Benign and Malignant Breast Mass Detection and Classification in Digital Mammography: The Effect of Subtracting Temporally Consecutive Mammograms

Kosmia Loizidou[1], Galateia Skouroumouni[2], Gabriella Savvidou[3], Anastasia Constantinidou[3], Christos Nikolaou[4] and Costas Pitris[1]

***Abstract*—Breast cancer remains one of the leading cancers worldwide and is the main cause of death in women with cancer. Effective early-stage diagnosis can reduce the mortality rates of breast cancer. Currently, mammography is the most reliable screening method and has significantly decreased the mortality rates of these malignancies. However, accurate classification of breast abnormalities using mammograms is especially challenging, driving the development of Computer-Aided Diagnosis (CAD) systems. In this work, subtraction of temporally consecutive digital mammograms and machine learning were combined, to develop an algorithm for the automatic detection and classification of benign and malignant breast masses. A private dataset was collected specifically for this study. A total of 196 images were gathered, from 49 patients (two time points and two views of each breast), with precisely annotated mass locations and biopsy confirmed malignant cases. For the classification, ninety-six features were extracted and five feature selection techniques were combined. Ten classifiers were tested, using leave-one-patient-out and 7-fold cross-validation. The classification performance reached 91.7% sensitivity, 89.7% specificity and 90.8% accuracy, using Neural Networks, an improvement, compared to the state-of-the-art algorithms that utilized sequential mammograms for the classification of benign and malignant breast masses. This work demonstrates the effectiveness of combining subtraction of temporally sequential digital mammograms, along with machine learning, for the automatic classification of benign and malignant breast masses.**

***Keywords*— Breast cancer, Computer-Aided Diagnosis (CAD), digital mammography, temporal subtraction, machine learning**

I. INTRODUCTION

The World Health Organization (WHO) estimates that, by 2030, there will be 2.7 million new Breast Cancer (BC) cases and 857 thousand women will die worldwide [1]. Mammograms are currently assessed by two radiologists, and a third, if consensus is not reached. However, the identification of breast masses is very challenging due to images of dense breast tissue with increased intensity and variations that are very similar to some abnormalities [2].

A breast mass can be radiologically classified as benign or suspicious depending on key parameters such as shape, intensity, texture, etc. [3]. One of the most challenging tasks for radiologists is to accurate classify benign vs. malignant masses, thus, Computer-Aided Diagnosis (CAD) systems are being developed, to assist in that task. Several algorithms have been proposed for the classification of breast masses [4]. However, in the majority of the studies, only the most recent mammogram is being used for the diagnosis, which does not allow comparison with prior images from the same woman. Such comparisons are routinely performed by the radiologists to discover new abnormalities, or regions changing rapidly between screenings, and are considered features indicative of malignancy. Temporal analysis is a technique proposed for the comparison of consecutive mammograms and has already been applied to breast mass detection and classification [5], [6]. This technique offers no benefit when the findings are new and with no traces of an abnormality in the prior mammogram.

In this work, an algorithm for the automatic classification of benign and malignant masses is introduced, based on the subtraction of temporally consecutive digital mammograms and machine learning. Temporal subtraction, developed by this group, has already been applied with great success, to the diagnosis of breast micro-calcifications [7]. A dataset was collected specifically for this study with a total of 196 images, from 49 patients. Mammograms were at first pre-processed, and then image registration, along with temporal subtraction, took place. Mass detection and segmentation followed and, subsequently, 96 features were extracted from each mass, which were ranked using 5 feature selection algorithms. After testing several classifiers and validation schemes, the masses were classified as benign or malignant.

II. MATERIALS AND METHODS

A new, custom, dataset was collected specifically for this study, since publicly available databases do not include sequential mammograms and in some cases, the available images are scanned or/and outdated. It also includes precise

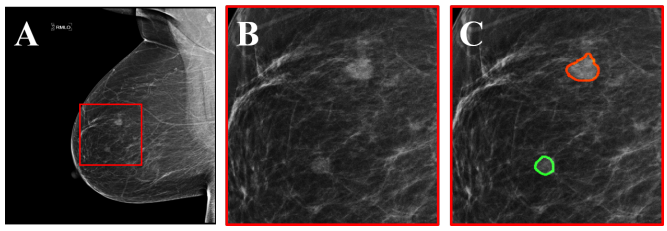

Fig. 1. Dataset example. **(A)** Mammogram of a 68-year-old woman with benign and malignant masses. **(B)** Zoomed region marked by the red square in **A**, including the masses. **(C)** The region in **B** with precise marking of mass locations (green for benign, red for malignant), as annotated by two expert radiologists.

