IEEE Journal of Translational Engineering in Health and Medicine

Received 19 July 2022; revised 10 October 2022; accepted 29 October 2022.
Date of publication 4 November 2022; date of current version 10 November 2022.

Digital Object Identifier 10.1109/JTEHM.2022.3219891

Automatic Breast Mass Segmentation and Classification Using Subtraction of Temporally Sequential Digital Mammograms

KOSMIA LOIZIDOU[1], (Student Member, IEEE), GALATEIA SKOUROUMOUNI[2], CHRISTOS NIKOLAOU[3], AND COSTAS PITRIS[1], (Member, IEEE)

[1]KIOS Research and Innovation Center of Excellence, Department of Electrical and Computer Engineering, University of Cyprus, 2109 Nicosia, Cyprus
[2]Radiology Department, German Oncology Center, 4108 Limassol, Cyprus
[3]Radiology Department, Limassol General Hospital, 4131 Limassol, Cyprus

CORRESPONDING AUTHOR: K. LOIZIDOU (cloizi01@ucy.ac.cy)

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.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Cyprus National Bioethics Committee.

ABSTRACT Objective: Cancer remains a major cause of morbidity and mortality globally, with 1 in 5 of all new cancers arising in the breast. The introduction of mammography for the radiological diagnosis of breast abnormalities, significantly decreased their mortality rates. Accurate detection and classification of breast masses in mammograms is especially challenging for various reasons, including low contrast and the normal variations of breast tissue density. Various Computer-Aided Diagnosis (CAD) systems are being developed to assist radiologists with the accurate classification of breast abnormalities. Methods: In this study, subtraction of temporally sequential digital mammograms and machine learning are proposed for the automatic segmentation and classification of masses. The performance of the algorithm was evaluated on a dataset created especially for the purposes of this study, with 320 images from 80 patients (two time points and two views of each breast) with precisely annotated mass locations by two radiologists. Results: Ninety-six features were extracted and ten classifiers were tested in a leave-one-patient-out and k-fold cross-validation process. Using Neural Networks, the detection of masses was 99.9% accurate. The classification accuracy of the masses as benign or suspicious increased from 92.6%, using the state-of-the-art temporal analysis, to 98%, using the proposed methodology. The improvement was statistically significant ($p\text{-value} < 0.05$). Conclusion: These results demonstrate the effectiveness of the subtraction of temporally consecutive mammograms for the diagnosis of breast masses. Clinical and Translational Impact Statement: The proposed algorithm has the potential to substantially contribute to the development of automated breast cancer Computer-Aided Diagnosis systems with significant impact on patient prognosis.

INDEX TERMS Breast cancer, Computer-Aided Diagnosis (CAD), machine learning, sequential mammograms, temporal subtraction.

I. INTRODUCTION

Breast Cancer (BC) accounts for 19% of new cancer cases (i.e. ∼1 in 5 of all new cancers) worldwide and 30% of cancers in women, constituting a major cause of morbidity and mortality. BC incidence rates continue to increase by about 0.5% per year. A large percentage of BCs begins in the ducts, but the processes by which the malignancy initially appears and later develops, can vary between patients [1]. Mammographic screening followed by appropriate disease management in the case of positive findings, has significantly improved patient prognosis [2]. Currently, the mammograms are evaluated by two radiologists (and a third if consensus is not reached), which is an indication of the challenges faced when attempting to identify probable abnormalities in the images. Increasing breast density, described using the Breast Imaging Reporting and Data System (BI-RADS) density categorization, makes the identification of breast masses even more difficult. Images of breast with dense tissue (BI-RADS c & d), exhibit increased intensity with variations that are very similar and obscure the abnormalities. The issue is further complicated by the fact that dense tissue may also be associated with an increased risk of BC [3].

Breast masses can be radiologically classified as benign or suspicious depending on key parameters such as size, perimeter, density, gradient, texture, etc. [4]. A benign mass has a smooth, round and well-defined boundary, compared to suspicious masses which are characterized by rough, blurry and spiculated boundaries [5]. Suspicious abnormalities are followed to confirm whether are malignant or benign. Classification of masses is one of the most challenging tasks for radiologists, not only because of their wide variation in size and shape, but also because of their low contrast. Furthermore, masses are usually surrounded and/or enclosed by other structures, such as normal tissue, blood vessels and muscle [6]. Computer-Aided Diagnosis (CAD) systems are being developed to aid the radiologist with this task.

The development of various algorithms for the detection and classification of breast masses in mammograms has been the subject of an intense research effort [2], [7], [8]. Since 2015, various deep learning techniques, such as Convolutional Neural Networks (CNNs) were also introduced, increasing the accuracy of detection and automating feature extraction [9], [10], [11]. However, using single mammograms does not allow comparison of the recent and prior images of the same patient. Such comparisons are routinely performed by radiologists in order to more effectively identify any abnormalities which have developed between screenings. Usually, new abnormalities, or Regions of Interest (ROIs), changing rapidly between screenings, are more likely to be suspicious. On the contrary, ROIs that remain unchanged, are more likely benign and harmless [12].

Temporal analysis was developed for the comparison of sequential mammograms and is already being applied for breast mass detection and classification. Usually, the detection of the masses is conducted on the most recent mammographic view and, with the use of registration algorithms, the corresponding location is identified in the prior image as well. Features are extracted from recent and prior images and, then, subtracted for the creation of a temporal feature vector that is used in the classification. The effectiveness of temporal analysis for the diagnosis of breast masses has been assessed thoroughly [13], [14], [15], [16], [17], [18]. Despite the fact that the findings are promising, temporal analysis offers no benefit, compared to using only the most recent mammographic view, when the findings are new with no traces of abnormality in the prior screening.

In this work, an algorithm for the segmentation and classification of breast masses is proposed, exploiting the subtraction of temporally sequential digital mammograms and machine learning. Temporal subtraction, developed by this group, has already been applied for the detection and classification of breast micro-calcifications with great success [19]. In this study, the concept was evaluated for the segmentation and classification of masses. The various steps of the method were modified and optimized based on the radiological characteristics of masses. A new dataset was created for the purposes of this study, which included 80 patients, of which 40 had at least one suspicious breast mass only in the recent mammogram but not in the prior. From the remaining 40, 12 had only benign masses and 28 did not have any masses at all in the recent mammograms. In total, 320 images were collected (two time points, i.e. recent and prior mammogram, and two views of the breast). With the introduction of temporal subtraction, the areas that remained unchanged between the screenings and the background were effectively removed, producing a new image with higher Contrast Ratio (CR) compared to the corresponding recent image. Subsequently, all the detected ROIs were either classified as normal tissue or true masses to eliminate False Positive (FP) detections. The true masses were further classified as benign or suspicious. For comparison, temporal analysis was also performed.

TABLE 1. Characteristics of the population selected for the study.

Variable	Normal *(n = 40)*	Suspicious *(n = 40)*	Total Population *(n = 80)*
Patient age			
Mean ± STD	54.9 ± 9.9	61.5 ± 8	58.2 ± 9.6
Median	53	62	57
Range	40-81	39-80	39-81
Interquartile range	48-58.7	57-67.5	50-64.7
BI-RADS breast density			
a	5	3	8
b	17	21	38
c	14	15	29
d	4	1	5
BI-RADS classification			
1	28	0	28
2	12	0	12
3	0	0	0
4a	0	5	5
4b	0	15	15
4c	0	17	17
5	0	3	3

The rest of the paper is organized as follows: Section II describes the dataset (II-A) and the segmentation and classification of masses using temporal subtraction (II-B) and temporal analysis (II-C). Section III describes the results with Sections III-A and III-B focusing on temporal subtraction and temporal analysis, respectively. Section IV includes a discussion of the findings and the main conclusions are provided in Section V.

