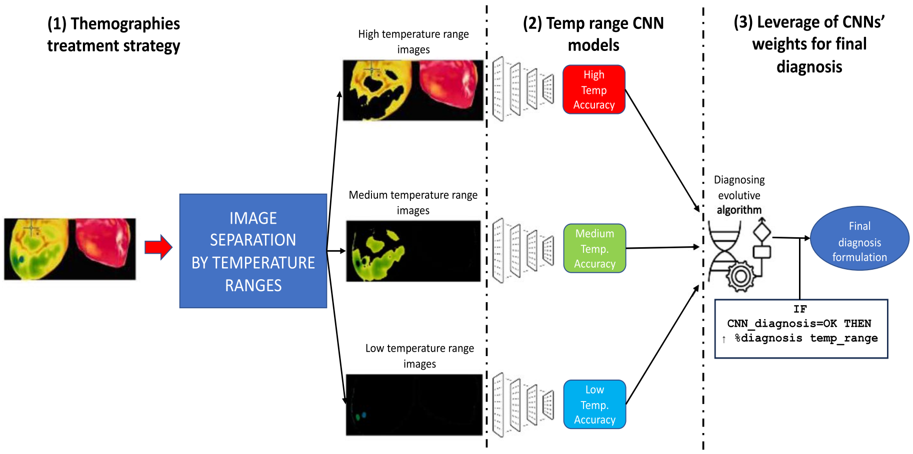

Fig. 2 Workflow followed by the hybrid model using the dynamic three ranges of temperature

features. Convolutional layers work with kernels or filters, representing neurons and their weights. The filters go through all the training images and calculate the element-wise product summed up, obtaining the output value, Eq. 1. The output represents an essential feature of the image. After passing the filter through the image, the final information is called the activation map, represented by a matrix of numbers. The hyperparameters to be tuned during the training stage are the number and size of filters.

$$Y_k = f.\left(W_k * X\right) \tag{1}$$

In the Equation above, X represents an input image, and W_k is the filter with the k_{th} features map, which applies the convolutional operator element by element (pixels multiplication).

The pooling layer is applied after the convolutional. This layer reduces the dimensionality by passing a filter through the activation map. There are two pooling layers: max (it computes the maximum value) and mean (it averages the activation map values).

Weighted-average diagnosing evolutive algorithm In this paper, we evaluate four use cases of a breast cancer diagnosis. Three of them are based on separating thermography into three or two images that correspond to temperature ranges and colour channels. In this case, we made the first diagnosis for each temperature or colour channel range and then a final diagnosis by obtaining a formula with an evolutive algorithm that weights the contribution of the previous CNN models with differential evolution (DE). DE, a population-based metaheuristic search algorithm, was proposed by [54].

The algorithm starts with an initial population of individuals (chromosomes), each a possible solution to the problem. The individuals with better fitness are considered the best ones and are selected as "parents" of a new generation. Then, genetic operators are applied to them to obtain the offspring. In this algorithm, it has been used the mutation, crossover, and selection operators. For the case of the three temperature ranges, individuals have three genes, and each of them represents the weighted contribution of the deep learning model diagnosis. The following Equation describes it.

```
Data: Population initialization: vector of 3 random weights (one for
      each temperature range)
Result: Weighted diagnosis for each temperature range
initialization;
while accuracy does no improve or N iterations have been carried do
    create a new mutated individual;
    combine this individual with the target vector (selected individual
      with best accuracy);
    evaluation of the new individual;
    evaluation of the whole set of individuals (accuracy);
end
```

Fig. 3 Pseudocode for the evolutive algorithm

$$TempRange_{diagnosis} = High\% * H_{prediction} + Medium\% * M_{prediction} + Low\% * L_{prediction} \tag{2}$$

The fitness of each chromosome is calculated by diagnosing the whole test set by weighting the predicted outcome of the three models with the value of the corresponding gene. The threshold for positivity is more than 50% membership in the positive class. The individuals with the best fitness (higher prediction accuracy) become the next generation's parents. The offspring are obtained by applying mutation, crossover, and selection operators in successive generations.

