Identification and Classification of Benign and Malignant Masses based on Subtraction of Temporally Sequential Digital Mammograms

Kosmia Loizidou¹, Galateia Skouroumouni², Gabriella Savvidou³, Anastasia Constantinioudou³, Christos Nikolaou⁴ and Costas Pitriss

Abstract—Breast cancer remains the leading cause of cancer deaths and the second highest cause of death, in general, among women worldwide. Fortunately, over the last few decades, with the introduction of mammography, the mortality rate of breast cancer has significantly decreased. However, accurate classification of breast masses in mammograms is especially challenging. Various Computer-Aided Diagnosis (CAD) systems are being developed to assist radiologists with the accurate classification of breast abnormalities. In this study, classification of benign and malignant masses, based on the subtraction of temporally sequential digital mammograms and machine learning, is proposed. The performance of the algorithm was evaluated on a new dataset created for the purposes of this study. In total, 196 images from 49 patients, with precisely annotated mass locations and biopsy confirmed malignant cases, were included. Ninety-six features were extracted and five feature selection algorithms were employed to identify the most important features. Ten classifiers were tested using leave-one-patient-out and 7-fold cross-validation. Neural Networks, achieved the highest classification performance with 90.85% accuracy and 0.91 AUC, an improvement compared to the state-of-the-art. These results demonstrate the effectiveness of the subtraction of temporally consecutive mammograms for the classification of breast masses as benign or malignant.

Keywords—Breast cancer, Computer-Aided Diagnosis (CAD), Digital mammography, Temporal subtraction, Machine learning

I. INTRODUCTION

The World Health Organization (WHO) estimates that, by 2025, new Breast Cancer (BC) cases will reach 2.5 million and at least 769 thousand women will die worldwide [1]. Currently, mammograms are evaluated by two radiologists, and a third if consensus is not reached. However, images of dense breast tissue exhibit increased intensity, with variations that are very similar to some abnormalities, making the identification of breast masses very challenging [2].

A breast mass can be radiologically classified as benign or suspicious depending on key parameters such as shape, intensity, texture, etc. [3]. Accurate classification of benign vs. malignant masses is one of the most challenging tasks for radiologists, thus, Computer-Aided Diagnosis (CAD) systems are being developed to assist in that task. Various algorithms have been developed for the classification of breast masses [4]. However, most use only the most recent mammogram for the diagnosis, which does not allow comparison with available prior images of the same patient. Such comparisons are routinely performed by the radiologists to identify newly developed abnormalities or regions changing rapidly between screenings and are considered critical in assuring correct diagnosis. Temporal analysis is a technique developed for the comparison of sequential mammograms and has already been applied to breast mass detection and classification [5], [6]. Although the results were promising, this approach offers no benefit when the findings are new and there are no traces of an abnormality in the prior image.

In this study, a new algorithm for the classification of benign and malignant masses is proposed, based on the subtraction of temporally sequential digital mammograms. Temporal subtraction, developed by this group, has already been applied, with great success, to the detection and classification of breast micro-califications [7]. The algorithm was evaluated on a new dataset, created specifically for this study. First, the images were pre-processed and image registration along with temporal subtraction were applied. Mass detection and segmentation followed. Subsequently, 96 features were extracted from each mass, which were ranked using 5 feature selection algorithms. After applying various classifiers and validation schemes, the masses were classified as benign or malignant and the most effective approach was identified.

II. MATERIALS AND METHODS

A. Dataset

This study required a new, custom, dataset since publicly available datasets do not include sequential mammograms and in some cases the mammograms are scanned and outdated. In addition, this dataset included precise annotation of each individual mass, which served as the ground truth (Fig. 1). The dataset included women 39 to 80 years of age.

Fig. 1. Dataset example. (A) Mammographic view of a 58-year-old woman with benign and malignant masses. (B) Zoomed regions marked by the red square in A, including the masses. (C) The regions in B with precise marking of mass locations (green for benign, red for malignant), as annotated by two expert radiologists.

1K. Loizidou and C. Pitriss are with the Department of Electrical and Computer Engineering, KIOS Research and Innovation Center of Excellence, University of Cyprus, Cyprus cloizidou@ucy.ac.cy
2G. Skouroumouni is with the German Oncology Center, Cyprus
3G. Savvidou and A. Constantinioudou are with the Medical School University of Cyprus and the Bank of Cyprus Oncology Centre, Cyprus
4C. Nikolaou is with the Limassol General Hospital, Cyprus

978-1-7281-2782-8/22/$31.00 ©2022 IEEE
Fig. 2. Example of temporal subtraction in a 48-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 CR has increased 5 times after subtraction.

and was collected from various screening centers across Cyprus. For every participant, two mammographic views, the Cranio-Caudal (CC, view from above) and Medio-Lateral Oblique (MLO, angled view) 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 confirmed as malignant with biopsies followed by histopathology. Fifteen cases exhibited only benign masses in the most recent mammograms. The remaining 34 patients had at least one biopsy confirmed malignant mass in the most recent screening. In all cases the prior mammograms were normal. The study was approved by the Cyprus National Bioethics Committee.

B. Mass Detection and Segmentation

The recent and prior mammograms were pre-processed, beginning with normalization to adjust the range of pixel intensity values. Contrast Limited Adaptive Histogram Equalization (CLAHE), gamma correction and border removal were also applied. CLAHE enhanced the contrast of the images by re-allocating its gray levels, operating on small regions of [8,8] tiles [8]. Contrast adjustment, using gamma correction, accounted for the non-linear mapping of image intensities [9]. Finally, border removal removed high intensity areas connected to the border, such as the pectoral muscle [10].