II. MATERIALS AND METHODS

A. DATASET

For this study, 80 full-field digital pair of mammograms were collected, between 2012 to 2020, from women 39 to 81 years of age, randomly selected from various screening centers in Cyprus. The prior mammograms originated between 2012 and 2018, with an average interval of 2.5 years

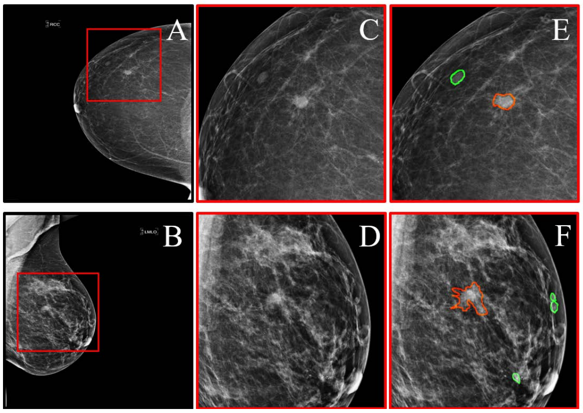

FIGURE 1. Dataset examples. (A) Mammographic view of a 68-year-old woman (BI-RADS breast density category a) with benign and suspicious masses. (B) Mammographic view of a 58-year-old woman (BI-RADS breast density category c) with benign and suspicious masses. (C, D) Zoomed regions marked by the red squares in A and B, showing masses. (E, F) The regions in C and D with precise marking of mass locations (green for benign, red for suspicious), as annotated by two expert radiologists.

between screenings. Normal cases were selected to form a matched group compared to those with suspicious findings. The study was approved by the Cyprus National Bioethics Committee.

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 dataset with a total of 320 images. A radiologist with ten years of experience selected the participants and assessed the mammograms, per BI-RADS classification, along with a second radiologist, with two years of experience. Half of the images came from participants without any suspicious findings. Of these, 12 women had only benign masses and 28 had no masses in the recent mammograms. The remaining 40 patients exhibited at least one suspicious mass in the most recent screening with a normal prior. Table 1, shows a summary of the study population. The dimensions of the mammograms were 4096 × 3328 pixels, in an 8-bit DICOM format.

It was necessary to compile a new dataset for this study since publicly available datasets do not include temporally sequential mammograms and in some cases the mammograms are scanned and/or outdated. In addition, this new dataset not only includes sequential digital mammograms, but also precise annotation of each individual mass (both benign and suspicious) that serves as the ground truth (Fig. 1). Such detailed annotations are rarely found in publicly available datasets. The dataset is publicly available (https://doi.org/10.5281/zenodo.7179856) [20].

B. BREAST MASS SEGMENTATION AND CLASSIFICATION USING TEMPORAL SUBTRACTION

The proposed methodology for the segmentation and classification of breast masses, using subtraction of sequential digital mammograms, is outlined in Figure 2.

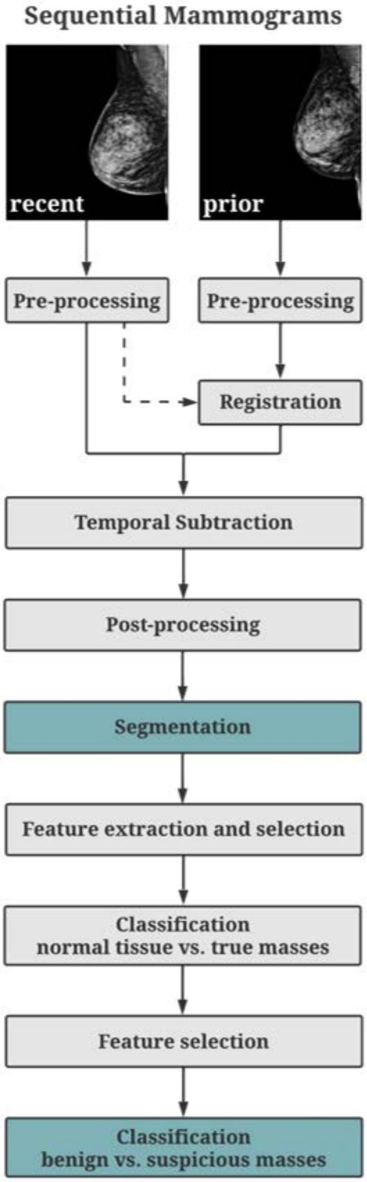

FIGURE 2. Proposed methodology for the automatic breast mass segmentation and classification using subtraction of temporally sequential digital mammograms.

1) IMAGE REGISTRATION, SUBTRACTION AND MASS SEGMENTATION

The recent and prior mammographic views were preprocessed in parallel, beginning with normalization to adjust the range of pixel intensity values. This step was followed by Contrast Limited Adaptive Histogram Equalization (CLAHE), gamma correction and border removal. CLAHE enhances the contrast of an image by re-allocating its gray levels, and operates on small regions, i.e. tiles, rather than the entire image. An important parameter in CLAHE is the clip limit, a contrast factor that prevents over-saturation of the image specifically in homogeneous areas. In this case, this parameter was set to 0.01. The application of CLAHE effectively diminished noise and edge-shadowing effects [21].

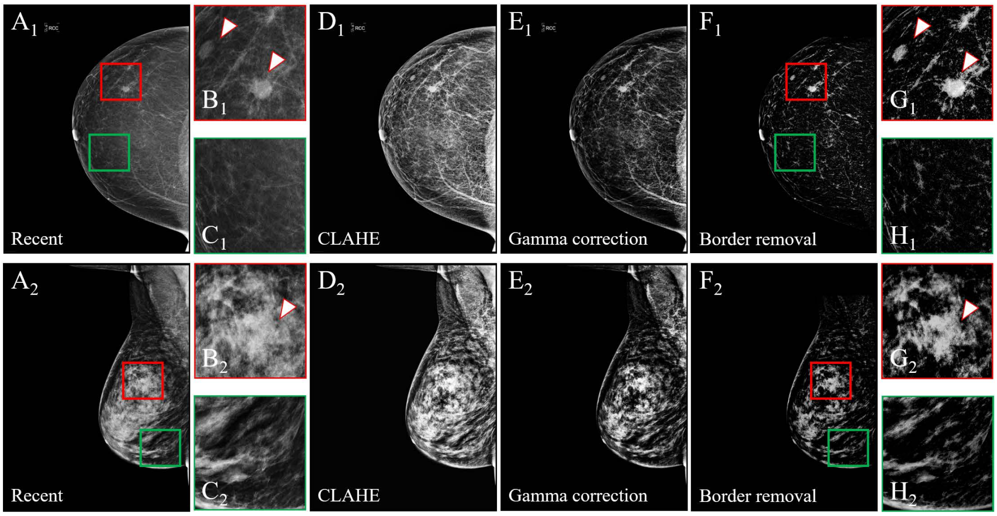

FIGURE 3. Effects of the pre-processing on two cases. (1) Mammographic view of a 68-year-old woman with BI-RADS breast density category a (Case 1, top row). (2) Mammographic view of a 50-year-old woman with BI-RADS breast density category c (Case 2, bottom row). For each case: (A) Original most recent image. (B) Zoomed region marked by the red square in A, showing an area with breast masses (indicated by the arrows). (C) Zoomed region marked by the green square in A showing an area without masses. (D) Image after CLAHE. (E) Image after gamma correction. (F) Final pre-processed image after border removal. (G) Zoomed region marked by the red square in F, showing the same area as B, after pre-processing. (H) Zoomed region marked by the green square in F, showing the same area as C, after pre-processing.

