Deep Learning Based Masses Segmentation Method in Mammograms

Tiberiu Marita, Raluca Brehar, Razvan Itu, Paul Trif, Delia Mitrea
Computer Science Department,
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
e-mail: Tiberiu.Marita@cs.utcluj.ro

***Abstract*—The current paper presents a method for automatic segmentation of potentially malignant abnormalities in 2D mammographic images based on pixel level classification using convolutional neural networks (CNN). Experiments were performed for masses segmentation by training U-Net models on the CBIS-DDSM dataset using original algorithms for data preprocessing in order to obtain an augmented and balanced training dataset and post-processing for semiautomatic evaluation. In terms of masses detection rates, a FPPI (False Positive Per Image) rate of 3.65 and a sensitivity (True Positive Rate – TPR) of 0.77 were obtained on the whole test data set, which are closed to the ones reported in the literature, proving that the proposed method has the potential to be integrated in mammography CAD applications for assisting screening programs in which the a main goal is a high true positive rate and a small false negative rate.**

Keywords – 2D mamograms; preprocessing; segmentation; automatic masses detection; deep learning; convolutional neural networks

I. INTRODUCTION

Breast cancer is one of the main diseases that affect women's life and is amongst the main causes of death related to cancer, if not discovered in the early stages. The pathology has an asymptomatic early phase that can be detected through mammography. The WHO (World Health Organization) concluded that 23% of cancers are breast cancer and 14% lead to death [1].

The main hardship in the detection of potentially malignant abnormalities (especially masses and micro-calcifications) is due to the acquisition process of the 2D mammograms (digitalized or film), meaning that a 3D structure is projected on a 2D plan, which leads to increased noise due to tissue overlapping, especially for dense glandular tissues. Therefore, 10% to 15% of cancers are invisible on mammograms, as the dense glandular tissue might hide 30%-50% of cancers [2].

The double interpretation (independent and simultaneous visualization of the mammograms by two radiologists) has been enforced as a standard in most of the screening programs, thus decreasing false-negatives detection [3]. This technique requires more time and additional expenses, as a digitized mammogram usually averages above 10 megapixels.

As an alternative to double blind interpretation, computer tools such as CADe (Computer- Aided Detection) or CADx (Computer Aided Diagnostics) can be used to provide a second opinion when mammograms are interpreted by a single radiologist. Amongst the limitations of these methods we mention: the high detection rate of false positives for CADe [4] or their limited-acceptance in clinical practice for CADx [5]. Most of the 2D mammography CAD Tools (3D mammography [6] is not the subject of the current paper due to limited access to such a technology) are able to detect abnormalities (as masses and calcification) based on traditional pattern recognition and machine learning algorithms as [7], [8], and [9] and are focused on reducing the false negative detection rate. Only a few tools such as ProFound AI, and SecondLook from iCAD [10] claim to use state of the art approaches based on Artificial Intelligence (AI) and Deep Learning (DL).

Recently, deep learning methods based on convolutional neural networks (CNN) have proven superior performance, including the medical imaging field. Instead of manually selected mathematical or heuristic features, these algorithms can automatically generate discriminative features at different levels of abstraction.

A hybrid mass classification method that combines feature extraction with a CNN network and a traditional SVM type classification method is presented in [11]. As training dataset for the neural network, the INbreast[12] data set was used, containing both mammograms digitized from films with a color depth of 8 bits/pixel and digital mammograms (16 bits/pixel). In [13] and [14] a method for mass detection is presented in several stages: in the first phase hypotheses are generated through a hybrid approach in which deep convolutional networks cascade with random decision tree type classifiers, followed by a segmentation stage based on deep learning models and a refinement stage using "level-set" methods. The authors report a true positive detection rate TPR = 0.96 at 1.2 FPPI (false positive per image) on the INbreast data set and TPR = 0.75 at 4.8 FPPI on the DDSM-BCRP data set [15].

A mass segmentation method that combines FCN (Fully Convolutional Networks) with the CRF method (Conditional Random Fields) and the contradictory learning technique (adversarial learning) is presented in [16]. The authors present comparative results obtained on the INbreast and DDSM-BCRP datasets between the various combinations of the three techniques used and other methods from the CNN-based literature. A method for mass detection is presented using a RetinaNet type deep convolutional network in [17]. The authors also use the transfer learning technique by which the weights of the network are pre-trained on one dataset and used in training and testing on another dataset. The authors also present a comparative analysis in terms of the FPR vs. FPPI metric of their own method in the context of the transfer learning technique on the INbreast dataset and on GURO dataset (their own data set), as well as compared with other segmentation methods based on deep networks or traditional methods on different datasets (INbreast, DDSM). In [18], a method for masses and micro-calcifications segmentation using dedicated neural network architectures (AlexNet, VGGNet, GoogleNet, ResNet) is presented. For each type of

network, a pixel classification model with 3 classes (normal, mass or calcification) was trained using the CBIS-DDSM dataset [19], with patches of size 224x224 extracted from the native resolution images around abnormalities marked in the binary masks (Ground Truth - GT), augmented by rotations and random mirroring operations. The predictions were made using the Class Activation Mapping (CAM) technique [20]. The authors reported a maximum accuracy of 92.42% on calcifications for the VGGNet architecture, a maximum accuracy of 95.06% on masses for the GoogleNet architecture and a maximum overall accuracy of 92.53% for VGGNet architecture without mentioning whether the evaluation was done at pixel or abnormality levels.