A very robust image registration algorithm is required in order to effectively subtract the prior from the recent image. Registration is very challenging since the mammograms vary significantly between screenings due to breast tissue changes, variations in breast compression and operating factors at the time of imaging [11]. In this study, 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 registration, the prior registered image was subtracted from the recent one. 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. Figure 2 shows an example of temporal subtraction. 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. For improved mass segmentation, unsharp-mask filtering was applied to enhance the high spatial frequencies [13]. Thresholding using Otsu’s method eliminated the low intensity areas. The threshold value was selected using the discriminant criterion and by optimizing the global classification rate. Finally, the margins of the breast masses were identified after applying morphological operations (erosion and closing). For the training of the algorithms, the ground truth provided by the radiologists was used.

C. Feature Extraction and Selection

Various features were extracted from the regions identified above. The features were selected based on characteristics that radiologists routinely check to assess if a mass is benign, or whether it warrants further investigation. In total, 96 features were extracted, divided in four major categories: shape-based, intensity-based, First-Order Statistics (FOS) and Gray Level Co-occurrence Matrix (GLCM) features. 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 (STD) 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 very important for effective classification. The methods that were compared included: feature importance using random forest and extra trees; Maximum Relevance-Minimum Redundancy (MRMR), SelectKBest and t-test. Since each feature selection algorithm is based on different principles, they result in different rankings of the features. Thus, to select the most statistically significant features and to assure high classification accuracy, 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 best features selected 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. As is often the case in real-life scenarios, the dataset was imbalanced, with unequal numbers of benign and malignant masses (58 vs. 84). 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.

D. Classification and Performance Evaluation

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).

In addition, 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, 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 information from the same patient included in both 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 result was visually enhanced images containing only the newly developed abnormalities or the regions that have changed significantly between screenings. Feature selection revealed the most significant features for the robust classification of breast masses as benign vs. malignant. The selected features were then 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 90.85% accuracy and 0.91 AUC (Table I). In addition, 7-fold cross-validation was used to prove the robustness of temporal subtraction (Fig. 3).
TABLE II
AN OVERVIEW OF THE STATE-OF-THE-ART TECHNIQUES USING SEQUENTIAL MAMMOGRAMS FOR THE CLASSIFICATION OF BREAST MASSES AS BENIGN OR MALIGNANT

| Method                | Dataset  | Classifier | AUC |
|-----------------------|----------|------------|-----|
| Timp et al. (2007)    | 465 pairs| SVM        | 0.77|
| Bozek et al. (2014)   | 60 pairs | LDA        | 0.9 |
| Ma et al. (2015)      | 95 pairs | LDA        | 0.9 |
| Kooi and Karssemeijer (2017) | 18566 cases | CNN | 0.88|
| Proposed methodology  | 49 pairs | NN         | 0.91|

IV. DISCUSSION

For the characterization of the breast masses as benign or malignant, 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 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.

Since this is the first demonstration of temporal subtraction, 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 (Table II), in terms of the AUC. However, unlike temporal analysis, temporal subtraction, proposed in this study, tracks and classifies newly developed abnormalities or regions that changed significantly between the screenings. Direct comparison of different algorithms is challenging due to differences in the method of cross-validation and performance evaluation [15].

A key limitation of this study is the relatively small dataset. Unfortunately, publicly available databases cannot be exploited since they do not contain sequential digital mammograms, nor they include detailed 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. CONCLUSIONS

In this study, a new algorithm was introduced for the segmentation and classification of benign and malignant breast masses using temporal subtraction of sequential digital mammograms and machine learning. Various feature were extracted and ranked with a combination of different feature selection techniques. The most statistically important features were then used for the classification. The highest classification performance was achieved using a NN with 90.85% accuracy and 0.91 AUC. These results are better than the state-of-the-art techniques that use sequential mammograms and temporal analysis (0.90 vs. 0.91 AUC). Encouraged by this initial results, further studies are planned to include more patients. With further expansion and improvement, the proposed algorithm has the potential to substantially contribute to the development of automated CAD systems with significant 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., “Estimating the global cancer incidence and mortality in 2018: Globocan sources and methods,” International journal of cancer, vol. 144, no. 8, pp. 1941–1953, 2019.
[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] H. M. Rangayyan, F. J. Ayres, and J. L. Desautels, “A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs,” Journal of the Franklin Institute, vol. 344, no. 3-4, pp. 312–348, 2007.
[4] S. Zahoot et al., “Breast cancer detection and classification using traditional computer vision techniques: A comprehensive review,” Current medical imaging, 2020.
[5] S. Timp, C. Varela, and N. Karssemeijer, “Temporal change analysis for characterization of mass lesions in mammography,” IEEE Transactions on Medical imaging, vol. 26, no. 7, pp. 945–953, 2007.
[6] T. Kooi and N. Karssemeijer, “Classifying symmetrical differences and temporal change for the detection of malignant masses in mammography using deep neural networks,” Journal of Medical Imaging, vol. 4, no. 4, p. 044501, 2017.
[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 majority over-sampling technique,” Journal of artificial intelligence research, vol. 16, pp. 321–357, 2002.
[15] 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.
[16] 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.