Subsequently, contrast adjustment using gamma correction was used to account for the non-linear mapping of image intensities, using:

$$I'(x,y) = l_{max}\left(\frac{I(x,y)}{l_{max}}\right)^{\gamma} \tag{1}$$

where l_{max} defines the maximum intensity of the input image, $I(x, y)$ is the intensity of each pixel in the input image and γ is the gamma parameter, which was set to 2 [22]. Border removal, the last step in the pre-processing, removed the high intensity areas connected to the border, such as the pectoral muscle, using the marker image B:

$$B(x,y) = \begin{cases} I'(x,y), & \text{if } (x,y) \text{ is on the border of } I'. \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

where, I' is the input image. Using B as a mask, a new image H that contained only the objects touching the border was reconstructed as;

$$H(x,y) = I'(x,y)B(x,y) \tag{3}$$

Subsequently, a new image C containing only the objects from the input image that do not touch the border was created as [23];

$$C(x,y) = I'(x,y) - H(x,y) \tag{4}$$

Figure 3, illustrates the effects of pre-processing for two cases, with different BI-RADS density categorization, in ROIs with and without breast masses.

For an effective subtraction between the recent and prior images, effective image registration is required. The mammograms vary significantly between screenings due to breast tissue changes, variations in breast compression and operating factors at the time of imaging [24]. Several image registration techniques have been applied in the past [25], with Affine being the most common [26]. In this study, Demons registration [27] was selected, since it can better account for the non-linear shape deformation of the breast.

Demons is a local registration technique that aligns the moving image (prior) to the fixed (recent) using regional similarity and location. In Demons, the registration is seen as a diffusion process affected by the optical flow formulation and most of the times includes a regularization term to assure continuity and smoothness [27]. It can be represented as an energy function, with respect to the update field u of a fixed image F, a moving image M and a transformation field s:

$$E^{s}_{corr}(u) = ||F - M \circ (s+u)||^2 + \left(\frac{\sigma_i^2}{\sigma_x^2}\right)||u||^2 \tag{5}$$

where, σ_i^2 is the noise of the image intensity and σ_x^2 the spatial uncertainty. With the application of a Taylor expansion, Eq. (5) is linearized and the energy function reaches its minimum when its gradient descent is zero. The registration must be solved iteratively since the update field is based on local information [28]. High intensity areas on the periphery of the breast were removed since they correspond to skin areas that cannot contain masses and were a results of misalignment. Figure 4 shows an example of temporal subtraction in a 50-year-old woman. To assess the effectiveness of the registration and temporal subtraction, the CR of the subtracted image was compared to the corresponding CR of the most recent image after pre-processing.

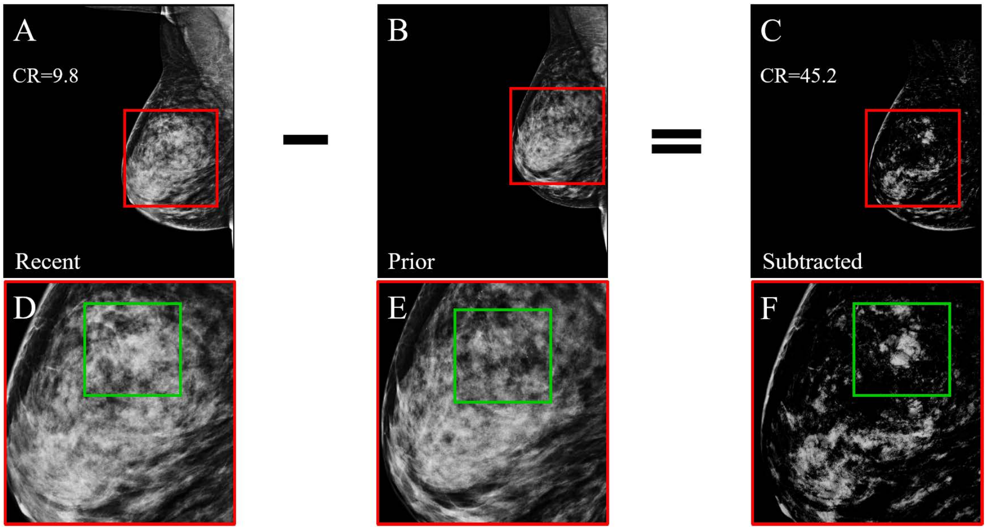

FIGURE 4. Example of temporal subtraction in a 50-year-old woman (BI-RADS breast density category c) with a suspicious mass. (A) Most recent mammographic view. (B) Prior mammographic view. (C) The result of subtracting the registered version of B from A. (D)–(F) Zoomed regions marked by the red squares in A–C, where the green squares enclose a new suspicious mass that was not subtracted. The Contrast Ratio (CR) has increased 9 times after subtraction.

The accuracy of the mass segmentation improved with unsharp-mask filtering. This spatial filter enhances a range of high frequencies using:

$$P_{sharp}(x, y) = C(x, y) + k * Q(x, y) \tag{6}$$

where, k is the scaling constant and its reasonable values k vary between 0.2 to 0.7. For this case, k was set to 0.5 after trial and error. $Q(x, y)$ calculated as follows:

$$Q(x, y) = C(x, y) - C_{smooth}(x, y) \tag{7}$$

where, $C(x, y)$ is the input image and $C_{smooth}(x, y)$ is the smoothed version of the input [29].

Subsequently, thresholding, using Otsu's method, was applied, to covert the grayscale image to binary and to eliminate the low intensity areas. Masses were preserved since they are brighter than the background. The appropriate threshold value was found using the discriminant criterion [30] and by optimizing the global classification rate. Finally, the margins of the breast masses were identified by applying morphological operations. The goal was to efficiently identify the masses inside the images, for better segmentation. Erosion, with a radius of 2 pixels, removed isolated pixels that did not correspond to masses in the binary image. Subsequently, closing was applied with a radius of 10 pixels, to connect isolated regions that constituted breast masses. The remaining ROIs were considered as possible breast masses.

2) FEATURE EXTRACTION AND SELECTION FOR CLASSIFICATION

Machine learning was employed to eliminate regions falsely detected as masses and to classify the breast masses as benign or suspicious according to their BI-RADS category. Features were selected for their ability to identify masses and further characterize them as benign or suspicious. In total, 96 features were extracted from the ROIs, divided in four major categories: shape-based, intensity-based, First-Order Statistics (FOS) and Gray Level Co-occurrence Matrix (GLCM) features. Table 2, describes the features that were extracted. Shape is a particularly important radiological feature for the diagnosis of masses and their differentiation as benign or suspicious [4]. A total of 15 shape-based features were extracted from each ROI. Breast masses usually exhibit higher intensity compared to the background and other regions. Intensity features were also extracted to reflect these differences. Furthermore, the texture of a breast mass provides diagnostically useful information reflected in the FOS [31] and GLCM [32] features of each region. Each GLCM feature was extracted at 0, 45, 90 and 135 degrees and the mean and standard deviation (STD) were obtained, resulting in 24 values for each different offset D. 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.

Normalization was applied to each feature row, in order to scale all the samples and adjust the range of their values. A row-wise application of least squares (l2) normalization was used to normalize all the features of a single ROI. Standardization was not applied since it was not found to improve the classification [33].

Feature selection is a necessary step to eliminate irrelevant and redundant features. Four commonly used algorithms were tested: hypothesis t-test [34], SelectKBest [33], feature importance [33] and Principal Component Analysis (PCA) [35], to identify the most important features with the highest contribution to the classification performance. A combination of paired t-test and feature importance, provided the best classification performance. Feature Importance provides a score for each extracted feature, in this case

TABLE 2. Features extracted for the diagnosis of breast masses.