[1]K. Loizidou and C. Pitris are with the Department of Electrical and Computer Engineering, KIOS Research and Innovation Center of Excellence, University of Cyprus, Cyprus cloizi01@ucy.ac.cy

[2]G. Skouroumouni is with the German Oncology Center, Cyprus

[3]G. Savvidou and A. Constantinidou are with the Medical School University of Cyprus and the Bank of Cyprus Oncology Center, Cyprus

[4]C. Nikolaou is with the Limassol General Hospital, Cyprus

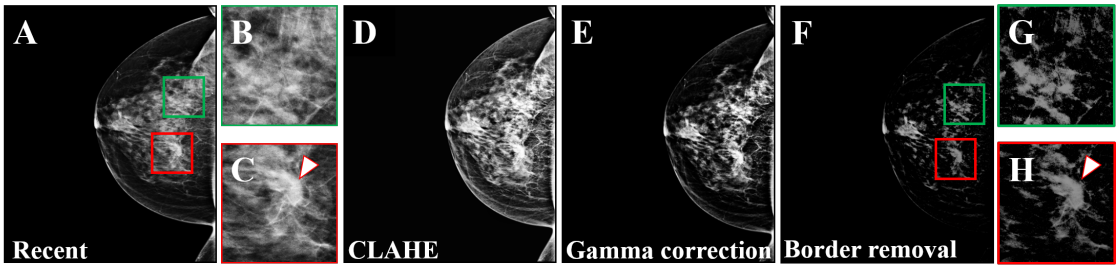

Fig. 2. Effect of pre-processing on a 69-year-old woman. **(A)** Most recent mammogram without any processing. **(B)** Zoomed region marked by the green square in **A**, showing an area without masses. **(C)** Zoomed region marked by the red square in **A**, showing an area with a mass (indicated by the arrow). **(D)** Recent image after CLAHE. **(E)** Following image after gamma correction. **(F)** Final pre-processed image after border removal. **(G)** Zoomed region marked by the green square in **F**, showing the same area as **B**, after the pre-processing. **(H)** Zoomed region marked by the red square in **F**, showing the same area as **C**, after the pre-processing.

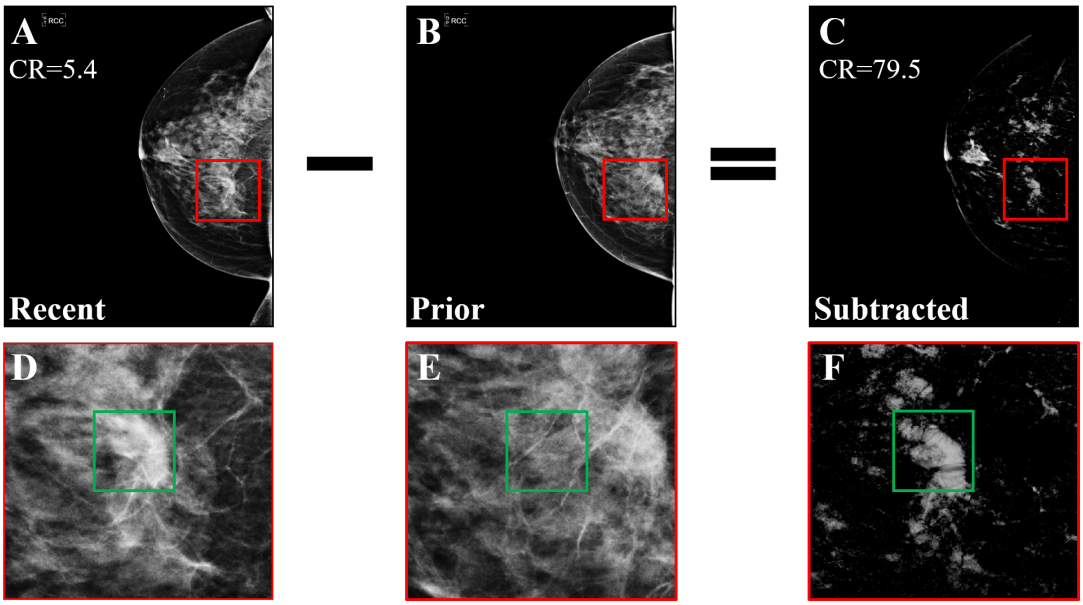

Fig. 3. Example of temporal subtraction in a 69-year-old woman with a malignant mass. **(A)** Most recent mammogram. **(B)** Prior mammogram. **(C)** The result of subtracting the registered version of **B** from **A**. **(D)-(F)** Zoomed regions marked by the red squares in **A-C**, where the green squares enclose a new malignant mass that was not subtracted. The contrast ratio (CR) has increased 14 times after subtraction.