The mechanism explained above can be applied in the three-temperature range case. At a particular moment, we have three different CNNs, one for each temperature range that has been trained with its respective thermographies separated by temperature ranges. These models can be used to diagnose patients with breast cancer and healthy people. This diagnosis uses an output vector of two classes that gives a percentage belonging to one category or another. The final diagnosis of each model is the class with the highest rate. We can create a definitive diagnosis in its most basic version by averaging the percentage belonging to one category. For example, suppose the high-temperature model diagnoses a patient with 39% cancer, the medium-range model with 42%, and the low-temperature model with 73%. In that case, the final diagnosis could be: $\frac{1}{3}0.39 + \frac{1}{3}0.42 + \frac{1}{3}0.73 = 51.3\%$. However, averaging the models makes no sense as their performance is not the same diagnosis. So, by using the evolutive algorithm whose fitness function considers the number of hits when diagnosing with the test set, we can leverage the influence of the three models by balancing their weights. Figure 3 describes how the algorithm works for the particular case of separating the infrared image into three temperature ranges.

4 Results

4.1 Data scrapping and treatment of images

Before training the convolutional models, we need to download and preprocess the images. As they were unavailable as a unique dump, we created a scraper in Python. Scraping is a technique that retrieves unstructured data from a website and stores it in a structured way, Saurkar et al. [48]. We have downloaded 5,343 images, but we have decided only to work with the frontal perspective as they are the majority. As some images seem equal, we have decided to check that none of them was repeated by comparing them pixel by pixel and discarding the repeated ones. Some parts of the images added noise, so we removed the colour legend and body parts like shoulders or neck, leaving the breast and the surrounding parts of the body. To focus on the thermal activity of the breast, we have left only this part of the image and coloured the background black. Finally, we have cropped them to 215 × 538 size.

Then, we applied two different treatment image strategies. First, we created a dynamic image separation for the two use cases that work with the temperature ranges. This method takes the maximum and minimum values of each temperature matrix and creates three equal ranges: low, medium, and high. It is called dynamic because maximum and minimum temperatures change from one patient to another. These ranges suppose three thermographies, one representing the low temperatures, another representing the medium ones, and the last with the high range. In the second case of the three-colour channels, we have just made this separation considering the red, green, and blue channels.

The final dataset comprises 96 images of healthy people and 99 patients with cancer, 195 in total.

4.2 Evaluation of the four use cases

At this point, we have four possible use cases depending on how images have been treated: first works with the original thermographies, second with three obtained images that correspond to the low, medium, and high ranges of temperature, and another, similar to this but only with the two more influential ranges. Finally, we have a model that separates the infrared images depending on the colour channel.

4.3 The training stages

Datasets have been split into training, validation, and testing. The two first help to fine-tune the best hyperparameters for the models. The test set checks the model's actual performance once it has been trained. This work used an 85–15% division corresponding to the training/validation and test sets. The training/validation test also performs with an 80–20 proportion while applying a fivefold cross-validation strategy. We have randomized the instances of the subsets to avoid biases. Summarizing the training/validation subset comprises 135 images for training (66 sick people and 69 healthy) and 30 images (15 for each class). Then, the test subset will contain 15 images for each category (breast cancer vs. healthy).

We have performed a grid search strategy during the training stages to find the optimal hyperparameters for the deep learning models. This method finds the best value of the hyperparameters by combining different ranges of possibilities, [9]. These values are in the following Table 1.

Hyperparameters of the final models All models are trained with the same datasets obtaining the same hyperparameters. The four models have an input layer with the size of the images. Then depending on the use case, we can use a filter layer that checks the separation of range temperatures by three or two and the segmentation of color channels.

Metrics performance of the four models We have used the accuracy to check how well the models of the four use cases perform when diagnosing breast cancer. Results have been compiled in Table 2 and can be interpreted using the bias-variance trade-off, [8].