Name	Description
Shape-based features	
Area	number of pixels
Circularity	how compact or circular is the region
Compactness	degree of deviation of the mass from a perfect circle
Convex area	number of pixels in the convex image
Eccentricity	ratio of distance between the foci of the ellipse and its major axis length
Equivalent diameter	diameter of the circle with the same area as the region
Euler number	subtraction of the number of objects and the number of holes in those objects
Extent	ratio of pixels inside the region to pixels in the bounding box
Filled area	number of pixels inside the filled image
Major axis length	length of the major axis of an ellipse that has same normalized central moment with the region
Minor axis length	length of the minor axis of an ellipse that has same normalized central moment with the region
Orientation	angle between the axis of the ellipse that has same second moments with the region
Perimeter	distance around the boundary of the region
Solidity	relative amount of pixels that appeared in both the convex hull and the region
Shape ratio	proportional relationship of the width to the height
Intensity-based features	
Maximum intensity	maximum of all the intensity values
Minimum intensity	minimum of all the intensity values
FOS features	
Average intensity	mean of all the intensity values
Entropy	texture measurement
Kurtosis	correlated with probability distribution
Skewness	asymmetry measurement
Smoothness	measures the relative smoothness of intensity
Standard deviation	variation from the average value
Variance	how far are the pixels values from the average
GLCM features	
Contrast	amount of local variations
Correlation	measurement of gray tone linear-dependencies
Energy	information measurement
Homogeneity	information about the distribution elements

96 scores. This score is calculated as the decrease in node impurity, weighted by the probability of reaching that node. As the score increases, the feature is marked as more important or relevant. Feature importance is an inbuilt class that comes with tree-based classifiers, thus extra tree classifier was applied in this study [33]. The optimal combination of features, using the 40 features with the highest importance value and the statistically significant features resulted from the t-test, was identified by optimizing the classification performance.

Unfortunately, the dataset created was imbalanced since (i) a large number of FP regions (39 FPs per image) were detected after image registration, subtraction and mass segmentation, resulting in the true masses being the minority class (148 true masses), and (ii) the number of benign and suspicious masses was not equal (52 vs. 96), reflecting their clinical incidence. Synthetic Minority Oversampling Technique (SMOTE) was implemented. SMOTE is a data augmentation approach which selects the instances closer to the feature space and creates new samples at points between them [36]. To avoid bias, SMOTE was applied only on the training set to automatically create new instances of the minority class. Hence, the data into the classifiers were balanced to increase the accuracy. Feature extraction was performed using MATLAB (R2020a; MathWorks, Massachusetts, U.S.A). The pre-processing of the features, the feature selection step and the classification, were performed using Python (version 3.7.7; Python Software Foundation, Delware, U.S.A).

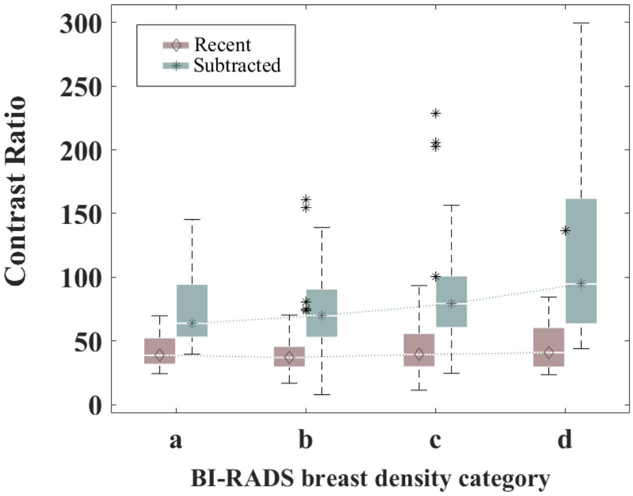

FIGURE 5. Plot comparing the contrast ratio of the processed recent image and the image created by temporal subtraction, for the four categories of breast density as defined by the BI-RADS.

3) TRAINING AND COMPARISON OF CLASSIFIER DESIGNS

Different classifiers have been proposed in the literature first to eliminate the falsely detected regions and then classify the breast masses as benign or suspicious [8]. In this study, nine popular classifiers were evaluated: Linear Discriminant Analysis (LDA) [37], k-Nearest Neighbor (k-NN) [38], Support Vector Machine (SVM) [39], Naive Bayes (NB) [38], Random Forest (RF) [40], AdaBoost (ADA) [41], Bagging (BAG) [42], Gradient Boosting (GB) [43], and Voting [41]. Ensemble models combined the predictions from multiple separate models to enhance the performance and reduce over-fitting.

In addition, different Neural Network (NN) architectures were evaluated using Python and Keras (version 2.3.1) [38]. All the available parameters of the network were tested and optimized based on the classification accuracy and the processing time. The goal was to built an efficient network with high classification performance and low processing time. For the classification of normal tissue vs. mass, the selected architecture consisted of 6 fully connected layers (96, 192, 384, 768, 768), with 990,386 trainable parameters. A Rectified Linear Unit (ReLU) was used as an activation function and batch normalization, along with adaptive dropout regularization (0.2-0.4), were included every 2 hidden layers. Gaussian noise was added, as a regularization term, in order to increase the robustness of the network. The batch size was set to 128,

TABLE 3. Comparison of the classification results of the segmented regions as normal tissue or true masses with temporal subtraction (TS) and temporal analysis (TA), using leave-one-patient-out cross-validation.

Classifier	Sensitivity [%]	Specificity [%]	Accuracy [%]	AUC
LDA	TS: 81.08 TA: 74.32	TS: 83.65 TA: 75.71	TS: 93.62 TA: 75.69	TS: 0.82 TA: 0.75
k-NN	TS: 92.57 TA: 60.14	TS: 98.38 TA: 87.39	TS: 98.31 TA: 87.07	TS: 0.95 TA: 0.74
SVM	TS: 65.54 TA: 71.62	TS: 87.96 TA: 79.69	TS: 87.70 TA: 79.59	TS: 0.77 TA: 0.76
NB	TS: 72.97 TA: 78.38	TS: 93.72 TA: 74.27	TS: 93.48 TA: 74.31	TS: 0.83 TA: 0.76
RF	TS: 88.51 TA: 39.86	TS: 99.68 TA: 98.13	TS: 99.55 TA: 97.45	TS: 0.94 TA: 0.69
ADA	TS: 77.03 TA: 68.24	TS: 89.33 TA: 86.24	TS: 89.19 TA: 86.03	TS: 0.83 TA: 0.77
BAG	TS: 83.78 TA: 43.92	TS: 99.43 TA: 97.29	TS: 99.24 TA: 96.67	TS: 0.92 TA: 0.71
GB	TS: 89.86 TA: 66.89	TS: 98.69 TA: 90.08	TS: 98.59 TA: 89.81	TS: 0.94 TA: 0.78
Voting	TS: 90.54 TA: 69.59	TS: 99.51 TA: 87.69	TS: 99.40 TA: 87.48	TS: 0.95 TA: 0.79
NN	**TS: 96.62** **TA: 66.89**	**TS: 99.93** **TA: 91.21**	**TS: 99.89** **TA: 90.93**	**TS: 0.98** **TA: 0.79**

TABLE 4. Comparison of the classification results of the true masses as benign or suspicious with temporal subtraction (TS) and temporal analysis (TA), using leave-one-patient-out cross-validation.

Classifier	Sensitivity [%]	Specificity [%]	Accuracy [%]	AUC
LDA	TS: 77.08 TA: 73.96	TS: 76.92 TA: 51.92	TS: 77.03 TA: 66.22	TS: 0.77 TA: 0.63
k-NN	TS: 84.38 TA: 80.21	TS: 96.15 TA: 55.77	TS: 88.51 TA: 71.62	TS: 0.90 TA: 0.68
SVM	TS: 80.21 TA: 90.63	TS: 98.08 TA: 36.54	TS: 86.49 TA: 71.62	TS: 0.89 TA: 0.64
NB	TS: 83.33 TA: 83.33	TS: 65.38 TA: 59.62	TS: 77.03 TA: 75.00	TS: 0.74 TA: 0.71
RF	TS: 84.38 TA: 77.08	TS: 82.69 TA: 69.23	TS: 83.78 TA: 74.32	TS: 0.83 TA: 0.73
ADA	TS: 87.50 TA: 77.08	TS: 90.38 TA: 57.69	TS: 88.51 TA: 70.27	TS: 0.89 TA: 0.67
BAG	TS: 87.50 TA: 70.83	TS: 82.69 TA: 57.69	TS: 85.81 TA: 66.22	TS: 0.85 TA: 0.64
GB	TS: 90.63 TA: 76.04	TS: 86.54 TA: 55.77	TS: 89.19 TA: 68.92	TS: 0.89 TA: 0.66
Voting	TS: 89.58 TA: 79.17	TS: 94.23 TA: 61.54	TS: 91.22 TA: 72.97	TS: 0.92 TA: 0.70
NN	**TS: 98.96** **TA: 93.75**	**TS: 96.15** **TA: 90.38**	**TS: 97.97** **TA: 92.57**	**TS: 0.98** **TA: 0.92**

the learning rate was set to 0.001 and the network was trained for 100 epochs. For the classification of benign vs. suspicious masses, the same network structure was used with appropriate modifications. In this case, batch normalization was not used and the learning rate was set to 0.0001, resulting in 985,922 trainable parameters.

