IL NUOVO CIMENTO **44 C** (2021) 138
DOI 10.1393/ncc/i2021-21138-x

COMMUNICATIONS: SIF Congress 2020

Explainability of a CNN for breast density assessment

C. SCAPICCHIO([1]), F. LIZZI([1])([2]) and M. E. FANTACCI([1])

([1]) *INFN, Sezione di Pisa and Dipartimento di Fisica, Università di Pisa - Pisa, Italy*
([2]) *Scuola Normale Superiore - Pisa, Italy*

received 15 January 2021

Summary. — Deep neural network explainability is a critical issue in Artificial Intelligence (AI). This work aims to develop a method to explain a deep residual Convolutional Neural Network able to automatically classify mammograms into breast density classes. Breast density, a risk factor for breast cancer, is defined as the amount of fibroglandular tissue compared to fat tissue visible on a mammogram. We studied the explainability of the classifier to understand the reasons behind its predictions, in fact with a deep multi-layer structure, it acts like a black-box. As there is no well-established method, we explored different possible analyses and visualization techniques. The main obtained results were the achievement of a performance improvement in terms of accuracy and a contribution to assess trust in the model. This is fundamental for a potential application in clinical practice.

1. – Introduction

A drastic reduction of breast-cancer–related death could be seen thanks to early tumors diagnose, since the introduction of screening programs based on Field Digital Mammography (FFDM) examination [1]. Mammographic density is defined as the amount of dense tissue in the breast, visible on a mammogram, compared to the amount of fatty tissue. The denser glandular tissue attenuates radiation more with respect to fat tissue that shows up darker because less attenuating. The importance of assessing breast density lies in three key points: a) the possibility of drawing up a personalized dosimetric index [2], b) the overlapping of dense tissue which causes a masking effect and c) breast density itself is an independent risk factor for breast cancer. Each exam is evaluated by a radiologist and assigned to a density class label, following the standard reported in the BIRADS (Breast Imaging Reporting and Data System) Atlas [3]: entirely fatty (A), scattered areas of fibroglandular density (B), heterogeneously dense (C) and extremely dense (D). Automatic methods, based on AI, have been developed in order to make the breast density classification reproducible and avoid the inter- and intraobserver variability. The aim of our work is to provide a method to explain the behaviour of a breast density

classifier based on a deep residual Convolutional Neural Network (CNN). It is able to classify the image by extracting related features, but acting as a black box with its deep multi-layer nonlinear structure. Since, in the field of AI, explainability is developing as a border issue and a new subject of research, there is no a well-established methodology to apply this concept to a deep neural network. We then explored possible ways to describe the internal processes that are responsible for the final classification score.

2. – Materials and methods

2˙1. *Data.* – Deep Learning is a data-driven approach. Learning directly from data, the network performance has a strong dependence on the dataset which it is trained on.

We used 1662 mammographic exams made available to us by the "Azienda Ospedaliero-Universitaria Pisana" (AOUP) and annotated by a radiologist with one of the four BIRADS density classes (A, B, C, D), which represents the ground truth [4]. The images have been acquired with GE Senograph DS imaging systems. The data were also randomly split into training set (80%), validation set (10%) and test set (10%).

2˙2. *ResNet model.* – The breast density classifier is a CNN based on a residual architecture. It is made of 41 convolutional layers, organized in residual blocks. It is described in detail, along with the hyperparameters chosen for the training, in a previous work [4]. To train, fit and evaluate the CNN we used the open-source software package Keras. The hardware used for the training was a K80 Nvidia GPU, made available by "Istituto Nazionale di Fisica Nucleare" of Pisa.

After a supervised training of the algorithm on the training set of labelled data, the classification performance was evaluated on the test set using *accuracy*, *recall* and *precision* as figures of merit.

2˙3. *Explainability.* – A deep convolutional neural network has a main drawback related to the difficulty in interpreting its internal processes. When a new unseen data sample is given to the model, it is hard to capture what makes it arrive at a particular classification decision. In association with high-accuracy results, understanding the reasons behind predictions is equally quite important in assessing trust in the model [5]. Explainability is defined as the ability to provide a qualitative understanding of deep neural networks for which it is possible to explain why they predict what they predict. There is not a well-established standard method to explain a DNN. Therefore, two methods have been explored. The first one consisted in analyzing how the output varies with the input. This method covered the study of how the preprocessing of the images influences the classifier performance. The second method consisted in visualizing what the network has learned through specific techniques.

2˙4. *How the output varies with the input.* – In the data pre-processing phase, we started with the background removal. An algorithm based on "marching squares" algorithm was implemented to automatically draw the breast skin edge line and crop the image at the line. The mammograms were then inspected one by one to exclude some visibly problematic images from the original dataset. The last significant step was the pectoral muscle segmentation. The muscle has pixel intensities similar to that of dense tissues and, therefore, it may be a confounding factor. A segmentation algorithm has been developed to automatically detect and remove the muscle.