Table 1 Grid search values

Hyperparameters	Values
Random state	1, 13, 17, 42, 99
Batch size	10, 20, 25
Dropout	0.4, 0.6
Number of kernels	16, 32, 64
Kernel size	2×2, 3×3, 4×4
Learning rate	5×10^{-3}, 5×10^{-4}, 5×10^{-5}

Table 2 Performance of the four use cases

Proposed model	Test accuracy
Only thermography	91.33% $\pm$ 2.15
Three colour channels	79.49% $\pm$ 2.56
Three temperature ranges	**93.85% $\pm$ 2.92**
Two most influential ranges	87.37% $\pm$ 4.71

The results above show that the best model is the hybrid using the low, medium, and high-temperature ranges with an approximate accuracy of 93.85%. As this model comprises deep learning models and an evolutive algorithm, we are studying the different parts in-depth to check that there is no overfitting and understand its function better. The training for the best model lasted 5 h and 30 min. We used an AMD Ryzen Threadripper 2950X 16-Core Processor with 4 GeForce RTX 2080 Ti to complete the process.

Analyzing the best model in-depth We have used specificity, sensitivity, and precision to interpret the model in-depth. As these metrics take into account the false positives and false negatives, they are very used in the medical field, [34]. Specificity obtains the ratio between the number of true negatives and the total of those predicted as true negatives and false positives. Sensitivity makes the same calculation but uses false negatives instead of false positives. Precision has been calculated by dividing the number of true positives between the sum of true positives and false positives. The metrics are formally defined in the following Equations. Using the three possibilities allows obtaining a good criterion about the model performance as the last two consider false positives and false negatives, which are very useful in the medical area,

$$Specificity = \frac{true\ negatives}{true\ negatives + false\ positives} \tag{3}$$

$$Sensitivity = \frac{true\ positives}{true\ positives + false\ negatives} \tag{4}$$

$$Precision = \frac{true\ positives}{true\ positives + false\ positives} \tag{5}$$

In this case, a false positive is a person who has no breast cancer and has been wrongly diagnosed with the illness. When this happens, it supposes a loss of money in unnecessary

Table 3 Evaluation of the high-temperature range model

	Train	Validation	Test
Accuracy	87.67% ± 6.65	86.67% ± 12.47	89.23% ± 5.85
Specificity	89.18% ± 9.53	92.00% ± 5.58	88.00% ± 8.00
Sensitivity	86.10% ± 16.88	81.33% ± 27.24	90.53% ± 10.79
Precision	89.82% ± 8.00	91.81% ± 6.31	88.91% ± 9.45

Table 4 Evaluation of the medium temperature range model

	Train	Validation	Test
Accuracy	84.83% ± 6.27	76.67% ± 4.08	79.49% ± 2.56
Specificity	87.21% ± 11.56	88.00% ± 8.69	72.00% ± 10.95
Sensitivity	82.37% ± 8.18	65.33% ± 10.95	87.37% ± 7.06
Precision	87.30% ± 10.33	85.59% ± 7.07	75.48% ± 5.72

Table 5 Evaluation of the low-temperature range model

	Train	Validation	Test
Accuracy	76.33% ± 5.39	76.67% ± 2.36	72.31% ± 3.34
Specificity	70.16% ± 3.74	84.00% ± 5.96	57.00% ± 4.47
Sensitivity	82.71% ± 10.27	69.33% ± 10.11	88.42% ± 11.41
Precision	72.73% ± 3.28	81.87% ± 4.38	66.08% ± 0.82

treatments and health damages for the patients. A false negative is a patient who has cancer and has not been diagnosed. This issue could lead people to have many health problems. As the best model is formed by three convolutional models, one for each temperature range, we are studying them separately. Tables 3, 4, and 5 compile this information for the low, medium, and high-temperature range models. They show the average of five experiments and their standard deviation.

Looking at the tables above, we can interpret how the model performs considering the bias-variance trade-off against the accuracy. Bias values are correct, except for the low-temperature model, as human performance accuracy with the help of a CAD is near 83%, [28]. In cases of diagnosis without the help of a computer, specificity is 60.3%, and sensitivity is 81.5%, [53]. In terms of variance, only in the case of the medium temperatures seems slightly broad but not problematic. Compared with the former paper, it makes sense that the low-temperatures model performs a little worse as these images contain very little information about the thermography. For the later paper, our models are always better, in some cases much better.