The classification took place in two rounds. First, the classifiers were trained and tested to separate the detected ROIs into normal tissue and true masses. This step eliminated the FP detections resulting from the erroneous identification of normal tissue as abnormal. In the second classification round, the true masses were classified as benign or suspicious. In both rounds Leave-One-Patient-Out (LOPO) cross-validation was used for the training. All the images associated with a single patient (CC and MLO views of recent and prior mammograms) were reserved for the testing group, while the images of the remaining patients were used for training, repeating until all the cases were classified. In addition to LOPO cross-validation, k-fold cross-validation was also applied to examine the classification performance. In a similar manner, the folds were created per patient and not by randomly dividing the ROIs. Grouping the data per patient is of great importance to avoid any bias during the classification procedure resulting from information from the same patient included in both the training and test set.

For the evaluation of the two classification rounds and all the cross-validations schemes, sensitivity, specificity, accuracy and the Area Under the receiver operating characteristics Curve (AUC) were calculated.

C. BREAST MASS SEGMENTATION AND CLASSIFICATION USING TEMPORAL ANALYSIS

In temporal analysis, prior mammographic views are utilized. When prior information is available for direct comparison by the clinicians, abnormalities can be identified at an earlier stage and the clinicians feel more confident of their assessment [44]. Prior and recent images are coarsely registered based on anatomical features (nipple, skin, center of mass) and the locations of recently identified masses in the prior images are identified by regional registration. Combining the features from all images resulted in an increase in specificity and reduced FPs rates. However, temporal analysis offers no benefit, over using just the most recent mammogram, when the findings are new and with no traces of abnormality in the prior screening [14].

For comparison purposes, temporal analysis was also applied on the same data to verify the benefit of temporal subtraction. For the results to be comparable, the same procedures as before were followed. Hence, image registration, feature subtraction, mass segmentation and then machine learning (feature extraction, feature selection and classification) were optimized for FP removal and breast mass classification as benign or suspicious, using the temporal analysis data.

III. RESULTS

A. BREAST MASS SEGMENTATION AND CLASSIFICATION USING TEMPORAL SUBTRACTION

1) IMAGE REGISTRATION, SUBTRACTION AND MASS SEGMENTATION

As already mentioned in Section II-B.1, the prior and recent images were pre-processed for enhancement and border removal (Fig. 3). To effectively register the prior and recent mammographic views, Demons registration was used. The performance of temporal subtraction was evaluated by comparing the CR of the subtracted image, to the corresponding CR of the original most recent image after pre-processing (Fig. 5). The CR increased ~2 times (41.9 vs. 81.2) with the introduction of registration and temporal subtraction, resulting in a visually enhanced image containing only the newly developed ROIs or the regions that have changed significantly between the screenings. The processing time for these operations was an average of ~15 minutes per image pair (Intel Core i7 2 GHz; Intel Corp., Santa Clara, CA, USA).

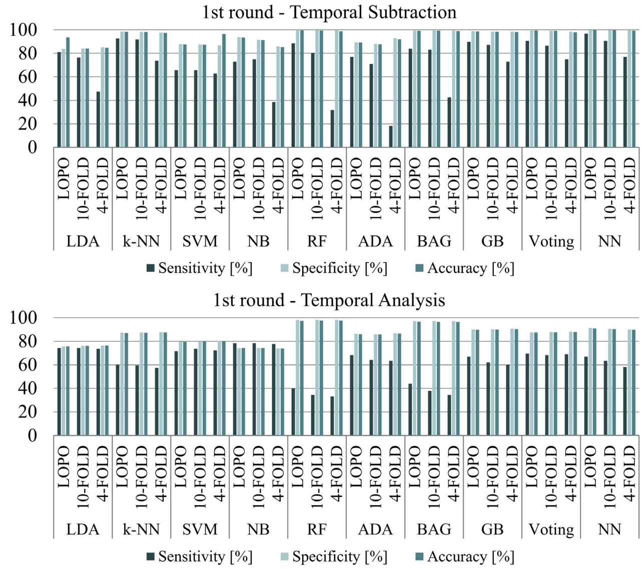

FIGURE 6. Classification results of the segmented regions as normal tissue or true masses using different classifiers and cross-validation methods. Top: Results using temporal subtraction. Bottom: Results using temporal analysis.

2) FEATURE EXTRACTION AND SELECTION FOR CLASSIFICATION

Feature selection revealed that the most significant features for the elimination of normal tissue misclassified as masses, i.e. the FPs, were: area, minimum intensity, average intensity, correlation 45° D_2, correlation 90° D_2, correlation 135° D_2, correlation mean D_2, contrast mean D_3, correlation 0° D_3, correlation 90° D_3, correlation 135° D_3, correlation mean D_3, compactness and the shape ratio. The most important features for the classification of the true masses as benign or suspicious were: major axis length, minor axis length, convex area, filled area, solidity, correlation 0° D_2, correlation 45° D_2, correlation 135° D_2, correlation mean D_2, correlation 0° D_3, correlation 135° D_3, correlation mean D_3, circularity.

3) TRAINING AND COMPARISON OF CLASSIFIER DESIGNS

First, the selected features were incorporated into various classifiers that were optimized for the elimination of FP detections, by classifying the detected ROIs as true masses or normal tissue, using LOPO cross-validation. The NN achieved the highest and most stable classification performance, with 99.9% accuracy and 0.98 AUC (Table 3). For the SVM, a linear kernel was used, k-NN was implemented with k set to 1, and for the Ensemble Voting, 1-NN, RF and GB were combined, in a hard voting scheme. Also, k-fold cross-validation was applied, using 4 and 10-folds. Overall, the performance remained approximately at the same level, proving the robustness of the algorithm. The overall classification results are shown in Fig. 6.

Subsequently, the true masses were classified as benign or suspicious using the selected features in a LOPO cross-validation scheme. The results are shown in Table 4. The highest and most robust performance was reached using a NN, with 98% accuracy and 0.98 AUC. For the SVM, a linear kernel was used and for k-NN, the number of neighbors was set to 7. In Ensemble Voting, 7-NN, SVM with a linear kernel, ADA, BAG and GB were combined, in a soft voting scheme. For the k-fold cross-validation, 4 and 13-fold cross-validation was applied, since of the 80 cases, only 52 had true masses. The results (Fig. 7) confirm the robustness of the algorithm.

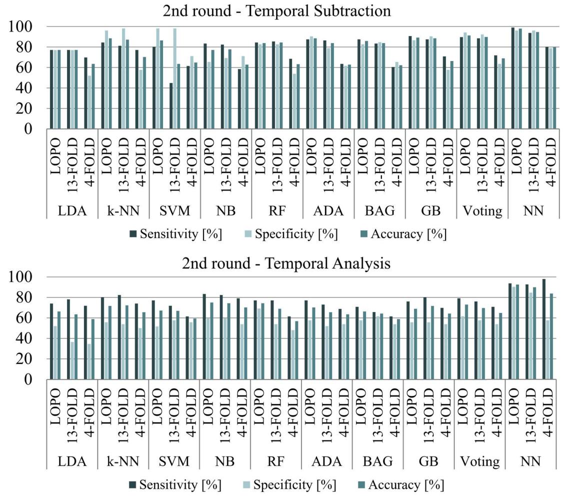

FIGURE 7. Classification results of the detected true masses as benign or suspicious using different classifiers and cross-validation methods. Top: Results using temporal subtraction. Bottom: Results using temporal analysis.

