Computer-aided diagnosis of masses in breast computed tomography imaging: deep learning model with combined handcrafted and convolutional radiomic features

Marco Caballo,[a] Andrew M. Hernandez,[b] Su Hyun Lyu,[c] Jonas Teuwen,[a,d] Ritse M. Mann,[a,e] Bram van Ginneken,[a] John M. Boone,[b,c] and Ioannis Sechopoulos[a,f,*]

[a]Radboud University Medical Center, Department of Medical Imaging, Nijmegen, The Netherlands
[b]University of California Davis, Department of Radiology, Sacramento, California, United States
[c]University of California Davis, Department of Biomedical Engineering, Sacramento, California, United States
[d]The Netherlands Cancer Institute, Department of Radiation Oncology, Amsterdam, The Netherlands
[e]The Netherlands Cancer Institute, Department of Radiology, Amsterdam, The Netherlands
[f]Dutch Expert Center for Screening, Nijmegen, The Netherlands

Abstract

Purpose: A computer-aided diagnosis (CADx) system for breast masses is proposed, which incorporates both handcrafted and convolutional radiomic features embedded into a single deep learning model.

Approach: The model combines handcrafted and convolutional radiomic signatures into a multi-view architecture, which retrieves three-dimensional (3D) image information by simultaneously processing multiple two-dimensional mass patches extracted along different planes through the 3D mass volume. Each patch is processed by a stream composed of two concatenated parallel branches: a multi-layer perceptron fed with automatically extracted handcrafted radiomic features, and a convolutional neural network, for which discriminant features are learned from the input patches. All streams are then concatenated together into a final architecture, where all network weights are shared and the learning occurs simultaneously for each stream and branch. The CADx system was developed and tested for diagnosis of breast masses ($N = 284$) using image datasets acquired with independent dedicated breast computed tomography systems from two different institutions. The diagnostic classification performance of the CADx system was compared against other machine and deep learning architectures adopting handcrafted and convolutional approaches, and three board-certified breast radiologists.

Results: On a test set of 82 masses (45 benign, 37 malignant), the proposed CADx system performed better than all other model architectures evaluated, with an increase in the area under the receiver operating characteristics curve (AUC) of 0.05 ± 0.02, and achieving a final AUC of 0.947, outperforming the three radiologists ($\text{AUC} = 0.814 - 0.902$).

Conclusions: In conclusion, the system demonstrated its potential usefulness in breast cancer diagnosis by improving mass malignancy assessment.

© *2021 Society of Photo-Optical Instrumentation Engineers (SPIE)* [DOI: 10.1117/1.JMI.8.2.024501]

Keywords: breast cancer; breast computed tomography; computer-aided diagnosis; deep learning; radiomics.

Paper 20325R received Nov. 30, 2020; accepted for publication Mar. 12, 2021; published online Mar. 29, 2021.

[*] Address all correspondence to Ioannis Sechopoulos, Ioannis.Sechopoulos@radboudumc.nl

1 Introduction

Radiomics is a growing medical image analysis field that aims at the extraction of mineable information from medical images, with the goal of developing computerized support systems to improve clinical decision making.[1]

Among many potential clinical applications, cancer diagnosis is an area where radiomics is growing substantially, where several advancements in the development of computer-aided diagnosis (CADx) systems are being proposed.[2–4] Although radiomics is being investigated in many oncological fields, as could be expected efforts have been focused on the cancers with the highest incidence and mortality, with lung,[5] colon,[6] and breast cancer[7] being at the forefront.[8] In breast cancer imaging, the potential of radiomic-based CADx systems has been studied for numerous modalities used in the clinical routine, including digital mammography, digital breast tomosynthesis, and breast magnetic resonance imaging (MRI).[9–11] Less work has, instead, been reported on more recently developed imaging technologies. One of these technologies is dedicated breast computed tomography (bCT), a modality that brings the advantages of high-resolution, high-contrast imaging to the breast, providing 3D volume data sets with contrast and resolution characteristics optimized for breast tissues and lesions.[12,13] bCT overcomes the tissue superimposition problems of mammography and, to a lesser extent, of digital breast tomosynthesis, resulting in improved visualization of breast masses.[14] Due to these characteristics, the application of radiomic-based CADx techniques to bCT may result in the generation of important information for breast mass characterization, potentially improving clinical performance.