The values of the other metrics (specificity, sensitivity, and precision) are consistent except for sensitivity for medium temperatures and all of them for low temperatures. The latter makes sense as the images of this model provide too little information so they could be biased depending on the data split. For high and medium temperatures, specificity is slightly better than accuracy. However, it should be highlighted that sensitivity has a significant standard deviation which confirms the problems with the biased subsets. This problem indicates that performance should be better with a dataset with more instances. In terms of interpretability, the better specificity shows a better performance minimizing the patients diagnosed with cancer that are healthy. This fact has an impact on medical expenditure.

Table 6 Contribution in the percentage of the temperature range models for calculating the final diagnosis

Model	Weight
High	34.57%
Medium	48.42%
Low	17.01%
Total	**100%**

Table 7 Weighted-average diagnosis

Metric	Final diagnosis
Specificity	97.00% ± 2.74
Sensitivity	90.55% ± 4.40
Precision	96.66% ± 3.06

Although the deep learning models present some issues due to the scarce data and the poor image representations of some cases, these problems could be solved by applying an evolutive algorithm for the final diagnoses. After using a weighted-average algorithm with the test dataset to the model, we can obtain the contribution of each temperature range CNN model and interpret which influences the most when calculating the final diagnosis. The best set of weightings thus obtained is compiled in Table 6.

As seen in the Table, the medium-range temperature influences the most in the workflow and contributes more when calculating the final diagnosis.

The other metrics for the diagnosis after using the weighted-average algorithm of the temperature range approach obtained with these weights are displayed in Table 7.

The approach using the three temperature ranges is better than those classifying thermographies by colour channel. We have to highlight that the accuracy is improved compared to human performance by about 10 points, [28]. By comparing the three metrics, again, we conclude that the model fails by diagnosing some people like cancer patients when they are healthy.

Comparison with different baselines To measure the accuracy of the best model, we have compared it with some well-known proposals in the computer vision field. In this case, we have used VGG16, VGG19, and Inception V3. VGG16 and VGG19 are two convolutional neural networks with 16 and 19 hidden layers, respectively, Simonyan and Zisserman [51]. Inception V3 was introduced by Szegedy et al. [52], and it is a CNN that uses convolutional factorization. The comparison of our best model with these baselines can be seen in Table 8.

Results demonstrate that our proposal improves the results of every baseline. In fact, all the baselines are far from obtaining our results in each metric.

Table 8 Comparison with baselines

Model	Accuracy	Specificity	Sensi5tivity	Precision
Our model	**93.85% ± 2.92**	**97.00% ± 2.74**	**90.53% ± 4.40**	**96.66% ± 3.06**
VGG16	69.23% ± 4.05	67.89% ± 4.95	70.75% ± 3.27	68.00% ± 6.71
VGG19	56.91% ± 6.46	57.43% ± 6.06	56.19% ± 7.19	51.33% ± 5.05
Inception v3	46.18% ± 4.76	27.17% ± 16.22	48.97% ± 2.36	84.00% ± 11.94

5 Conclusions and future works

The present work evaluates four different methods to diagnose breast cancer using thermographies. The different use cases depend on how the information is preprocessed and the usage of different techniques: deep learning models and evolutive algorithms. We have presented four preprocessing methods: use the whole thermography, segment it by the high, medium, and low-temperature ranges, same but only with the most influential (high and medium) and segment by RGB channels. The four different preprocessing methods lead to four diagnostic models: the first uses a regular CNN model, the second and third use one CNN model for each range of temperature and an evolutive algorithm for the final diagnosis, and the last one uses one CNN model for each colour channel and another evolutive algorithm. Results show that the best model is the one that segments the thermographies by temperature range which obtains an accuracy of 94% approximately. The contribution of the work is not only providing a model with good performance. Apart from providing an innovative method, we have demonstrated that some data-related problems can be solved by applying the evolutive algorithm. Another contribution is the interpretability we can obtain from the model, explaining which temperature range is the most efficient for diagnosis. In this way, we conclude that the medium-range temperatures are those that help the model in making an accurate diagnosis.

The main limitation of the approaches is related to the nature of the data. First, we have found that the dataset is scarce in some ways and results could be improved. In this way, we have also found some poor-quality images and repeated instances, which leads us to think that a more refined method when collecting the data would give us more insights. Another limitation related to a cancer use case is that we cannot measure the impact of false positives if we do not have new medical tests to know how the patient has evolved. By collecting this information, we can check if some of these false positives have evolved into cancer diagnoses. As we consider it very important, we proposed it as future work.