In the current paper a generic method for automatic segmentation of potentially malignant abnormalities in 2D mammographic images based on the deep learning paradigm is proposed. The method is generic and can be applied for the detection of both type of abnormalities: masses and calcifications. However in the current paper only experiments and results for masses segmentation are presented.

A preprocessing step of the training data set is proposed that provides $2^n x 2^n$ size (n configurable) image patches (ROIs), centered around the annotated GT (Ground Truths) from the training dataset. In order to prevent over-fitting, data augmentation is performed by generating an additional set of 4 images cropped for the same binary mask, where the centers of the $2^n x 2^n$ ROIs are shifted by a random value.

For the abnormalities segmentation a pixel level classification model based on the U-Net CNN architecture is trained on the downscaled ROIs (by a ratio of 4). Different variations of the model were tested by modifying the loss function type (binary cross-entropy, dice coefficient loss) and its metrics (accuracy, dice coefficient). To increase the trained model's performance balancing the training set is performed by giving different weights to black and white pixels (there are several times more black pixels compared to white pixels in the training images).

For the image segmentation step, the test images are divided in patches of the same resolutions ($2^n x 2^n$ ROIs further down sampled by a ratio of 4) using a non-overlapping sliding window scheme. The prediction map is generated by applying the inference with the obtained models for each patch and the full test images are resembled from the binarized predictions.

For evaluation of the results, both pixel level segmentation and abnormalities level detection rates are estimated in a semi-automatic manner.

II. The Architecture of the Proposed Segmentation Method

The architecture of the proposed method is presented in figure 1. The architecture is generic and can be duplicated for both types of abnormalities detection: masses and calcifications. A publicly available dataset (CBIS-DDSM [19]) that contains digitized film mammograms was used for training the pixel level classification models. Deep learning models based on the U-Net Convolutional Neural Network (CNN) architecture were trained and tested.

The system is composed of an off-line phase for training the models and an on-line phase to generate the prediction map for the images from the test set. Processing sequences are the same for both types of abnormalities, but the training process is conducted on different set of images and are as follows:

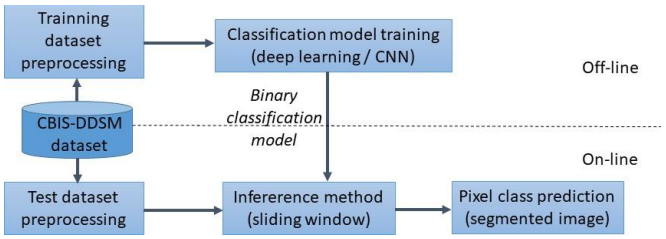


Figure 1. Architecture of the proposed method.

The main steps of the method are:

1. Preparing the dataset for the training phase
2. Training the pixel level classification models
3. Preprocessing the test dataset in a similar manner as the training set
4. Segmentation phase: generating the prediction map for the pixel level segmentation on the test images.

III. Automated Masses Segmentation

A. Data preprocessing

The CBIS-DDSM [19] dataset contains training and testing data for two types of abnormalities: masses and calcifications. The images are kept in DICOM files and are structured compliant to the standard [27]: Calc-Test Full Mammogram Images (DICOM), Calc-Test ROI and Cropped Images (DICOM), Calc-Training Full Mammogram Images (DICOM), Calc-Training ROI and Cropped Images (DICOM), Mass-Test Full Mammogram Images (DICOM), Mass-Test ROI and Cropped Images (DICOM), Mass-Training Full Mammogram Images (DICOM) and Mass-Training ROI and Cropped Images (DICOM).

The "Training Full" folders contain grayscale images (16bits/pixel) with the original mammogram. "Training ROI and Cropped" contain files with the binary mask having the abnormality marked (8 bits/pixel) and the grayscale image of the rectangular region of interest (ROI) that bounds the abnormality which has no use for our method and are discarded. Most of the binary masks and grayscale ROIs are in the same folder in the folder tree, and their numbering (000000 or 000001) is random, the only way to differentiate them being the field that holds the information regarding the number of bits per pixel. There have been solitary cases where the binary mask and the ROI were in the same parent directory, but in different subfolders. There is a possibility for a "Full" mammographic image to have many abnormalities, which leads to multiple binary masks images. For the test dataset there is an analogous structure for "Test Full" and "Test ROI and Cropped".

An algorithm was used to restructure the datasets, with the following steps:

- Parsing the folder tree for the abnormality (masses or calcifications);
- Renaming the images with distinct names which index the patient's case identifier (*F XXXXXX*.dcm for the "Full" images, M XXXXXX*.dcm for the binary masks), structured in folders that are abnormality specific and phase specific (training, testing) ;
- Deleting the solitary cases where the "Full" image and the binary mask have different spatial dimensions.

On the 1253 images with distinct annotated mass regions available in the training dataset, there have been done some preprocessing steps, with the main purpose of selecting the region of interest (ROI) that will be fed to the network, together with the corresponding binary mask, in the training process.