For the development of CADx systems in medical imaging, two types of design have been predominantly used in literature. They are based on either an engineered design and development of handcrafted radiomic descriptors or on convolutional neural networks (CNNs). The former approach involves a multi-step pipeline for segmentation of the region of interest, feature extraction and selection, and final classification.[9,15] The latter is instead able to directly extract the most appropriate features from the input radiological image and learn to perform the diagnostic task in an end-to-end fashion.[16]

Despite the demonstrated overall higher performance of CNNs over handcrafted image analysis pipelines, interest in both types of radiomic frameworks has been growing exponentially over the last decade.[15] While strengths and weaknesses of both methods have been repeatedly demonstrated by numerous studies,[15] less work has been reported that evaluates the combination of the two strategies, studies their potentially most effective merging approach, or directly compares their performance using the same patient image dataset. Some investigators have combined the two approaches through a feature-level fusion, where handcrafted and convolutional radiomic features are extracted separately from the same image, concatenated, and fed to a late-stage classifier to obtain the final predicted outcome.[17–21] Other studies have instead opted for decision-level merging, where the two feature sets are first processed by independent machine learning models, and then the outcomes are combined through voting, averaging, or other fusion approaches.[22–25]

Most of these studies focus on diagnosis,[22–25] with a minority in the fields of detection,[17] cancer survival,[18,19] and segmentation,[21] and reported overall benefit when the handcrafted and convolutional approaches were combined. However, to the best of our knowledge, most of the previously proposed studies consider handcrafted features and CNNs as separate branches, which are merged at a later stage either via feature concatenation, or via decision-level fusion.

In this work, we developed a CADx system that incorporates both strategies in a single deep learning model, where learning occurs simultaneously for both the handcrafted and the convolutional branch. The CADx system was developed to estimate the likelihood of malignancy in bCT mass images, and leverages multiple computational blocks, in parallel and in cascade, devoted to automated segmentation, feature extraction, and analysis. The classification performance of the developed model was assessed on a test set of breast masses acquired with multiple bCT systems, and then compared to that of both handcrafted and convolutional approaches applied separately and to the performance of additional feature- and decision-level fusion methods.

2 Materials and Methods

The proposed deep learning system for computerized mass diagnosis works by processing multiple two-dimensional (2D) views extracted from the same mass using multiple streams, each of which is composed of two major branches. One branch processes the input image patch through a handcrafted feature-based pipeline, including automatic mass segmentation, radiomic feature extraction and selection, and a multilayer perceptron (MLP) neural network. The other branch consists of a single CNN that converts the input image patch into the corresponding feature vector after convolutional and pooling operations. The two branches are merged together into a single stream, where backpropagation occurs simultaneously for both branches during training and outputs a predicted malignancy probability on a mass image patch-basis. The predicted probabilities from multiple 2D mass patches are then merged by evaluating different output fusion techniques, to convert them into a single per-mass classification outcome.

2.1 *Breast CT Systems and Image Dataset*

The first cohort of patient data sets used in this study (Dataset 1) was acquired using the first and second generation dedicated bCT systems designed and developed at the University of California (UC), Davis (California). Their use in clinical studies was performed under several Institutional review board-approved protocols.[14,26] Each bCT patient scan consisted of a total of 500 projections acquired over a 360-deg orbital extent, which were reconstructed using a filtered backprojection algorithm at an isotropic voxel size of 0.38 mm. The reconstructed volumes were then corrected for shading artifacts using a maximum-likelihood polynomial fitting approach in the reconstruction space.[27]