TABLE 5. Comparison of the Sensitivity (SE), Specificity (SP) and Accuracy (AC) of the proposed algorithm with and without the application of SMOTE, for the 5 classifiers with the best performance, using leave-one-patient-out cross-validation.

Classifier	normal vs. masses				benign vs. suspicious			
	without SMOTE		with SMOTE		without SMOTE		with SMOTE	
k-NN	SE:	59.46	SE:	92.57	SE:	90.63	SE:	84.38
	SP:	99.89	SP:	98.38	SP:	88.46	SP:	96.15
	AC:	99.42	AC:	98.31	AC:	89.86	AC:	88.51
ADA	SE:	39.86	SE:	77.03	SE:	88.54	SE:	87.50
	SP:	99.72	SP:	89.33	SP:	76.92	SP:	90.38
	AC:	99.02	AC:	89.19	AC:	84.46	AC:	85.81
GB	SE:	75.00	SE:	89.86	SE:	93.75	SE:	90.63
	SP:	99.95	SP:	98.69	SP:	75.00	SP:	86.54
	AC:	99.66	AC:	98.59	AC:	87.16	AC:	89.19
Voting	SE:	65.54	SE:	90.54	SE:	92.71	SE:	89.58
	SP:	99.99	SP:	99.51	SP:	78.85	SP:	94.23
	AC:	99.59	AC:	99.40	AC:	87.84	AC:	91.22
NN	SE:	89.40	**SE:**	**96.62**	SE:	96.88	**SE:**	**98.96**
	SP:	99.98	**SP:**	**99.93**	SP:	94.23	**SP:**	**96.15**
	AC:	99.85	**AC:**	**99.89**	AC:	95.95	**AC:**	**97.97**

In both rounds, the use of SMOTE resulted in a higher performance, as shown in Table 5, which exemplifies the importance of balancing the classification.

Figure 8 illustrates an example of the outcome of the algorithm for breast mass segmentation and classification using temporal subtraction. On the most recent mammographic view of a 58-year-old woman, two benign (green line) and one suspicious (red line) masses were all correctly classified, without any FPs detected in this image.

B. BREAST MASS SEGMENTATION AND CLASSIFICATION USING TEMPORAL ANALYSIS

Temporal analysis is the current state-of-the-art technique for the detection and classification of breast masses using sequential mammograms. The most important features for the first classification round (normal tissue vs. true masses)

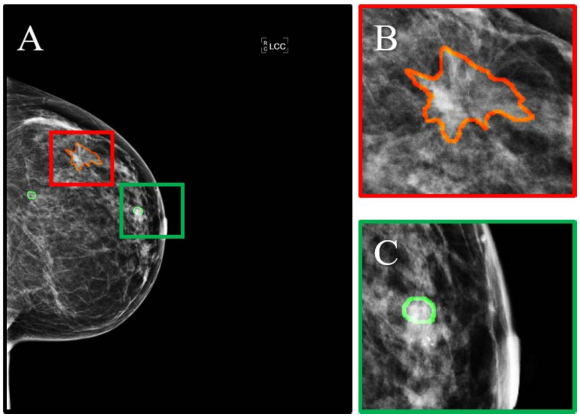

FIGURE 8. Results of the classification of the masses as benign or suspicious. (A) Recent mammographic view of a 58-year-old woman (BI-RADS breast density category c) with benign and suspicious masses. (B) Zoomed region marked by the red square in A, showing a suspicious mass. (C) Zoomed region marked by the green square in A, showing a benign mass.

were: skewness, kurtosis, STD, variance, correlation 45° D_2, correlation 90° D_2, correlation mean D_2, contrast 45° D_3, contrast 90° D_3, contrast 135° D_3, contrast mean D_3, correlation 90° D_3, correlation mean D_3 and smoothness. Using those in a LOPO cross-validation scheme, the NN reached 90.9% accuracy and 0.79 AUC (Table 3). The SVM was implemented using a linear kernel, the k-NN was performed with 11 neighbors and Ensemble Voting used LDA, Quadratic Discriminant Analysis (QDA), SVM with a linear kernel, Multi-Layer Perceptron (MLP) and GB, in a hard voting scheme. The application of 4 and 10-fold cross-validation resulted in ~5 drop in the classification performance (Fig. 6).

For the classification of the true masses as benign or suspicious, the same procedure was followed. The features with the highest contribution were: correlation 0° D_1, correlation mean D_1, contrast STD D_2, correlation 45° D_2, correlation STD D_2, energy STD D_2, homogeneity STD D_2, correlation 0° D_3, correlation 45° D_3, correlation mean D_3, correlation STD D_3, energy STD D_3 and smoothness. NN achieved 92.6% accuracy and 0.92 AUC using the selected features in a LOPO cross-validation scheme (Table 4). For this task, the SVM was implemented using a polynomial kernel, k-NN was implemented using 5 neighbors and the Ensemble Voting using 5-NN, RF and MLP, in a soft voting scheme. As before, since only 52 patients had true masses, the application of 4 and 13-fold cross-validation resulted in a slight drop of in terms of sensitivity, specificity and accuracy (Fig. 7).

IV. DISCUSSION

In this study, an automated algorithm for the segmentation and classification of breast masses, using subtraction of temporally sequential mammograms, is presented. The algorithm begins with pre-processing, image enhancement and border removal. Demons registration of prior to recent images was effective in adequately tracking the temporal changes between the screenings. After registration and subtraction, the CR of the subtracted images improved ~2 times, compared to the most recent mammographic view with pre-processing. This increase is clinically important since it could allow radiologists to better visualize the recent changes in the mammograms by eliminating unchanged and diagnostically insignificant features. Moreover, the radiologists could identify subtle abnormalities that otherwise would have been obscured by the background tissue and they could track the changes without having to manually refer back to the prior images.

For the classification of the detected ROIs as normal tissue or true masses, the classification accuracy was 99.9% which was achieved using a NN with a LOPO cross-validation scheme. The result was an average of 0.06 FP detections per image. Out of 148 true masses, 5 were misclassified as normal tissue and 9 normal regions were incorrectly identified as masses. For the characterization of the true masses as benign or suspicious, NN reached 98% accuracy using LOPO cross-validation, with an average of 0.012 FPs per image.

Despite the misclassifications, the actual clinical consequences would have been minimal if the algorithm was actually applied. While out of 52 benign masses, 2 were wrongly detected as suspicious, those patients also had other suspicious masses so they would have been followed up with biopsy despite the result. Similarly, out of 96 suspicious masses, 1 was misclassified as benign, but again since the patient had another suspicious mass, her care would not have been compromised as she would have been biopsied irrespectively. Although unlikely, a mass that is not changing between screenings could be subtracted and disappear from the final image. However, this does not impose any clinical consequences since in such cases only follow-up is recommended [45].

In addition to LOPO cross-validation, k-fold cross-validation was also applied to evaluate the robustness of the algorithm. The classification performance varied only slightly, depending on k, indicating the likely potential of the algorithm to correctly classify new data.

An important factor for an effective classification is a balanced dataset. In this study, the application of SMOTE, addressed the imbalance in the dataset, resulting in more robust results, especially in terms of sensitivity. For the classification of the detected ROIs as normal tissue or true masses, the improvement using SMOTE was more evident, since the dataset was more imbalanced. Without SMOTE, the sensitivity was low (except in the case of the NN), since the algorithm was biased to detect the normal ROIs. In the second classification round, from the 148 masses, 52 were benign and 96 were suspicious. Hence, the dataset was not as imbalanced. However, SMOTE, again, improved the classification performance.

The results presented indicate that temporal subtraction can be an effective technique for the segmentation and classification of breast masses using sequential mammogram pairs. 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. With temporal analysis, the NN

TABLE 6. Comparison between different state-of-the-art algorithms using sequential mammograms for the classification of breast masses as benign or malignant.