The idea behind the process implies that a rectangle that bounds the abnormality in every binary mask and its center point are found. Then, an image of size $2^n x 2^n$ pixels is cropped from both the binary mask and from the corresponding ”Full” grayscale image, trying to center it in the bounding rectangle’s center. If the $2^n x 2^n$ ROI image extends outside the border of the binary mask and its corresponding grayscale image, the $2^n x 2^n$ ROIs are cropped as much as the border allows, keeping the abnormality as much as possible contained in the cropped image.

This results in a $2^n x 2^n$ binary mask containing only the abnormality, and then the corresponding grayscale ROIs cropped from the ”Full” image. For data augmentation and to prevent overfitting, an additional set of 4 images are cropped for the same binary mask, where the center of $2^n x 2^n$ ROI is shifted by a random value, as seen in figure 2.

Before passing the data to the training process, the $2^n x 2^n$ images are downscaled by a ratio of 4 in order to avoid overloading the hardware resources used in the training process (single RTX 2080 Ti GPU) and the pixel intensities are normalized to the [0..1] interval

The pseudo-code for the binary-mask and corresponding grayscale ROI pairs generation process for each abnormality is presented below:

*I. For each **mask** in **binary_masks**:*
- *a. Read the corresponding **grayscale_image***
- *b. If **mask.dimensions==grayscale_image.dimensions** and **patch_dim < dimensions**:*
  1. ***Bounding_rect = get_abnormality_contour** (from mask)*
  2. *Get **Bounding_rect** center*
  3. *Crop a square region of **patch_dim** size around the **center***
  4. *Crop other **random_images** (4, for the proposed method) - square regions, where the center from (3) is moved by a random offset*
  5. *Apply the same cropping to the **grayscale counterpart** for each patch cropped from the **binary mask***
  6. *Save the cropped **ROI** and **mask** patches*

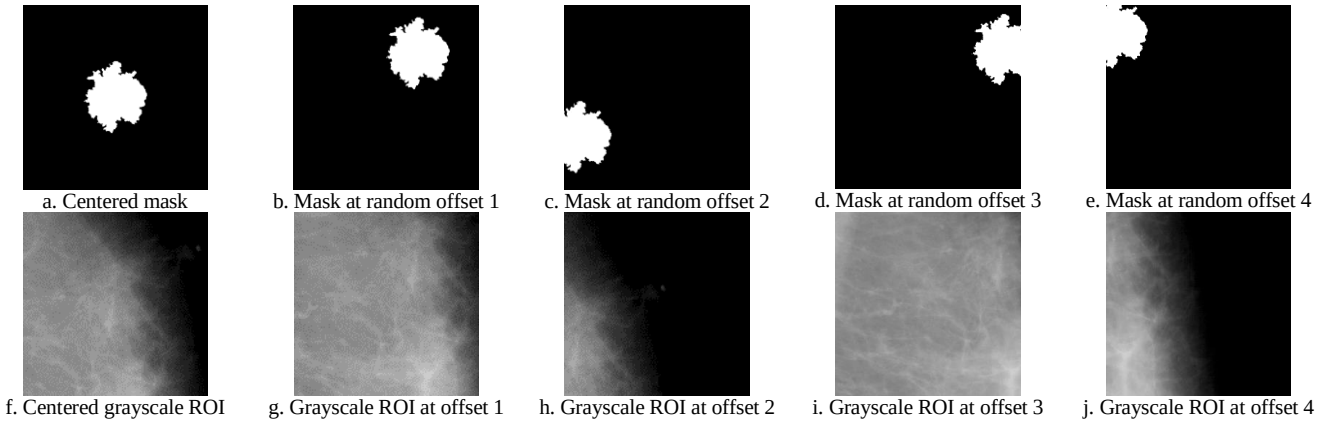


Figure 2. Ilustration of binary-mask and grayscale ROI pairs genereation process for each abnormality from the training dataset.

B. Training the pixels classification model

The architecture that has been used for the training process is U-Net [22] (presented in figure 3), which is based on an ”encoder-decoder” architecture, with 7 encoding and 7 decoding layers (including the final output layer). The main operation that supports a Convolutional Neural Network (CNN) is the convolution, which downscales the size of the image and leads later on to activation maps. These are then processed and transformed using nonlinear functions (usually, Rectified Linear Unit (ReLU) activation functions).

This U-net model has 91.238.241 trainable parameters and 24.256 non-trainable ones. Different variations of the model were tested by modifying the loss function type (binary cross-entropy, dice coefficient loss) and its metrics (accuracy, dice coefficient). The optimizer used was Adam and the network was set to train for a maximum of 100 epochs, with a batch size of 8. The initial number of output filters of the convolution was 32. Another important factor that contributed to the performance increasing the of the trained model was balancing the training set (giving different weights to black and white pixels, taking into account that there are several times more black pixels compared to white pixels). The initial training set was split into training and validation with a ratio of 85% to 15%, resulting in 5325 training samples and 940 validation samples. Besides the centered abnormality, each image fed to the network also had another four counterparts where the center of the rectangle bounding the abnormality has been moved by a random offset, as presented in section III.A. This step represents the augmentation applied on the initial 1253 images.