The second dataset (Dataset 2) was acquired with a bCT system[28] (Koning Corp., West Henrietta, New York) installed at Radboud University Medical Center (Nijmegen, The Netherlands), with similar geometry, detector, and dose characteristics as those at UC Davis.[29] Each bCT scan consisted of 300 projections being acquired over the full angular range and reconstructed with an isotropic voxel size of 0.273 mm using filtered backprojection. During reconstruction, images were corrected for cupping artifacts using a proprietary correction method. The dimensions of a bCT volume depend on the size needed to encompass the whole breast, with the voxel size being fixed, and the number of voxels varying according to the image field of view. The typical overall dimensions were ∼20 cm along each anatomical direction (coronal, axial, and sagittal).

Both image datasets were acquired by trained radiographers, as part of other ethics board-approved patient trials on bCT. Informed, written consent was acquired from all participating women. Dataset 1, from UC Davis, consisted of a total of 462 images. The minimum age to participate in the trial was 35 years. 170 women who had suspicious (BI-RADS® 4 and 5) lesions at mammography were recruited and imaged between October 2005 and November 2010. Both the ipsilateral and the contralateral breast were imaged, with some acquisitions performed both without ($n = 340$) and with ($n = 122$) injection of contrast-enhancing material.

Dataset 2, from Radboudumc, consisted of a total of 156 images acquired between October 2016 and November 2019. These were acquired from women at least 50 years old with a suspicious finding detected at mammographic screening (BI-RADS® 0, 4, or 5). Only the ipsilateral breast was imaged, and no contrast-enhancing material was used for any acquisition.

In the UC Davis trial, exclusion criteria were women with a recent breast biopsy, history of chronic asthma, positive urine pregnancy test or currently breast-feeding, suspected or confirmed pregnancy, inability to lie motionless for up to 20 s, inability to understand the risk and benefits of the study, and prisoners. For the Radboudumc trial, exclusion criteria were the presence of the suspicious lesion in the axillary tail, bilateral mastectomy, suspected or confirmed pregnancy, prior breast cancer or breast biopsy in the recalled breast in the last 12 months, presence of palpable lesions, frailty, breastfeeding, or inability to cooperate.

For this study, only unenhanced images with visible breast masses contained in the field of view, without any severe artifacts and with patient and radiological information available, were selected. Masses were identified and localized on the images by experienced breast radiologists, and all solid masses were biopsy-proven.

After case selection, the final dataset used in this study consisted of a total of 284 masses (192 from Dataset 1, 92 from Dataset 2) collected from 211 patient images (138 from Dataset 1, 73 from Dataset 2). Of all masses, 106 were malignant (77 from Dataset 1, 29 from Dataset 2), and 178 were benign (115 from Dataset 1, 63 from Dataset 2).

Additional details about geometry, detector, and acquisition protocols for the bCT systems used in this study were previously reported.[29]

2.2 *Image Preprocessing*

After image collection, all cases were preprocessed to obtain a consistent dataset from the two bCT systems, including the upscaling through linear interpolation of the images from Dataset 1 to the same voxel size as Dataset 2 (0.273 mm), and the compression of the dynamic range to 8 bits for all cases.

Subsequently, all masses were divided into training, validation, and test set in a random manner (stratifying on the class, benign/malignant), in an approximate proportion of 65% for training, 5% for validation, and 30% for testing. Specifically, the training set consisted of 187 masses (81 and 43 benign, and 51 and 12 malignant from Dataset 1 and Dataset 2, respectively); the validation set consisted of 15 masses (6 and 3 benign, and 4 and 2 malignant from Dataset 1 and Dataset 2, respectively); the test set consisted of 82 masses (28 and 17 benign, and 22 and 15 malignant from Dataset 1 and Dataset 2, respectively). Details on individual masses of the test set, used to validate all our methods, are reported in Table 1.

