UDC 004.932

**Kamguia F.H., Goncharova A.B.**

# Determination of malignancy of neoplasm from mammograms

**1. Introduction.** Cancer in general and breast cancer in particular has been a great threat to health and life for many years now [1]. GLOBOCAN data from 2020 shows that female breast cancer has exceeded all other cancers in terms of new cases, with an estimated 2.3 million new cases (11.7%) [2].This can explain the increase in scientific works for early diagnosis, among which much work has been done on the classification of mammograms. With the coming of convolutional neural networks (CNN), this technology has been applied in computer vision problem resolution with impressive results: object detection [3, 4]. In the classification of mammograms with success, or in the detection of neoplasms in mammograms [5]. This paper proposes to determine the malignancy of neoplasm in a given mammogram using convolutional neural network. We successfully trained the model with pre-processed images and obtained state-of-art results. The obtained result from the model permits us to show area of interest on the input image and give a degree of malignancy expressed in percentage.

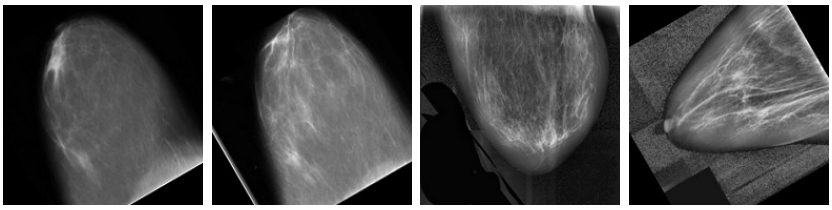
**2. Problem statement.** Determining the malignancy of neoplasm in a medical image (mammogram), by building a model based on CNN.

**3. Dataset.** The data set used in our experiment was derived from the Digital Database for Screening Mammography, a publicly available data set maintained collaboratively by the University of South Florida [6, 7]. Which has both mediolateral oblique (MLO) and craniocaudal (CC) views of each breast (all gray scale images). Each image is tagged with text-based disease definition labels (where each image can have multi-level labels) derived from radiology reports. Data augmentation and contrast-limited adaptive histogram equalization was used to pre-process these images. The obtained data set has 13128 images resized

*Kamguia Feukwi Herve* – graduate student, Saint-Petersburg state University; e-mail:st093241@student.spbu.ru, phone: +7(911)9301950

*Goncharova Anastasia Borisovna* – Ph.D., Associate Professor, Saint-Petersburg state University; e-mail: a.goncharova@spbu.ru, phone: +7(812)4284726

to $277 \times 277$ pixels [8].The images were randomly divided into training, verification, and testing sets (60%, 20%, and 20% of the full data set), with the testing set not used in the training phase.



(a) (b) (c) (d)

**Figure 1.** Examples of mammographic images of Benign (a-b) and Malignant lesions (c-d)

**4. Model.** With CNN models, convolution layers extract data from images to learn features (characteristics). Pullup reduces the dimensions of the receptive fields as inputs to each layer, however it goes deeper (in dimension) to learn more precisely what features are being identified by the convolution network. In order to counter the problem of overlapping tissues, CNN uses a to-and-fro learning process (forward and backward learning) [9]. In order to build our model we used the TensorFlow library which is built on the Keras library. TensorFlow ease the development of machine learning models with the help of its sequential API. TensorFlow is one of the most used libraries in the recent years to build machine learning models thanks to its parallelism, graphical support and scalability [10, 11].

The proposed model is a sequential set of sixteen layers. The model consists of 3 Conv2D layers, 3 MaxPooling2D layers, 6 ModuleWrapper layers, 2 Dense layers,1 Flatten layer and 1 Rescaling layer. The Rescaling layer takes in the input image and scale it to the desired dimension in our $244 \times 244$ pixels. It is worth mentioning that the training and validation dataset were autotune using the TensorFlow autotune function to ensure that the dataset does not become a bottleneck while training the model, there were later on normalised in the range of (0,1) which is an ideal range for the model. In a model with backpropagation process batch normalisation can be used to regularize inputs. It also helps avoid overfitting. Batch normalisation layers were added after each activation layer.Adam Optimization Algorithm was used with a learning rate of

0.001, batch size of 32, 10 Epochs and other default values. The result of the CNN model is shown in Figure 2.