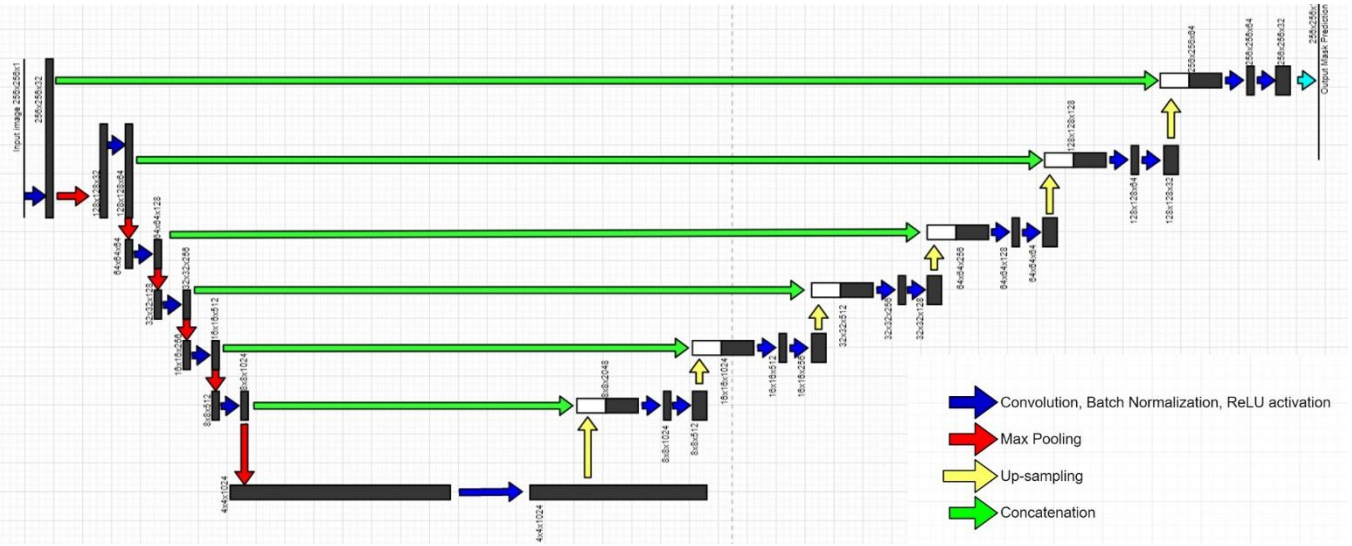
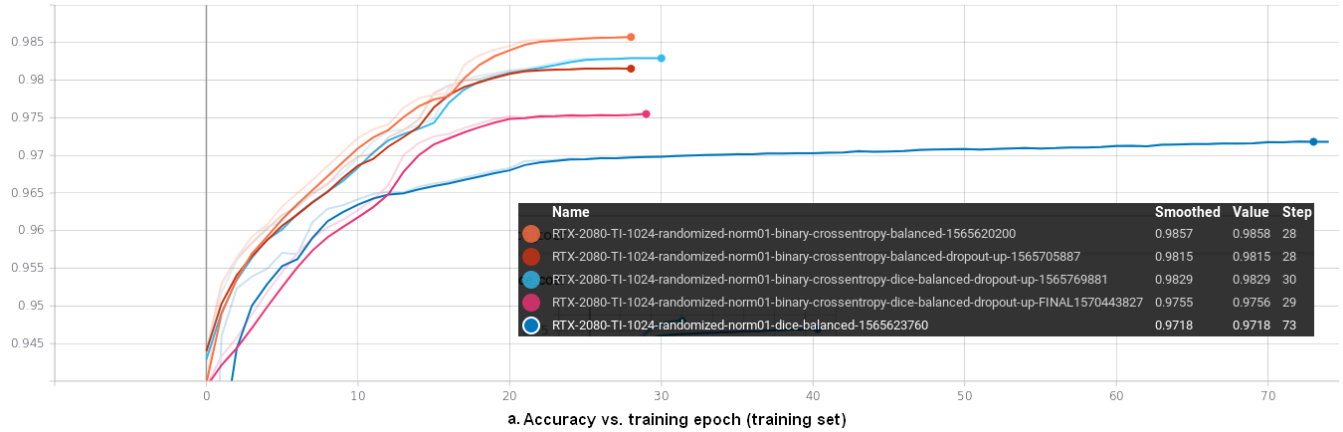


Figure 3. U-Net Architecturere for training patches of size of 256x256 obtined by downsampling 1024x1024 ROIs.