To simplify the classification of three-dimensional (3D) masses into a 2D problem, and at the same time to increase the dataset size to prevent overfitting, the CADx system was designed to work with multiple input 2D image patches that can characterize the lesion over multiple image views.[29,30] For this, nine square patches were extracted from each mass. The direction of each patch was parallel to one of the nine symmetry planes of an imaginary cube circumscribing the mass, corresponding to the coronal, sagittal, axial, and six oblique views, and intersecting the mass center. The size of each patch was kept at 128×128 pixels for all directions.

These dimensions were sufficient to enclose each mass in our dataset in patches of a fixed size. Masses were not rescaled to the size of the patches, so as to retain the original lesion size information.

Some examples of breast masses and their respective nine views are shown in Fig. 1.

Table 1 List and details of test set breast masses ($N = 82$).

Malignant masses ($n = 37$)	Invasive ductal carcinoma	19
	Ductal carcinoma *in situ*	6
	Invasive lobular carcinoma	2
	Invasive mammary carcinoma	3
	Adenocarcinoma	1
	Combination of tumor types	6
Benign masses ($n = 45$)	Cyst	29
	Fibroadenoma	10
	Atypical ductal hyperplasia	1
	Blunt duct adenosis	1
	Hamartoma	1
	Lymph node	2
	Fibrocystic change	1

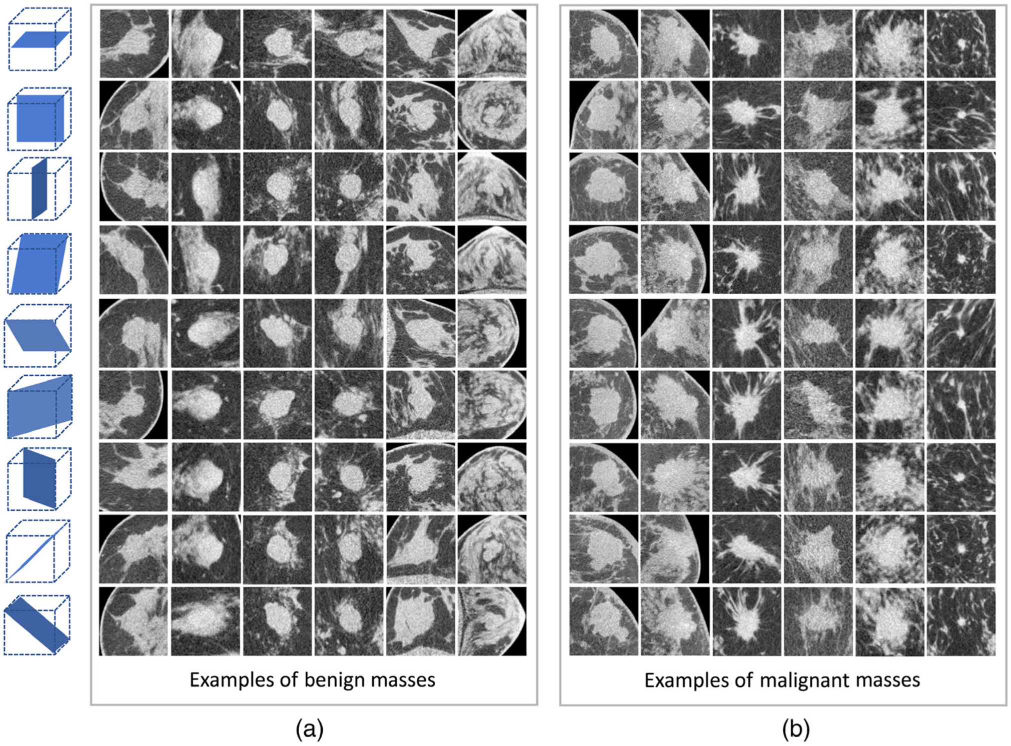

Fig. 1 Examples of breast masses and corresponding nine 2D patches (dimensions of 128×128 pixels) from different planar views through the 3D volume. (a) Examples of benign masses and (b) examples of malignant masses. Each column corresponds to the nine views extracted from the same mass.