For future work, we started to compile our data to check the method with a more extensive and stratified dataset. This data will be associated with mammographies, so we could make an in-depth analysis of both tests with deep learning models. Also, we need to collect the medical tests of the same patients in different moments, so we can check if they are developing the disease or not. Another significant contribution could be the validate our predictions by a medical specialist in the area. This information lets us understand in detail the most critical parts of the thermography to make an accurate diagnosis and also which factors make the model fail. This contribution would be a big benefit for the field as many physicians have problems understanding and interpreting information in thermographies. The method could also be tested in other medical tests like electroencephalograms as they can be segmented by channels or brainwaves (alpha, beta, and theta) and the contribution of each weighted.

Acknowledgements The work leading to these results has received funding from the "Programa estatal de generación de conocimiento y fortalecimiento científico y tecnológico del sistema de I+D+i", in the context of the project "Cribado coste-efectivo de cáncer de mama mediante mamografía, ecografía y termografía" (*PID2019-110686RB-I00*).

We also want to thank Avanade Ibérica for providing a scholarship under the Avanade-UFV Artificial Intelligence Chair agreement.

Data availability Data is openly available in a public repository that does not issue DOIs.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Abdalla HEM, Esmail MY (2018) Brain tumor detection by using artificial neural network. 2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), pp 1–6
2. Abdallah YMY, Alqahtani T (2019) Research in medical imaging using image processing techniques. In Medical Imaging-Principles and Applications. IntechOpen
3. Afantenos S, Karkaletsis V, Stamatopoulos P (2005) Summarization from medical documents: a survey. Artif Intell Med 33(2):157–177
4. Ahuja AS (2019) The impact of artificial intelligence in medicine on the future role of the physician. PeerJ 7:e7702
5. Al Husaini MAS, HadiHabaebi M, Gunawan TS, Islam MR (2021) Self-detection of early breast cancer application with infrared camera and deep learning. Electronics 10(20):2538
6. Alshehri A, AlSaeed D (2023) Breast cancer diagnosis in thermography using pre-trained vgg16 with deep attention mechanisms. Symmetry 15(3):582
7. Bayomi N, Nagpal S, Rakha T, Fernandez JE (2021) Building envelope modeling calibration using aerial thermography. Energy Build 233:110648
8. Belkin M, Hsu D, Ma S, Mandal S (2019) Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proc Natl Acad Sci 116(32):15849–15854
9. Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. J Mach Learn Res 13(2)
10. Chebbah NK, Ouslim M, Benabid S (2023) New computer aided diagnostic system using deep neural network and SVM to detect breast cancer in thermography. Quant InfraRed Thermogr J 20(2):62–77
11. Daoud H, Bayoumi MA (2019) Efficient epileptic seizure prediction based on deep learning. IEEE Trans Biomed Circuits Syst 13(5):804–813
12. Domino M, Borowska M, Kozłowska N, Zdrojkowski Ł, Jasiński T, Smyth G, Maśko M (2021) Advances in thermal image analysis for the detection of pregnancy in horses using infrared thermography. Sensors 22(1):191
13. Egwom OJ, Hassan M, Tanimu JJ, Hamada M, Ogar OM (2022) An LDA–SVM machine learning model for breast cancer classification. BioMedInformatics 2(3):345–358
14. Ensafi M, Keyvanpour MR, Shojaedini SV (2022) A New method for promote the performance of deep learning paradigm in diagnosing breast cancer: improving role of fusing multiple views of thermography images. Heal Technol 12(6):1097–1107
15. Farooq S, Del-Valle M, Dos Santos MO, Dos Santos SN, Bernardes ES, Zezell DM (2023) Rapid identification of breast cancer subtypes using micro-FTIR and machine learning methods. Appl Opt 62(8):C80–C87
16. Fusco R, Piccirillo A, Sansone M, Granata V, Rubulotta MR, Petrosino T, Barretta ML, Vallone P, Di Giacomo R, Esposito E et al (2021) Radiomics and artificial intelligence analysis with textural metrics extracted by contrast-enhanced mammography in the breast lesions classification. Diagnostics 11(5):815
17. Ghafarpour A, Zare I, Zadeh HG, Haddadnia J, Zadeh FJS, Zadeh ZE, Kianersi S, Masoumzadeh S, Nour S (2016) A review of the dedicated studies to breast cancer diagnosis by thermal imaging in the fields of medical and artificial intelligence sciences. Biomed Res 27(2):543–552
18. Gogoi UR, Majumdar G, Bhowmik MK, Ghosh AK (2019) Evaluating the efficiency of infrared breast thermography for early breast cancer risk prediction in asymptomatic population. Infrared Phys Techn 99:201–211
19. Hawkes PW (2004) Advances in imaging and electron physics. Elsevier
20. Hofvind S, Ponti A, Patnick J, Ascunce N, Njor S, Broeders M, Giordano L, Frigerio A, Törnberg S (2012) False-positive results in mammographic screening for breast cancer in Europe: a literature review and survey of service screening programmes. J Med Screen 19(1_suppl):57–66
21. Houssein EH, Emam MM, Ali AA (2021) An efficient multilevel thresholding segmentation method for thermography breast cancer imaging based on improved chimp optimization algorithm. Expert Syst Appl 185:115651