annotation of each individual mass, which served as the ground truth (Fig. 1). The mammograms were collected from various screening centers across Cyprus and for every participant two mammographic views, the Cranio-Caudal (CC) and Medio-Lateral Oblique (MLO) were included. Two images from two sequential screening rounds, resulted in a database with a total of 196 mammograms. Two radiologists selected and assessed the images to mark the masses as benign or suspicious. The suspicious cases were then confirmed as malignant with biopsies, followed by histopathology. Thirty-four patients had at least one biopsy-confirmed malignant mass in the most recent screening. The remaining 15 patients exhibited only benign masses in the most recent mammograms. In all cases, the prior mammograms were normal. The study was approved by the Cyprus National Bioethics Committee.

The recent and prior mammograms were processed in parallel. Pre-processing started with the normalization, to adjust the range of pixel intensity values, followed by Contrast Limited Adaptive Histogram Equalization (CLAHE) [8], gamma correction [9] and border removal [10]. Figure 2, displays the effect of pre-processing.

For an effectively subtraction between the prior and the recent image, a very robust image registration algorithm is required. Registration is very challenging, since the mammograms vary significantly between screenings due to variations in breast compression and breast tissue changes [11]. In this work, Demons registration [12] was selected, since it can better account for the non-linear shape deformations of the breast. Demons is a local registration technique that aligns the moving image (prior) to the fixed (recent), using regional similarity and location [12]. Following, the prior registered image was subtracted from the recent one and the high intensity areas on the periphery of the breast were removed, since they correspond to skin regions that cannot contain masses and were a result of misalignment (Fig. 3). To evaluate the performance of pre-processing, registration and temporal subtraction, the Contrast Ratio (CR) of the subtracted image was compared to the CR of the recent image after pre-processing. Unsharp-mask filtering [13] was then applied,

TABLE I
COMPARISON OF THE CLASSIFICATION RESULTS OF THE MASSES AS BENIGN OR MALIGNANT IN A LEAVE-ONE-PATIENT-OUT CROSS-VALIDATION SCHEME

Classifier	Sensitivity [%]	Specificity [%]	Accuracy [%]	AUC
LDA	76.19	74.14	75.32	0.76
9-NN	83.33	91.38	86.62	0.87
SVM	52.38	98.28	71.13	0.75
NB	91.67	43.10	71.38	0.67
RF	84.52	81.03	83.10	0.83
ADA	84.52	75.86	80.99	0.8
BAG	82.14	82.76	82.39	0.81
GB	83.33	87.93	85.21	0.86
Voting	86.90	87.93	87.32	0.87
NN	**91.67**	**89.66**	**90.85**	**0.91**

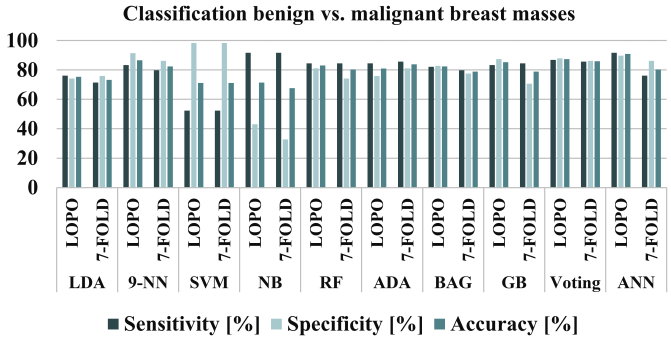

Fig. 4. Comparison of the classification results of the masses as benign or malignant using various classifiers and cross-validation methods.

to enhance the high spatial frequencies. Thresholding using Otsu's method eliminated the remaining low intensity areas. The threshold value was selected by optimizing the global classification rate. Subsequently, morphological operations were applied to identify the margins of the breast masses. For the training of the algorithms, the ground truth provided by the radiologists was used.

In total 96 features were extracted from the detected regions, divided in four major categories: shape-based, intensity-based, First-Order Statistics (FOS) and Gray Level Co-occurrence Matrix (GLCM) features. The selection of these features was based on characteristics that radiologists routinely check to assess if a mass is benign, or whether it warrants further investigation. They included: area, circularity, compactness, convex area, eccentricity, equivalent diameter, Euler number, extent, filled area, major and minor axis length, orientation, perimeter, solidity, shape ratio, average, minimum and maximum intensity, entropy, kurtosis, skewness, smoothness, standard deviation, variance, contract, correlation, energy and homogeneity. For each GLCM feature the mean and standard deviation were obtained. To determine the most appropriate offset, three different values were tested ($D_1 = 5$, $D_2 = 15$ and $D_3 = 25$). Thus, a total of 72 GLCM features were extracted.