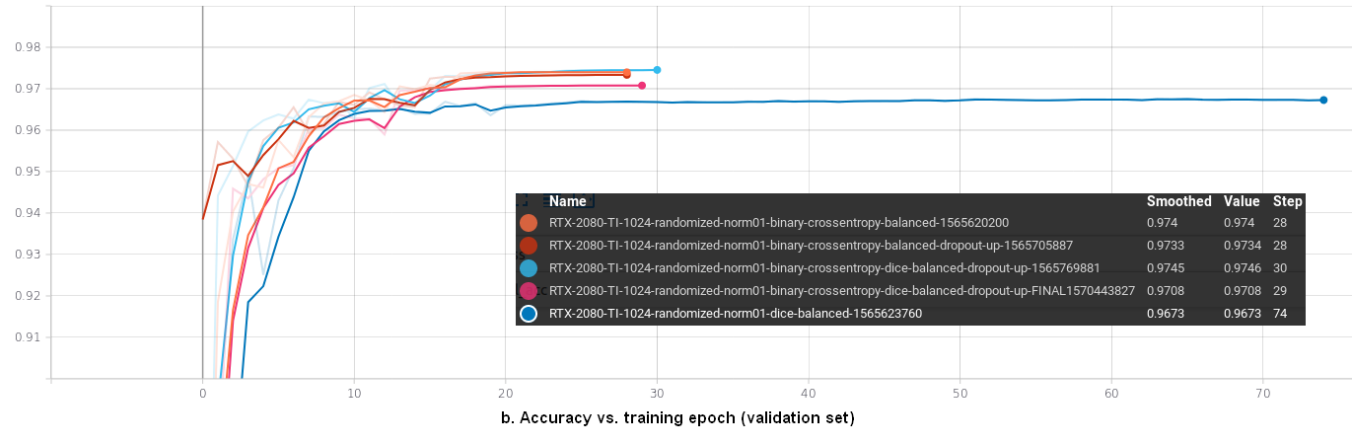
In the following the accuracy and the loss of the trained models as provided by Tensoarboard[23] vs. the number of epochs used for training is presented for both training and validation sets.

Figure 3 indicates the accuracy of the model on the training set and on validation set respectively, while figure 4 shows the loss of the model on the training set and on validation set. The different colors on the graph stand for how the model has been performing during the training and are as follows:

- **Orange**: [0..1] normalization for input images, with binary crossentropy as the loss function
- **Cherry-colored**: [0..1] normalization, binary crossentropy with a dropout value of 0.25
- **Light-blue** [0..1] normalization, binary crossentropy and dice coefficient as loss functions , with a dropout of 0.25
- **Pink**: same as above, with some minor tweaks regarding the dropout value and number of convolutions in the "decoder" layer
- **Dark-blue**: [0..1] normalization, dice coefficient as loss function



a. Accuracy vs. training epoch (training set)



b. Accuracy vs. training epoch (validation set)

Figure 4. Accuracy vs. training epochs: a. Training set; b. Validation set

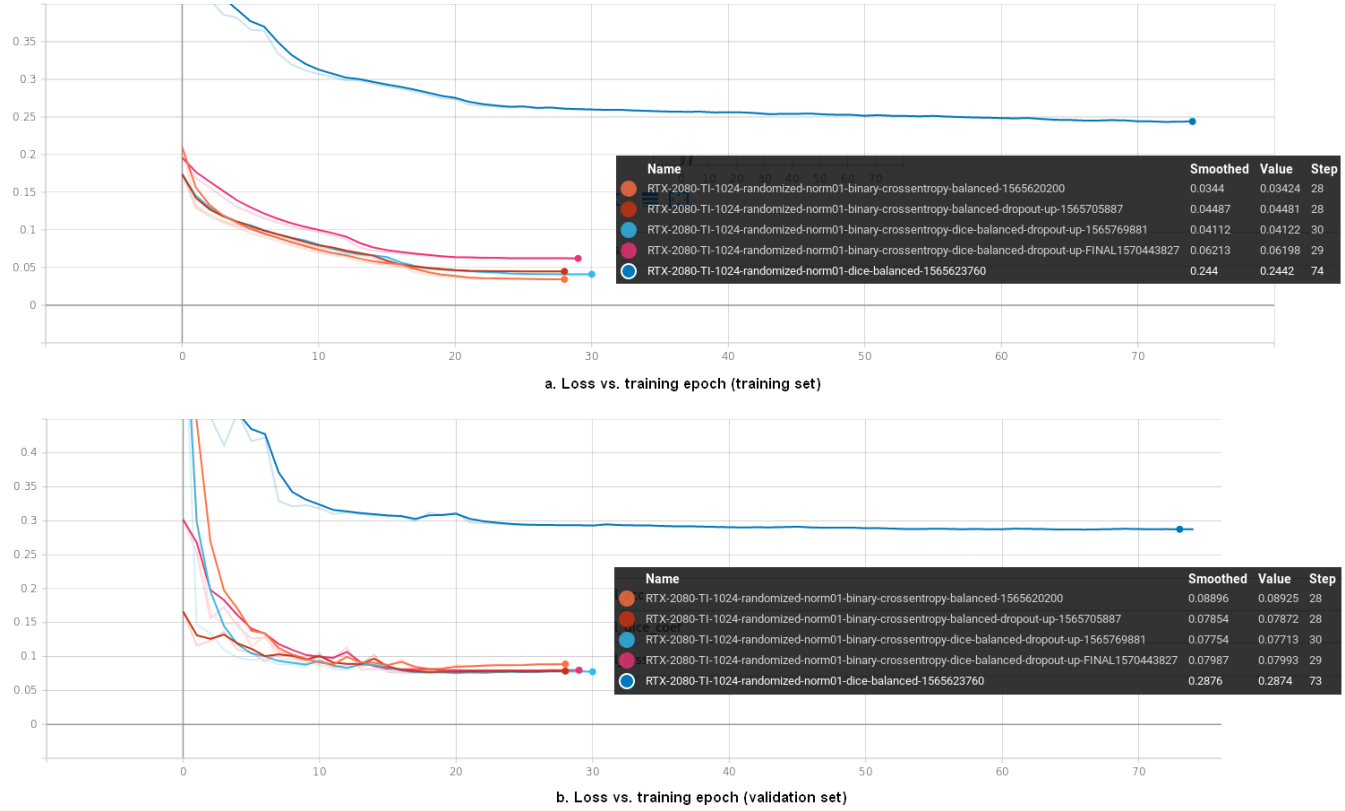


Figure 5. Loss vs. training epochs: a. Training set; b. Validation set

IV. Experimental Results and Evaluation

A. *Pixel level segmentation accuracy*

For the validation of the pixel level segmentation, the most common evaluation metric was used: Intersection over Union[28].