This nine-view approach allowed to capture the masses from different angles, implicitly containing 3D information for each case, and provided a first augmentation strategy that, as opposed to traditional affine transformations, can be used for handcrafted radiomic analyses since the radiomic signature can change over different mass planes.

All extracted patches were then manually annotated using the polyline toolbox in ImageJ® (LOCI, National Institutes of Health, Bethesda, Maryland) by an image analysis scientist with over three years of experience in breast image analysis, under the supervision of a board-certified breast radiologist with experience in bCT imaging. The pairs of image patches and their respective manual annotations in the training and validation set were used to develop and optimize the segmentation model, and for handcrafted radiomic feature extraction (described later). The manually annotated patches from the test set were instead only used as a ground truth to evaluate the automatic segmentation performance.

After annotation, to address the skewness in class proportion, which could bias the prediction of deep learning algorithms toward the most frequently represented class, the training set was corrected to a balanced ratio between benign and malignant examples. For this, the malignant masses in the training set were oversampled by generating an additional example from each malignant image patch. Instead of simply copying the examples, affine data augmentation was used, with each augmented case being generated through random rotation (angle between 5 deg and 35 deg), random horizontal or vertical mirroring, and random shearing (shear ratio between 10% and 30%). During each transformation, possible missing data at the boundary of the augmented patches, for example, deriving from rotation or shearing operations, were handled through nearest-neighbor interpolation.

2.3 *U-Net for Automatic Mass Segmentation*

Prior to the extraction of handcrafted radiomic features, a U-Net[31] was trained to perform the mass patch segmentation. In the training phase, this encoder–decoder architecture learns to map

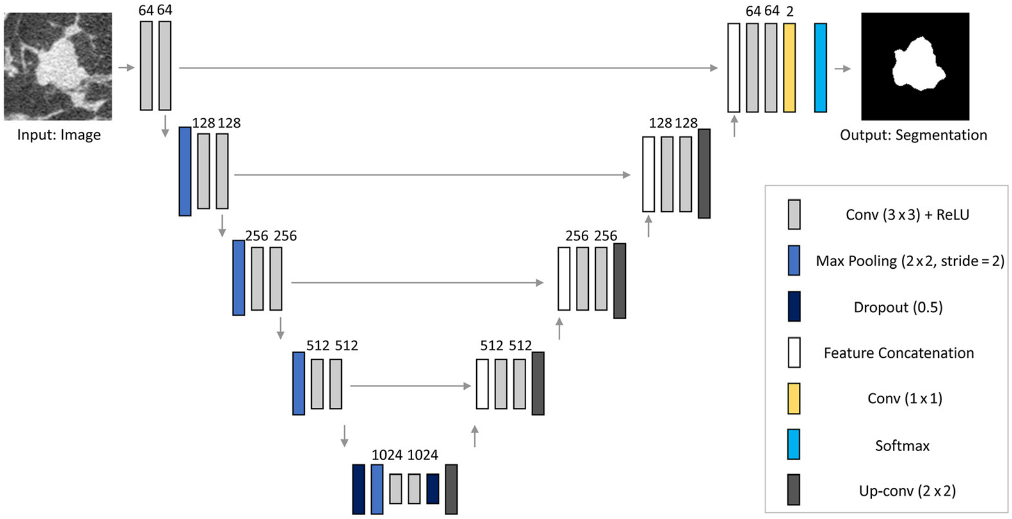

Fig. 2 U-Net architecture implemented and used for breast mass segmentation on a 2D patch-basis. The numbers above each convolutional block indicate the number of filters used for that convolutional operation.