Feature selection is critical for an effective classification. Five feature selection algorithms were compared including: feature importance, using random forest and extra trees, Maximum Relevance-Minimum Redundancy (MRMR), SelectKBest and t-test. Since each algorithm is based on different principles, they result in different rankings of the features. Thus, to select the most significant features and to assure high classification performance, the rankings were combined by applying a majority rule (i.e. keep the common features from all the methods) and a new feature vector was created for the classification. The selected features were: major and minor axis length, convex area, solidity, extent, perimeter, correlation 0° D1, correlation 45° D1, correlation 0° D2, correlation 45° D2, correlation 135° D2, correlation mean D2, correlation 0° D2, correlation 45° D3, correlation 135° D3, correlation mean D3, circularity, compactness and shape ratio. Synthetic Minority Oversampling Technique (SMOTE) was applied to create new instances of the minority class in the training set [14]. Least squares (l2) normalization was applied to the features of each mass, to scale all the samples and adjust the range of their values.

Nine classifiers were evaluated: Linear Discriminant Analysis (LDA), k-Nearest Neighbor (k-NN), Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), AdaBoost (ADA), Bagging (BAG), Gradient Boosting (GB) and Voting, using Python (v. 3.7.7) and Scikit-learn (v. 0.23.1).

Different Neural Network (NN) architectures were also evaluated using Python (v. 3.7.7) and Keras (v. 2.3.1). All the available parameters of the network were tested and optimized based on the classification accuracy. The selected architecture consisted of 1 fully connected layer, with 6,050 trainable parameters. A Rectified Linear Unit (ReLU) was used as an activation function and batch normalization, along with dropout regularization (0.2), were included. Gaussian noise was added after dropout, as a regularization term, in order to increase the robustness of the network. The batch size was set to 128, the learning rate was 0.0001 and the network was trained for 100 epochs. The features were added to the network without any pre-processing due to the limited sample size and the complexity of the network.

For the training, Leave-One-Patient-Out (LOPO) cross-validation was applied. All the images associated with a single patient were combined as a test set, while the images of the remaining patients were used as a training set, repeating until all the 49 cases were classified. In addition to LOPO cross-validation, 7-fold cross-validation was also applied, to verify the classification performance. In a similar manner, the folds were created per patient and not by randomly dividing the masses. Grouping the data per patient is of critical importance to avoid any bias resulting from the same patient included both in the training and test set. Sensitivity, specificity, accuracy and the Area Under the receiver operating characteristics Curve (AUC) were calculated to evaluate the effectiveness of the classification.

III. EXPERIMENTAL RESULTS

After registration and subtraction, the average CR of the subtracted images increased by ∼30%, compared to the corresponding average CR of the most recent images even after pre-processing. The selected features were incorporated

into the classifiers that were optimized using LOPO cross-validation. The optimization resulted in a radial basis function kernel for the SVM, 9-nearest neighbors for the k-NN, and for the ensemble voting, 9-NN, BAG and GB were combined, in a soft voting scheme. NN achieved the highest and most robust classification performance, with 91.67% sensitivity, 89.66% specificity and 90.85% accuracy (Table I). In addition, 7-fold cross-validation was used to prove the robustness of the algorithms (Fig. 4).

IV. DISCUSSION

For the classification of breast masses, NN reached 90.85% accuracy using LOPO cross-validation, with an average of 0.06 false positives and 0.07 false negatives per image. Out of 58 benign masses, 6 were wrongly detected as malignant, affecting 3 patients. Similarly, out of 84 malignant masses, 7 were misclassified as benign, again in 3 patients. In addition to LOPO cross-validation, 7-fold cross-validation was also applied, to evaluate the robustness of the algorithm. The performance dropped slightly, since 42 patients were used in each training round, compared to the 48 patients in the LOPO scheme. This drop exemplifies the need for additional training data, but also proves the potential of the algorithm to correctly classify new data.