Intersection over Union, or the Jaccard index is a statistic used for gauging the similarity and diversity of sample sets. The Jaccard index measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cap Y|}{|X|+|Y|-|X \cap Y|} \quad (1)$$

The Sørensen–Dice coefficient can be used instead of the Jaccard index, which is defined as follows:

$$DSC(X,Y) = \frac{2|X \cap Y|}{|X|+|Y|} = \frac{|X \cap Y|}{|X-Y|+|Y-X|+2|X \cap Y|} = \frac{2TP}{2TP+FP+FN} \quad (2)$$

In the case of mammograms segmentation, the set Y can be considered to be the set of object pixels from the ground truth (GT) mask, while X can represent the pixels from the prediction mask, that have binarized with an adequate threshold. The Dice index can be computed as follows:

1. Binarize the prediction matrix;
2. Intersect the ground truth (GT) matrix with the binarized prediction matrix by applying a logical AND operator and compute the cardinality of the intersection (TP);
3. Make the union between GT and the binarized prediction matrix by applying a logical OR operator and compute the cardinality of the reunion (TP + FP + FN);
4. Compute the Dice index with the following formula:

$$DSC(X,Y) = \frac{2*TP}{TP+(TP+FP+FN)} = \frac{2*Area_{intersectiion}}{Area_{intersection}+Area_{union}} \quad (3)$$

This operation can be automated for the entire test dataset by simple or weighted addition (running average), of the two type of areas:

$$\overline{DSC}(X,Y) = \frac{2*\sum_1^N Area_{intersection}}{\sum_1^N Area_{intersection}+\sum_1^N Area_{union}} \quad (4)$$

For evaluating the segmentation algorithm on a different set of images, the test set provided in the CBIS-DDSM [19] was used, containing 348 grayscale images and 365 binary masks. The inference function was computed on patches extracted from the full-sized images by a non-overlapping sliding window scheme ($2^n x 2^n$ patches down-sampled afterwards by a ratio of 4). One important fact is that if a patch has the mean of the pixels intensities lower than an empirical value (meaning that there is mostly black in that patch), that patch will not be given to the network for the inference (see examples from figure 6.c). Furthermore, the average intensity of a left and right side strip of the grayscale image was computed to determine on which part the breast appears. Finally the prediction obtained for each patch is binarized and reassembled in the full sized segmented image.

The pseudo-code for the inference algorithm is presented below:

*I. For each **image** in **grayscale_images**:*
- *a. Compute means in left and right area of **image***
- *b. If **left_mean** >= **right_mean**:*
  - *1. Crop **image** from left to right to have a round number of patches on image width*
  - *2. Crop **image** at the top and bottom to have a round number of patches on image height*
- *c. Else:*
  - *1. Crop **image** from right to left to have a round number of patches on image width*
  - *2. Crop **image** at the top and bottom to have a round number of patches on image height*

d. ***Resize*** *each* ***cropped image*** *and normalize the values between [0..1] (for prediction)*
e. ***Nr_of_imgs_width*** *:=* ***image_width / training_patch_dim***
f. ***Nr_if_imgs_height := image_height / training_patch_dim***
g. ***x_start*** *= 0,* ***y_start*** *= 0*
h. *Initialize(y)*
i. *Initialize prediction_mask to a black image*
j. *For j in (0..****nr_of_imgs_height****):*
   1. *Initialize(x)*
   2. *For i in (0..****nr_of_imgs_width****):*
      i. *Crop a square of size* ***training_patch_dim***
      ii. *Compute* ***mean, min, max*** *of the region*
      iii. *If (****mean*** *=>* ***threshold****) and (****max*** *–* ***min*** *>0):*
         - ***pred_mask_part****=model.predict(****patch****)*
         - ***prediction_mask*** *= or (****prediction_mask, pred_mask_part****)*
      iv. *update(x)*
   3. *Update(y)*
k. *Save the* ***prediction_mask***

The obtained IoU and Dice metrics computed at pixel level on the segmented images are presented in Table 1:

TABLE I. TABLE TYPE STYLES

	IoU	**DSC**
Pixel level segmentation (entire test dataset)	0.17	0.29
Pixel level segmentation (only at the GT true positives)	0.534	0.699

The first line is the evaluation on the entire test data set and the obtained low rates are due to the high false positives rate of the pixel level segmentation. The second line of the table was computed only at the true positives level (existing abnormalities), meaning that it shows the accuracy of the prediction only where the GT indicates an abnormality.