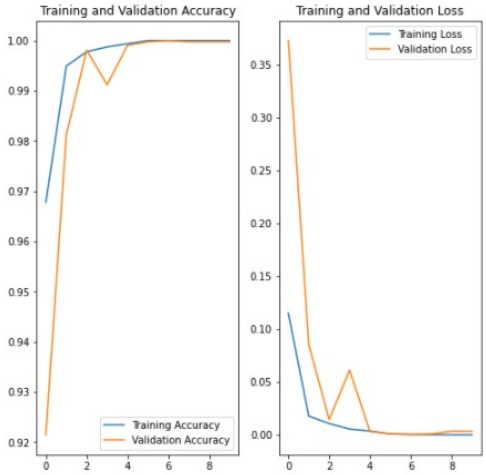



**Figure 2.** Accuracy and Loss functions value on the training and validation sets

**5. Experimental results and discussion.** Table 1 illustrates the results of applying the model to the test data set.

**Table 1.** Classification results of the proposed CNN model

| | Predicted Classes | | |
|---|---|---|---|
| True Classes | | Malignant | Benign |
| | Malignant | 1431 | 2 |
| | Benign | 1 | 1192 |

A saliency map of some images is shown in Figure 3, which allows us to estimate what each area contributes to the result of our model. Diagnostic medical characteristics obtained are as follows; sensitivity of CNN model determines the ability to identify breast cancer patients - 0.998; specificity determines the ability to identify patients with a benign breast mass - 0.999; predictability of a positive result - 0.999; predictability of a negative result - 0.998; accuracy - 0.999. Neural network model characteristics obtained are as follows; precision - 0.9989, recall - 0.9988. An analysis of the model's quality revealed that the proposed CNN model

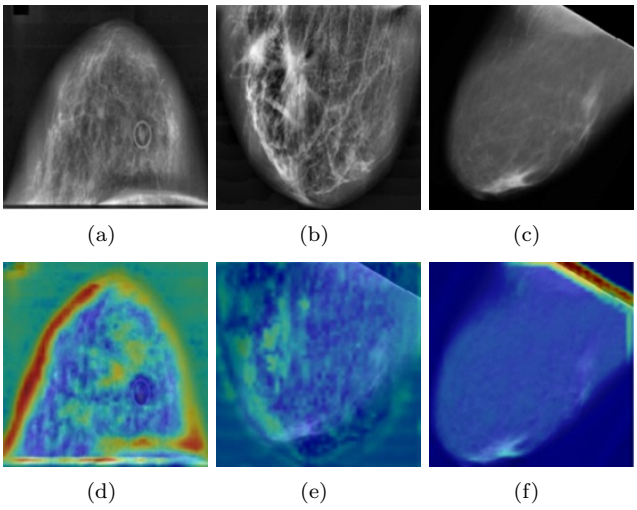



**Figure 3.** Example of initial mammograms (Malignant (a,b) and Benign (c)) and corresponding solidity maps (Malignant (d,e) and Benign (f))

classified benign and malignant images with a very high degree of accuracy.

**6. Conclusion.** This paper presents the architecture of an ultra-precise neural network for solving the problem of determining the oncology of a neoplasm by mammograms. The use of batch normalisation layers and dataset cache in our model helped overcome the problem of bottle neck, which is common in medical computer vision problems. Proposed CNN model showed outstanding results as the calculation of sensitivity, specificity, positive and negative values can show.

## References

1. Bray F., Laversanne M., Weiderpass E. et al.The ever-increasing importance of cancer as a leading cause of premature death worldwide// Cancer. In press. 2021. Vol. 127. No 16. P. 3029–3030.

2. Sung H., Ferlay J., Siegel R. L.et al. Jemal A., Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality

Worldwide for 36 Cancers in 185 Countries // CA: A Cancer Journal for Clinicians. 2021. Vol. 71. No 3. P. 209–249.

3. Girshick R., Donahue J., Darrell T. et al. Rich feature hierarchies for accurate object detection and semantic segmentation // Proc. of the Int. Conf. IEEE on computer vision and pattern recognition. 2014. P. 580–587.

4. Carneiro G., Nascimento J., Bradley A. P. Unregistered multiview mammogram analysis with pre-trained deep learning models // In International Conference on Medical Image Computing and Computer-Assisted Intervention. 2015. P. 652–660.

5. Agarwal R., Diaz O., Yap M. H. et al. Deep learning for mass detection in Full Field Digital Mammograms // Comput. Biol. Med. 2020 Vol. 121. P. 103–774.

6. Bowyer K., Kopans D., Moore R. et al. The Digital Database for Screening Mammography // Proc. of the 5th Int. Workshop on Digital Mammography. 2001. P. 212–218.

7. Heath M., Bowyer K., Kopans D. et al. Current status of the Digital Database for Screening Mammography in Digital Mammography // Proc. of the 4th Int. Workshop on Digital Mammography. 1998. P. 457–460.

8. Lin T., Huang M.-L. Dataset of Breast mammography images with Masses // Mendeley Data. 2020. Vol. 5. [Internet resource]: `URL:http://medicalresearch.inescporto.pt/breastresearch/index.php/Get_INbreast_Database` (date: 02.09.21).

9. Preet B., Sharma K. Classification of mammogram images by using CNN classifier // Proc. of the Int. Conf. on Advances in Computing, Communications and Informatics (ICACCI). 2016. P. 1743–2749.

10. TensorFlow. [Internet resource]: `URL:https://www.tensorflow.org` (date: 02.09.21).

11. Keras. [Internet resource]: `URL:https://keras.io` (date: 12.12.21).