This is the first demonstration of temporal subtraction for the classification of breast masses, thus, direct comparison with other studies is not possible. The current state-of-the-art in the analysis of sequential mammograms is temporal analysis. The results in this study are slightly better than those reported in the literature for the classification of benign vs. malignant masses using sequential mammograms (0.91 AUC vs. 0.9 Bozek et al., 2014 [5] and 0.90 Ma et al., 2015 [6]), in terms of the AUC. Additionally, temporal analysis offers no benefit over using just the recent mammogram, when the findings are new and there is no traces of an abnormality in the prior image. Temporal subtraction proposed in this study, overcomes this limitation, since it tracks and classifies newly developed abnormalities, or regions that changed significantly between the screenings. Unfortunately, direct comparison of different algorithms is challenging due to differences in the method of cross-validation applied [5], [15].

A key limitation of this work, is the relatively small dataset. Publicly available databases cannot be exploited, since they do not contain sequential digital mammograms, nor they include detail annotation of each individual mass. Other limitations include the fact that the patients with benign masses were not followed for further diagnostic evaluation and, although the masses were identified by two expert radiologists, differences might appear if more experts perform the same task.

V. CONCLUSION

In this work, an algorithm for the automatic classification of benign and malignant breast masses based on subtracting temporally consecutive mammograms and machine learning was proposed. Ninety-six features were extracted and using five feature selection techniques, the most statistically significant features were included in the classification. The highest classification performance was 90.85% accuracy and it was achieved using a NN and leave-one-patient-out cross-validation. Compared to the state-of-the-art techniques that use sequential mammograms and temporal analysis, the results in this study were superior (0.90 vs. 0.91 AUC). However, given the relatively small dataset, further studies must be conducted with more cases and different cross-validation methods. The findings of this study, if expanded and improved have the potential to encourage the development of automated CAD systems, with a major impact on patient prognosis.

ACKNOWLEDGMENT

This research was funded by the European Union's Horizon 2020 research and innovation program under grant agreement No. 739551 (KIOS CoE) and from the Republic of Cyprus through the Directorate General for European Programs, Coordination and Development.

REFERENCES

[1] J. Ferlay *et al.*, "Global cancer observatory: cancer today. lyon, france: international agency for research on cancer." https://gco.iarc.fr/, 2018.

[2] D. A. Spak *et al.*, "Bi-rads® fifth edition: A summary of changes," *Diagnostic and interventional imaging*, vol. 98, no. 3, pp. 179–190, 2017.

[3] A. Oliver *et al.*, "A review of automatic mass detection and segmentation in mammographic images," *Medical image analysis*, vol. 14, no. 2, pp. 87–110, 2010.

[4] S. Z. Ramadan, "Methods used in computer-aided diagnosis for breast cancer detection using mammograms: a review," *Journal of healthcare engineering*, vol. 2020, 2020.

[5] J. Bozek *et al.*, "Use of volumetric features for temporal comparison of mass lesions in full field digital mammograms," *Medical physics*, vol. 41, no. 2, 2014.

[6] F. Ma *et al.*, "Computer aided mass detection in mammography with temporal change analysis," *Computer Science and Information Systems*, vol. 12, no. 4, pp. 1255–1272, 2015.

[7] K. Loizidou *et al.*, "Digital subtraction of temporally sequential mammograms for improved detection and classification of microcalcifications," *European radiology experimental*, vol. 5, no. 1, pp. 1–12, 2021.

[8] S. Agrawal *et al.*, "Detection of breast cancer from mammograms using a hybrid approach of deep learning and linear classification," in *2018 International Conference on Smart City and Emerging Technology (ICSCET)*, pp. 1–6, IEEE, 2018.

[9] S.-C. Huang, F.-C. Cheng, and Y.-S. Chiu, "Efficient contrast enhancement using adaptive gamma correction with weighting distribution," *IEEE Transactions on Image Processing*, vol. 22, no. 3, pp. 1032–1041, 2013.

[10] R. Gonzalez, R. Woods, and S. Eddins, *Digital image processing using MATLAB*. Gatesmark Publishing, 2nd ed., 2010.

[11] K. Marias *et al.*, "A registration framework for the comparison of mammogram sequences," *IEEE Transactions on Medical Imaging*, vol. 24, no. 6, pp. 782–790, 2005.

[12] X. Pennec, P. Cachier, and N. Ayache, "Understanding the "demon's algorithm": 3d non-rigid registration by gradient descent," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 597–605, Springer, 1999.

[13] H. P. Chan *et al.*, "Digital mammography. roc studies of the effects of pixel size and unsharp-mask filtering on the detection of subtle microcalcifications.," *Investigative radiology*, vol. 22, no. 7, pp. 581–589, 1987.

[14] N. V. Chawla *et al.*, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[15] R. Rouhi *et al.*, "Benign and malignant breast tumors classification based on region growing and cnn segmentation," *Expert Systems with Applications*, vol. 42, no. 3, pp. 990–1002, 2015.