B. *Abnormality level detection accuracy*

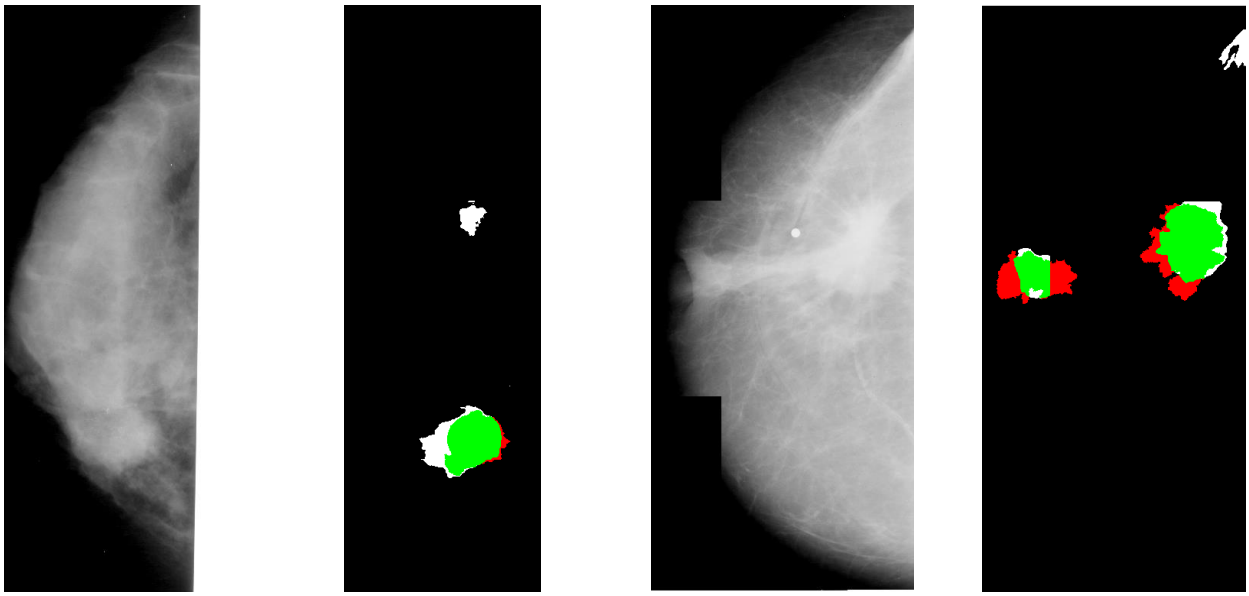
In this section the focus of the evaluation is on the entire abnormality, not on pixel level. The full image was reassembled from each prediction of 256x256 pixels made from patches of the same size from the test image.

An algorithm for automatic color coding of the true positives, false positives and false negatives at pixel level was proposed, described by the pseudo-code bellow:

I. *For each cropped* ***mask*** *and its coresponding binarized* ***prediction_map****:*
   a. *Initialize* ***final_prediction_map*** *with* ***prediction_map***
   b. ***And****(****mask****,* ***prediction_map****) => color with green in* ***final_prediction_map***
   c. ***Difference****(****mask****,* ***prediction_map****) => color with red in* ***final_prediction_map***
   d. *Save the* ***final_prediction_map***

Based on the automatic color coding the evaluation of the detection accuracy at abnormality level is performed by manually/visually counting each colored label on the entire test dataset.

An example on how the prediction performs is shown in figure 6, where the green zones are the intersection of the GT with the predicted mask (true positives), white areas are false positives and the red zones are GT with missed predictions (false negatives).



a. Grayscale image (ex.1) b. Mass segmentation (ex.1) c. Grayscale image (ex.2) d. Mass segmentation (ex. 2)

Figure 6. Segmentation results showing true positives (green regions), false positives (white regions) and false negatives (red).

In terms of masses detection rates, a FPPI (False Positive per Image) rate of 3.65 and a sensitivity (True Positive Rate – TPR) of 0.77 were obtained on the whole test data set, which are closed to the ones reported in the literature[13],[14] but still not as good as in the case of commercial mammography CAD systems[8].

V. CONCLUSIONS

In the current paper a method for automatic segmentation of potentially malignant abnormalities in 2D mammographic images based on pixel level classification using convolutional neural networks (CNN) was proposed. Experiments were performed for masses segmentation by training U-Net models on the CBIS-DDSM dataset using original algorithms for data pre- and post-processing.

A preprocessing step of the training data set was proposed that provides $2^n x 2^n$ size (n configurable) image patches (ROIs) centered on the annotated GT (Ground Truths). For data augmentation and to prevent over-fitting, an additional set of 4 images were cropped for the same binary mask, where the centers of the $2^n x 2^n$ ROIs were shifted by a random value.

For the masses segmentation a pixel level classification model based on the U-Net CNN architecture was trained on the downscaled ROIs (by a ratio of 4). Different variations of the model were tested by modifying the loss function type (binary cross-entropy, dice coefficient loss) and its metrics (accuracy, dice coefficient). To increase the trained model's performance, balancing the training set was performed by giving different weights to black and white pixels. The initial training set was split into training and validation with a ratio of 85% to 15%, obtaining an average accuracy of 0.98 / 0.97 and an average loss of 0.04 / 0.08 on the training / validation sets respectively.