22. Huang TS et al (1996) Computer vision: Evolution and promise. cern european organization for nuclear Research-reports-CERN, pp 21–26
23. Jasti N, Bista S, Bhargav H, Sinha S, Gupta S, Chaturvedi SK, Gangadhar BN (2019) Medical applications of Infrared thermography: A narrative review. J Stem Cells 14(1):35–53
24. Ji Y, Li H, Edwards AV, Papaioannou J, Ma W, Liu P, Giger ML (2019) Independent validation of machine learning in diagnosing breast Cancer on magnetic resonance imaging within a single institution. Cancer Imaging 19(1):1–11
25. Jiang Y, Edwards AV, Newstead GM (2021) Artificial intelligence applied to breast MRI for improved diagnosis. Radiology 298(1):38–46
26. Kadhim RR, Kamil MY (2023) Comparison of machine learning models for breast cancer diagnosis. IAES Int J Artif Intell 12(1):415
27. Kakileti ST, Madhu HJ, Krishnan L, Manjunath G, Sampangi S, Ramprakash HV (2020) Observational study to evaluate the clinical efficacy of Thermalytix for detecting breast cancer in symptomatic and asymptomatic women. JCO Glob Oncol 6:1472–1480
28. Keyserlingk JR, Ahlgren PD, Yu E, Belliveau N, Yassa M (2000) Functional infrared imaging of the breast. IEEE Eng Med Biol Mag 19(3):30–41
29. Khan AA, Arora AS (2021) Thermography as an economical alternative modality to mammography for early detection of breast cancer. J Healthc Eng 2021
30. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105
31. Lahane SR, Chavan N, Madankar M (2021) Classification of thermographic images for breast cancer detection based on deep learning. Ann Rom Soc Cell Biol 25(6):3459–3466
32. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
33. LeCun Y, Boser B, Denker J, Henderson D, Howard R, Hubbard W, Jackel L (1989) Handwritten digit recognition with a back-propagation network. Advances in neural information processing systems, 2
34. Lee Y, Kwon J, Lee Y, Park H, Cho H, Park J (2018) Deep learning in the medical domain: predicting cardiac arrest using deep learning. Acute and Critical Care 33(3):117
35. Ma J, Shang P, Lu C, Meraghni S, Benaggoune K, Zuluaga J, Zerhouni N, Devalland C, Al Masry Z (2019) A portable breast cancer detection system based on smartphone with infrared camera. Vibroengineering PROCEDIA 26:57–63
36. Morales-Cervantes A, Kolosovas-Machuca ES, Guevara E, Reducindo MM, Hernández ABB, Garc\'\ia MR, González FJ (2018) An automated method for the evaluation of breast cancer using infrared thermography. EXCLI J 17:989
37. Nogales A, Pérez-Lara F, Morales J, García-Tejedor ÁJ (2023) How Do Thermography Colors Influence Breast Cancer Diagnosis? A Hybrid Model of Convolutional Networks with a Weighted Average Evolutionary Algorithm. In: Arai K (ed) Intelligent Systems and Applications. IntelliSys 2022. Lecture Notes in Networks and Systems, vol 543. Springer, Cham. https://doi.org/10.1007/978-3-031-16078-3_17
38. Park HJ, Kim SM, La Yun B, Jang M, Kim B, Jang JY, Lee JY, Lee SH (2019) A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of breast masses on ultrasound: added value for the inexperienced breast radiologist. Medicine 98(3)
39. Pavithra PR, Ravichandran KS, Sekar KR, Manikandan R (2018) The effect of thermography on breast cancer detection. Syst Rev Pharm 9(1):10–16
40. Petrillo A, Fusco R, Di Bernardo E, Petrosino T, Barretta ML, Porto A, Granata V, Di Bonito M, Fanizzi A, Massafra R et al (2022) Prediction of breast cancer histological outcome by radiomics and artificial intelligence analysis in contrast-enhanced mammography. Cancers 14(9):2132
41. Raghupathi W, Raghupathi V (2014) Big data analytics in healthcare: promise and potential. Health Inform Sci Syst 2(1):1–10
42. Rim B, Sung N-J, Min S, Hong M (2020) Deep learning in physiological signal data: A survey. Sensors 20(4):969
43. Roslidar R, Rahman A, Muharar R, Syahputra MR, Arnia F, Syukri M, Pradhan B, Munadi K (2020) A review on recent progress in thermal imaging and deep learning approaches for breast cancer detection. IEEE Access 8:116176–116194
44. Russell SJ, Norvig P (2016) Artificial intelligence: a modern approach. Pearson Education Limited London, UK, Malaysia
45. Sánchez-Cauce R, Pérez-Mart\'\in J, Luque M (2021) Multi-input convolutional neural network for breast cancer detection using thermal images and clinical data. Comput Methods Prog Biomed 204:106045