Method	Dataset	Classifier	AUC
Hadjiiski et al., 2001 [44]	140 pairs	LDA	0.88
Timp et al., 2007 [14]	465 pairs	SVM	0.77
Bozek et al., 2014 [45]	60 pairs	LDA	0.9
Ma et al., 2015 [47]	95 pairs	LDA	0.90
Kooi and Karssemeijer, 2017 [17]	18366 cases	CNN	0.88
Proposed Temporal Analysis	**80 cases**	**NN**	**0.92**
Proposed Temporal Subtraction	**80 cases**	**NN**	**0.98**

reached 90.9% accuracy using LOPO cross-validation, when distinguishing between normal tissue and masses. However, 49 masses incorrectly identified as normal tissue and there was an average of 6.9 FP detections per image. The results for the diagnosis of benign vs. suspicious masses were better, but still inferior to temporal subtraction. The NN achieved 92.6% accuracy with 0.03 FPs per image (two times higher compared to temporal subtraction). Five benign masses were misclassified as suspicious, which would have resulted in 1 unnecessary biopsy, and 6 suspicious masses were incorrectly identified as benign, resulting in the incorrect classification of 2 patients as healthy, while having suspicious masses. These results are consistent and even slightly better than those reported in the literature (Table 6). The average accuracy for the classification of benign vs. suspicious masses improved by ~5% with the introduction of temporal subtraction. The improvement using the proposed algorithm is statistically significant (p-value < 0.05).

Various groups have developed feature-based and deep learning approaches, for the classification of breast masses as benign or malignant, using the most recent mammographic view only. Despite the fact that the results are promising, the use of the most recent image alone does not allow for the evaluation of the information that has changed between the screenings. Furthermore, with the introduction of temporal subtraction, the classification accuracy increases. Table 7 compares different state-of-the-art algorithms for the classification of breast masses as benign or malignant, with the proposed method.

However, direct comparison of different studies is challenging due to differences in the method of cross-validation applied in each. In several cases, the ROIs were randomly divided into training and test sets, thus regions from the same image and the same patient were included in both the training and testing sets [45], [46]. In this study, to avoid such bias, the cross-validation was performed per patient and not per ROI.

Despite the very promising results of this study, the relatively small dataset remains a limitation. More sequential mammograms are required to confirm the effectiveness and robustness of the algorithm. Unfortunately, publicly available databases cannot be exploited, since they neither contain sequential mammograms, nor they include the detailed annotation of each individual mass, as in this study. Other limitations include the fact that the healthy participants were not followed for further diagnostic evaluation and that, although the suspicious masses were identified by two expert radiologists, differences might appear if more experts perform the same task.

TABLE 7. Comparison between different state-of-the-art algorithms for the classification of breast masses as benign or malignant.

Method	Dataset	Classifier	Accuracy
Rouhi et al., 2015 [46]	2781 images	MLP	96.5%
Al-masni et al., 2018 [48]	600 pairs	FC-NN	97%
Al-antari et al., 2018 [49]	410 pairs	CNN	95.6%
Arora et al., 2020 [11]	1318 ROIs	Deep NN	88%
Aly et al., 2020 [50]	107 cases	YOLO	89.5%
Gnanasekaran et al., 2020 [51]	1416 images	CNN	96.5%
Proposed Temporal Subtraction	**80 cases**	**NN**	**97.97%**

V. CONCLUSION

In this study, a method for the automatic breast mass segmentation and classification using subtraction of temporally sequential digital mammograms, was developed. With the application of temporal subtraction, the regions that remained unchanged between screenings, along with the background, were effectively removed, resulting in a new image with higher CR, compared to the corresponding recent mammographic view. The performance of the classification of breast masses as benign vs. suspicious increased by ~5%, compared to temporal analysis (98% vs. 92.6% accuracy) and by 1% compared to single view studies in the literature (98% vs. 97% accuracy), proving the effectiveness of the proposed technique.

The method presented here, as many other proposed accurate and efficient CAD approaches, have yet to be translated to clinical practice. Although their performance is encouraging, they are still not reliable enough to be accepted as standalone clinical tools by the BC community. An important factor limiting the acceptance of such systems are the datasets used. The unavailability of large-scale publicly available databases, forces researchers to independently collect private data, resulting in various datasets with different properties and imbalanced classes. Hence, the results reported in the literature, although achieving great performances, cannot be generalized and can only be used in a supporting role as a second reader in clinical practice.

Encouraged by this initial results, further studies are planned to include more patients with an extended age range. With further expansion and improvement, the proposed algorithm can provide even more detailed, BI-RADS-based, classification and 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] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2021," *CA, Cancer J. Clinicians*, vol. 71, no. 1, pp. 7–33, Jan. 2021, doi: 10.3322/caac.21654.

[2] K. Ganesan, U. R. Acharya, C. K. Chua, L. C. Min, K. T. Abraham, and K. H. Ng, "Computer-aided breast cancer detection using mammograms: A review," *IEEE Rev. Biomed. Eng.*, vol. 6, pp. 77–98, 2012.

[3] D. A. Spak, J. Plaxco, L. Santiago, M. Dryden, and B. Dogan, ''Bi-rads®fifth edition: A summary of changes,'' *Diagnostic Interventional Imag.*, vol. 98, no. 3, pp. 179–190, Dec. 2017.

[4] A. Oliver et al., ''A review of automatic mass detection and segmentation in mammographic images,'' *Med. Image Anal.*, vol. 14, no. 2, pp. 87–110, Feb. 2010.

[5] R. M. Rangayyan, F. J. Ayres, and J. E. L. Desautels, ''A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs,'' *J. Franklin Inst.*, vol. 344, nos. 3–4, pp. 312–348, May 2007.

[6] W. Xie, Y. Li, and Y. Ma, ''Breast mass classification in digital mammography based on extreme learning machine,'' *Neurocomputing*, vol. 173, pp. 930–941, Jan. 2016.

[7] S. J. S. Gardezi, A. Elazab, B. Lei, and T. Wang, ''Breast cancer detection and diagnosis using mammographic data: Systematic review,'' *J. Med. Internet Res.*, vol. 21, no. 7, Jul. 2019, Art. no. e14464.

[8] S. Zahoor, I. U. Lali, M. A. Khan, K. Javed, and W. Mehmood, ''Breast cancer detection and classification using traditional computer vision techniques: A comprehensive review,'' *Current Med. Imag. Formerly Current Med. Imag. Rev.*, vol. 16, no. 10, pp. 1187–1200, Jan. 2021.

[9] L. Zou, S. Yu, T. Meng, Z. Zhang, X. Liang, and Y. Xie, ''A technical review of convolutional neural network-based mammographic breast cancer diagnosis,'' *Comput. Math. Methods Med.*, vol. 2019, pp. 1–16, Mar. 2019.

[10] Z. Wang et al., ''Breast cancer detection using extreme learning machine based on feature fusion with CNN deep features,'' *IEEE Access*, vol. 7, pp. 105146–105158, 2019.

[11] R. Arora, P. K. Rai, and B. Raman, ''Deep feature–based automatic classification of mammograms,'' *Med. Biol. Eng. Comput.*, vol. 58, pp. 1–13, Mar. 2020.

[12] F. Ma, M. Bajger, S. Williams, and M. J. Bottema, ''Improved detection of cancer in screening mammograms by temporal comparison,'' in *Proc. Int. Workshop Digital Mammography*. Cham, Switzerland: Springer, 2010, pp. 752–759.

[13] L. Hadjiiski et al., ''Improvement in radiologists' characterization of malignant and benign breast masses on serial mammograms with computer-aided diagnosis: An ROC study,'' *Radiology*, vol. 233, no. 1, pp. 255–265, 2004.

[14] S. Timp, C. Varela, and N. Karssemeijer, ''Temporal change analysis for characterization of mass lesions in mammography,'' *IEEE Trans. Med. Imag.*, vol. 26, no. 7, pp. 945–953, Jul. 2007.