For the image segmentation step, the test images were divided in patches of the same resolutions ($2^n x 2^n$ ROIs further downsampled by the same ratio of 4) using a non-overlapping sliding window scheme. The prediction map was generated by applying the inference with the obtained models for each patch and the full test images were resembled from the binarized predictions.

For evaluation of the results both pixel level segmentation and abnormalities level detection rates were estimated in a semi-automatic manner. At pixel level segmentation the average Dice index computed on the pixels of the whole test data set was low (0.29) due to the high global pixel level false positives rate. When only regions with abnormalities (regions with GT positives) were taken into consideration, the average Dice index increased to 0.7. In terms of abnormalities detection rates, a FPPI (False Positive per Image) rate of 3.65 and a sensitivity of 0.77 were obtained on the whole test data set, similar to the ones reported in the literature.

The obtained results proved that the proposed method has the potential to be integrated in mammography CAD applications for assisting screening programs in which the main goal is a high true positive detection rate and a small false negative detection rate of potentially malignant abnormalities.

Further work will be performed for experimenting the proposed method for masses detection on the CBIS-DDSM annotated dataset using segmentation models based on several other CNN architectures (AlexNet, VGGNet, GoogleNet, ResNet, ERFNet), in order to find the best segmentation model. The experiments will be also replicated for the micro-calcifications detection on the same dataset.

As a further step for integrated the proposed segmentation models in a mammographic CAD application, the models parameters have to be tuned by building an own annotated dataset based on 2D digital mammograms used in the current radiological practice.

ACKNOWLEDGMENT

This work was supported by a grant of the Romanian Ministry of Research and Innovation, CCCDI - UEFISCDI, project number PN-III-P2-2.1-CI-2018-1362, within PNCDI III.

REFERENCES

[1] Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., Murray, T., & Thun, M. J. (2008). Cancer statistics, 2008. CA: a cancer journal for clinicians, 58(2), 71-96.

[2] Ikeda, D., & Miyake, K. K. (2016). Breast Imaging: The Requisites E-Book. Elsevier Health Sciences.

[3] L.Tab´ar, et al. (2011). Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. Radiology, 260(3), 658-663.

[4] Nishikawa, R. M., & Gur, D. (2014). CADe for early detection of breast cancer - current status and why we need to continue to explore new approaches. Academic radiology, 21(10), 1320-1321.

[5] Doi, K. (2007). Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Computerized medical imaging and graphics, 31(4-5), 198-211.

[6] SenoClaire TM, http://www3.gehealthcare.com/static/senoclaire/

[7] ImageChecker CAD, https://www.hologic.com/hologic-products/breast-skeletal/image- analytics

[8] Parascript AccuDetect, https://www.parascript.com/wp-content/uploads/2015/10/AccuDetect-Brochure-2015.pdf

[9] CADOne - https://threepalmsoft.com/products/cadone/

[10] CAD: ProFound AI, SecondLook - https://www.icadmed.com/what-is-cad.html

[11] Arevalo, J., González, F. A., Ramos-Pollán, R., Oliveira, J. L., & Lopez, M. A. G. (2016). Representation learning for mammography mass lesion classification with convolutional neural networks. Computer methods and programs in biomedicine, 127, 248-257.

[12] Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., & Cardoso, J. S. (2012). Inbreast: toward a full-field digital mammographic database. Academic radiology, 19(2), 236-248.

[13] Dhungel, N., Carneiro, G., & Bradley, A. P. (2015, November). Automated mass detection in mammograms using cascaded deep learning and random forests. In 2015 international conference on digital image computing: techniques and applications (DICTA) (pp. 1-8). IEEE.

[14] Dhungel, N., Carneiro, G., & Bradley, A. P. (2017). A deep learning approach for the analysis of masses in mammograms with minimal user intervention. Medical image analysis, 37, 114-128.

[15] Heath, M., Bowyer, K., Kopans, D., Moore, R., & Kegelmeyer, W. P. (2000, June). The digital database for screening mammography. In Proceedings of the 5th international workshop on digital mammography (pp. 212-218). Medical Physics Publishing.

[16] Zhu, W., Xiang, X., Tran, T. D., & Xie, X. (2016). Adversarial deep structural networks for mammographic mass segmentation. arXiv preprint arXiv:1612.05970.

[17] H. Jung, Bet al. (2018). Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network. PloS one, 13(9), e0203355.

[18] Xi, P., Shu, C., & Goubran, R. (2018, June). Abnormality detection in mammography using deep convolutional neural networks. In 2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA) (pp. 1-6). IEEE.

[19] Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., & Rubin, D. L. (2017). A curated mammography data set for use in computer-aided detection and diagnosis research. Scientific data, 4, 170177.

[20] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2921-2929).

[21] The DICOM standard https://www.dicomstandard.org/

[22] Ronneberger, O., Fischer, P., & Brox, T. N. Convolutional networks for biomedical image segmentation. In Paper presented at: International Conference on Medical Image Computing and Computer-Assisted Intervention2015. (pp. 234-241).

[23] TensorBoard: TensorFlow's visualization toolkit, https://www.tensorflow.org/tensorboard

[24] Intersection over Union (Jacard index) https://en.wikipedia.org/wiki/Jaccard index