46. de Santana MA, Pereira JMS, da Silva FL, de Lima NM, de Sousa FN, de Arruda GMS, de CF de Lima R, da Silva WWA, do Santos WP (2018) Breast cancer diagnosis based on mammary thermography and extreme learning machines. Res Biomed Eng 34:45–53
47. Sathish D, Kamath S, Prasad K, Kadavigere R (2019) Role of normalization of breast thermogram images and automatic classification of breast cancer. Vis Comput 35(1):57–70
48. Saurkar AV, Pathare KG, Gode SA (2018) An overview on web scraping techniques and tools. Int J Future Revolut Comp Sci Commun Eng 4(4):363–367
49. Seo H (2021) Infrared thermography for detecting cracks in pillar models with different reinforcing systems. Tunn Undergr Space Technol 116:104118
50. Silva LF, Saade DCM, Sequeiros GO, Silva AC, Paiva AC, Bravo RS, Conci A (2014) A new database for breast research with infrared image. J Med Imaging Health Inform 4(1):92–100
51. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
52. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, ..., Rabinovich A (2015) Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9)
53. Kakileti ST, Manjunath G, Madhu H, Lakshmi Krishnan HVR (2022) A Systematic Study to Evaluate the Benefit of using a Visualization Tool for Breast Thermography. Thermol Int 31(2):53–61
54. Storn R, Price K (1997) Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. J Global Optim 11(4):341–359
55. Tello-Mijares S, Woo F, Flores F (2019) Breast cancer identification via thermography image segmentation with a gradient vector flow and a convolutional neural network. J Healthc Eng, 2019
56. Wiley V, Lucas T (2018) Computer vision and image processing: a paper review. Int J Artif Intell Res 2(1):29–36
57. Wu T, Sultan LR, Tian J, Cary TW, Sehgal CM (2019) Machine learning for diagnostic ultrasound of triple-negative breast cancer. Breast Cancer Res Treat 173(2):365–373
58. Zuluaga-Gomez J, Al Masry Z, Benaggoune K, Meraghni S, Zerhouni N (2021) A CNN-based methodology for breast cancer diagnosis using thermal images. Comput Methods Biomech Biomed Eng: Imaging Vis 9(2):131–145