[15] A. A. Roelofs et al., ''Importance of comparison of current and prior mammograms in breast cancer screening,'' *Radiology*, vol. 242, no. 1, pp. 70–77, 2007.

[16] C. M. Hakim et al., ''Effect of the availability of prior full-field digital mammography and digital breast tomosynthesis images on the interpretation of mammograms,'' *Radiology*, vol. 276, no. 1, pp. 65–72, 2015.

[17] T. Kooi and N. Karssemeijer, ''Classifying symmetrical differences and temporal change for the detection of malignant masses in mammography using deep neural networks,'' *J. Med. Imag.*, vol. 4, no. 4, 2017, Art. no. 044501.

[18] Y. Zheng, C. Yang, and A. Merkulov, ''Breast cancer screening using convolutional neural network and follow-up digital mammography,'' *Proc. SPIE*, vol. 10669, May 2018, Art. no. 1066905.

[19] K. Loizidou, G. Skouroumouni, C. Pitris, and C. Nikolaou, ''Digital subtraction of temporally sequential mammograms for improved detection and classification of microcalcifications,'' *Eur. Radiol. Experim.*, vol. 5, no. 1, pp. 1–12, Dec. 2021.

[20] *Breast Masses Dataset With Precisely Annotated Sequential Mammograms—Zenodo*. Accessed: Oct. 13, 2022, doi: 10.5281/zenodo.7179856.

[21] S. Agrawal, R. Rangnekar, D. Gala, S. Paul, and D. Kalbande, ''Detection of breast cancer from mammograms using a hybrid approach of deep learning and linear classification,'' in *Proc. Int. Conf. Smart City Emerg. Technol. (ICSCET)*, Jan. 2018, pp. 1–6.

[22] S.-C. Huang, F.-C. Cheng, and Y.-S. Chiu, ''Efficient contrast enhancement using adaptive gamma correction with weighting distribution,'' *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 1032–1041, Mar. 2013.

[23] R. Gonzalez, R. Woods, and S. Eddins, *Digital Image Processing Using MATLAB*, 2nd ed. Knoxville, TN, USA: Gatesmark Publishing, 2010, pp. 180–193.

[24] K. Marias, C. Behrenbruch, S. Parbhoo, A. Seifalian, and M. Brady, ''A registration framework for the comparison of mammogram sequences,'' *IEEE Trans. Med. Imag.*, vol. 24, no. 6, pp. 782–790, Jun. 2005.

[25] Y. Guo, R. Sivaramakrishna, C.-C. Lu, J. S. Suri, and S. Laxminarayan, ''Breast image registration techniques: A survey,'' *Med. Biol. Eng. Comput.*, vol. 44, nos. 1–2, pp. 15–26, Mar. 2006.

[26] Y. Díez et al., ''Revisiting intensity-based image registration applied to mammography,'' *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 5, pp. 716–725, Nov. 2011.

[27] X. Pennec, P. Cachier, and N. Ayache, ''Understanding the, 'demon's algorithm': 3D non-rigid registration by gradient descent,'' in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, 1999, pp. 597–605.

[28] H. Lu et al., ''Multi-modal diffeomorphic demons registration based on point-wise mutual information,'' in *Proc. IEEE Int. Symp. Biomed. Imag. Nano Macro*, Dec. 2010, pp. 372–375.

[29] H. P. Chan, C. J. Vyborny, H. MacMahon, C. E. Metz, K. Doi, and E. A. Sickles, ''Digital mammography. ROC studies of the effects of pixel size and unsharp-mask filtering on the detection of subtle microcalcifications,'' *Investigative Radiol.*, vol. 22, no. 7, pp. 581–589, 1987.

[30] N. Otsu, ''A threshold selection method from gray-level histograms,'' *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. SMCS-9, no. 1, pp. 62–66, Feb. 1979.

[31] V. Kumar and P. Gupta, ''Importance of statistical measures in digital image processing,'' *Int. J. Emerg. Technol. Adv. Eng.*, vol. 2, no. 8, pp. 56–62, 2012.

[32] R. M. Haralick, K. Shanmugam, and I. Dinstein, ''Textural features for image classification,'' *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[33] F. Pedregosa et al., ''Scikit-learn: Machine learning in Python,'' *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[34] P. Diehr, D. C. Martin, T. Koepsell, and A. Cheadle, ''Breaking the matches in a pairedt-test for community interventions when the number of pairs is small,'' *Statist. Med.*, vol. 14, no. 13, pp. 1491–1504, Jul. 1995.

[35] H. Abdi and L. J. Williams, ''Principal component analysis,'' *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.

[36] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, ''SMOTE: Synthetic minority over-sampling technique,'' *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, Jan. 2002.

[37] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. R. Mullers, ''Fisher discriminant analysis with kernels,'' in *Proc. Neural Netw. Signal Process., IEEE Signal Process. Soc. Workshop*, 1999, pp. 41–48.

[38] N. I. R. Yassin, S. Omran, E. M. F. El Houby, and H. Allam, ''Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review,'' *Comput. Methods Programs Biomed.*, vol. 156, pp. 25–45, Mar. 2018.

[39] C. Cortes and V. Vapnik, ''Support-vector networks,'' *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[40] L. Breiman, ''Random forests,'' *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[41] E. Bauer and R. Kohavi, ''An empirical comparison of voting classification algorithms: Bagging, boosting, and variants,'' *Mach. Learn.*, vol. 36, pp. 105–139, Jul. 1999.

[42] L. Breiman, ''Bagging predictors,'' *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[43] J. H. Friedman, ''Stochastic gradient boosting,'' *Comput. Statist. Data Anal.*, vol. 38, no. 4, pp. 367–378, Feb. 2002, doi: 10.1016/S0167-9473(01)00065-2.

[44] L. Hadjiiski, B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie, and M. Gurcan, ''Analysis of temporal changes of mammographic features: Computer-aided classification of malignant and benign breast masses,'' *Med. Phys.*, vol. 28, no. 11, pp. 2309–2317, Nov. 2001.

[45] J. Bozek, M. Kallenberg, M. Grgic, and N. Karssemeijer, ''Use of volumetric features for temporal comparison of mass lesions in full field digital mammograms,'' *Med. Phys.*, vol. 41, no. 2, Jan. 2014, Art. no. 021902.

[46] R. Rouhi, M. Jafari, S. Kasaei, and P. Keshavarzian, ''Benign and malignant breast tumors classification based on region growing and CNN segmentation,'' *Expert Syst. Appl.*, vol. 42, no. 3, pp. 990–1002, 2015.

[47] F. Ma, L. Yu, G. Liu, and Q. Niu, ''Computer aided mass detection in mammography with temporal change analysis,'' *Comput. Sci. Inf. Syst.*, vol. 12, no. 4, pp. 1255–1272, 2015.

[48] M. A. Al-masni et al., ''Simultaneous detection and classification of breast masses in digital mammograms via a deep learning Yolo-based cad system,'' *Comput. Methods Programs Biomed.*, vol. 157, pp. 85–94, Jun. 2018.

[49] M. A. Al-Antari, M. A. Al-Masni, M.-T. Choi, S.-M. Han, and T.-S. Kim, ''A fully integrated computer-aided diagnosis system for digital X-ray mammograms via deep learning detection, segmentation, and classification,'' *Int. J. Med. Inform.*, vol. 117, pp. 44–54, Sep. 2018.

[50] G. H. Aly, M. Marey, S. A. El-Sayed, and M. F. Tolba, ''YOLO based breast masses detection and classification in full-field digital mammograms,'' *Comput. Methods Programs Biomed.*, vol. 200, Mar. 2021, Art. no. 105823.

[51] V. S. Gnanasekaran, S. Joypaul, P. Meenakshi Sundaram, and D. D. Chairman, ''Deep learning algorithm for breast masses classification in mammograms,'' *IET Image Process.*, vol. 14, no. 12, pp. 2860–2868, Oct. 2020.