2˙5. *Off-line visualization.* – The second strategy we followed was the adoption of visualization techniques. These techniques aim to identify which discriminative pixels in the image influence the final prediction. We want to verify through an image that the attention of the network is focused on the dense regions of the mammogram to demonstrate that the classification result is given on the bases of the feature we expect. In fact, even if right, the outcome of a classifier could be unreliable. For instance, because of bias in the training set, the network could "look" at wrong features [5]. Visualization tools have been applied after training the model without altering its architecture. We used the *visualizecam* utility function, provided by Keras to generate a gradient-based class activation map, which is an image indicating the input regions whose change would most contribute towards maximizing the output. This function is based on the grad-CAM specific technique [6]. The Grad-CAM $L^c_{Grad\text{-}CAM}$ is a weighted combination of forward activation maps, followed by a ReLU:

$$L^c_{Grad\text{-}CAM} = ReLU(\sum_k \alpha^c_k A^k), \quad \text{with} \quad \alpha^c_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}}. \tag{1}$$

The neuron importance weight α^c_k highlights the "importance" of the feature map k for a target class c. To obtain these weights, we first compute the gradient of the score for class c, y^c, with respect to feature map activations A^k of a convolutional layer, *i.e.*, $\frac{\partial y^c}{\partial A^k}$. Then these gradients flowing back are global-average-pooled over the width and height dimensions (indexed by i and j, respectively). The places where the gradient is large let us exactly define the region that has a large impact on the final score decision. Thus we obtain a heatmap that indicates which areas of an image are being used by the model for discrimination among classes.

3. – Results

We established some pre-processing steps to optimize the training data. By a comparison between the main figures of merit, we observed a significant improvement in the classifier performance if we use preprocessed data (accuracy: 83.1%, recall: 80.1% and precision: 87.9%) with respect to using the images in their original form (accuracy: 75.3%, recall: 72.1% and precision: 76.4%). Then, we obtained the heatmaps through the grad-CAM technique. The maps have been generated using input images of the four classes. After generating the maps for all the images in the test set, their evaluation has been done qualitatively, which means by matching what is highlighted in the map with the dense regions in the original images. We observed that the matching is acceptable for images with B, C and D class labels (fig. 1). For A class mammograms the maps activate almost always at the edge of the breast. This is reasonable because the A class is the one corresponding to the lowest density and it is as if the classifier did not recognize any dense region and focused its attention on a different feature, such as the edge. This demonstrated that the "attention" of the classifier is focused on the dense region, as we expected.

The last result regards the pectoral muscle removal. After applying the above-mentioned segmentation algorithm on the original images, we obtained a new dataset in which the muscle has been replaced by the average gray value. We trained the CNN model both on images with (accuracy: 79.9 %, recall: 78.1% and precision: 81.1%) and without the pectoral muscle (accuracy: 82.0%, recall: 80.3% and precision: 83.3%). By

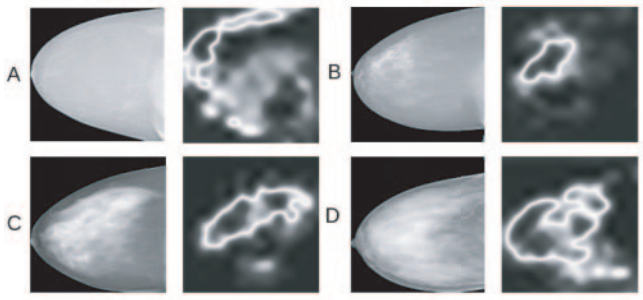

Fig. 1. – Example of comparison between the original image and the heatmap for the 4 classes.

comparing the Grad-CAMs generated in the two cases, we observed that, in most cases, muscle removal helps in guiding the network to focus on the right breast area and after segmentation the pixels forming part of the muscle are no longer highlighted.

4. – Conclusion

In this work a new method to optimize a CNN classifier for breast density assessment and, especially, to explain its behaviour has been explored. It has been discovered that an appropriate data preparation is essential to improve the classification performance. From the starting results obtained in the first work of CNN development [4], where no images preprocessing step was considered, we obtained a significant improvement in terms of figures of merit. Moreover, through the heatmaps, it has been confirmed that the "attention" of the algorithm is focused on the expected regions of the image, *i.e.*, the dense areas. We also found that segmenting and removing the pectoral muscle helps the model in focusing on the dense regions of the breast and in improving the performance also in terms of figures of merit. This study has been fundamental to assess trust in a developed model, which is crucial for the application of these deep algorithms in medical routine, and to open a new way of applying the explainability concept to a deep neural network, which could be improved in the future.

REFERENCES

[1] Muhammad M., *An Introduction to Medical Physics* (Springer Nature) 2018, Chapt. 7, pp. 199–220, ISBN:9783319615387.
[2] Traino A. C. *et al.*, *Eur. Radiol. Exp.*, **1** (2017) 28.
[3] Sickles E. *et al.*, *ACR BI-RADS® Atlas*, in *Breast Imaging Reporting and Data System* (American College of Radiology, Reston, VA) 2013, pp. 39–48.
[4] Lizzi F. *et al.*, *Residual Convolutional Neural Networks to Automatically Extract Significant Breast Density Features*, in *International Conference on Computer Analysis of Images and Patterns* (Springer) 2019, pp. 28–35.
[5] Ribeiro M. T. *et al.*, *"Why should I trust you?" Explaining the predictions of any classifier*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (ACM) 2016, pp. 1135–1144.
[6] Selvaraju R. R. *et al.*, *Grad-cam: Visual explanations from deep networks via gradient-based localization*, in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE) 2017, pp. 618–626.