between the input image patch and the respective manually annotated ground truth, learning the segmentation task in a supervised fashion. Specifically, the first half of the architecture encodes the input image into a feature vector via convolutional and max pooling operations. The second half recovers the information via nearest-neighbour upscaling and convolutional filters. The information is propagated between the two parts by concatenating the output of the encoding convolutional blocks with each respective decoding branch, allowing the preservation of high spatial detail. The encoding part consisted of four blocks of two 3×3 convolutions followed by ReLU activations, with a subsequent 2×2 max pooling layer (stride equal to 2). The decoding part consisted of four up-sampling blocks followed by a 2×2 convolutional layer that halved the number of channels, a concatenation with the corresponding features from the encoding part, and two consecutive 3×3 convolutional layers followed by ReLU activations. Dropout[32] (probability of 0.5) was used in the bottleneck of the network, before the last max pooling layer and before the first up-convolution, for regularization. The last block of the network consisted in a softmax layer that outputs the final probability map.[31] To train the network, the Adam optimization method[33] was used on mini-batches of 16 examples, with binary cross-entropy as loss function, and with an initial learning rate of 10^{-3}, decayed exponentially every 10 epochs for a total of 50 epochs.

The complete network architecture and the number of filters used in each convolutional layer are shown in Fig. 2. The trained network was applied to segment the test set mass patches, and the pairs image patches—segmented masks were used for handcrafted radiomic feature extraction on the test set.

2.4 *Radiomic Feature Extraction and Selection*

Handcrafted radiomic features were extracted using a previously proposed algorithm[29] able to quantify mass characteristics related to texture, shape and contour, and margin, resulting in a total of 1354 descriptors. Briefly, the algorithm calculates: (i) the image texture inside the mass and along its boundary, by quantifying the mass heterogeneity using first-order,[34] Haralick,[35] run length,[36] structural and pattern (including local binary pattern,[37] Hessian[38] and Laws masks,[39] and fractal dimension[40]), and Gabor descriptors;[41] (ii) the mass degree of irregularity and spiculation, using Fourier descriptors applied to the mass boundary and centroid distance function,[42–44] and automatic mapping of spiculae and lobes; and (iii) the mass degree of infiltration and potential differences in peritumoral compartments, using radial gradient analysis and by evaluating textural differences over different peripheral mass region.[45–47]

Given the large initial dimension of the feature space, feature selection was performed to minimize the risk of overfitting and to remove potentially uninformative features. For this, all radiomic features were first extracted from the mass patches of the training set, and the less informative features were discarded through a three-step feature reduction process,[29] which included stability and statistical analysis, and the final application of the Relieff algorithm.[48] Stability analysis was performed by assessing the intra-class correlation of the radiomic features with respect to segmentation variability of multiple breast radiologists and to the image acquisition settings of the different bCT systems. This allowed to retain only those features whose value was minimally affected by potential differences in mass segmentation due to inter-reader variability and by differences in imaging conditions. Statistical analysis was performed with univariate analyses, corrected for multiple comparisons, using the Mann–Whitney U-test, to discard features with low diagnostic value on the training set. Finally, the Relieff algorithm was used to rank and select the most informative features, with the aim of reducing the feature set dimensionality for model overfitting prevention.

Complete details about feature extraction and selection were previously reported.[29] The process resulted in the vast majority of features being discarded, with the number of selected features dropping from 1354 to 36. These selected features, detailed and discussed in our previous study,[29] quantify the texture ($n = 3$), shape ($n = 1$), and margin definition ($n = 26$) and diversity ($n = 6$) of breast masses, allowing for the capture of multiple imaging biomarkers related to malignancy.

2.5 *CADx System*

After developing the segmentation model and extracting the handcrafted radiomic features, a CADx system was developed for automatic classification of breast mass patches. Given an input mass patch, the proposed architecture processes the patch through two major branches: the handcrafted-based, and the convolutional-based (Fig. 3). The former is composed of the cascade of automatic segmentation performed by the U-Net, a module for the extraction of the selected radiomic features, and an MLP network aiming at processing the mass handcrafted radiomic descriptors. The MLP is composed of four fully connected (FC) layers (size: 64; 64; 32; and

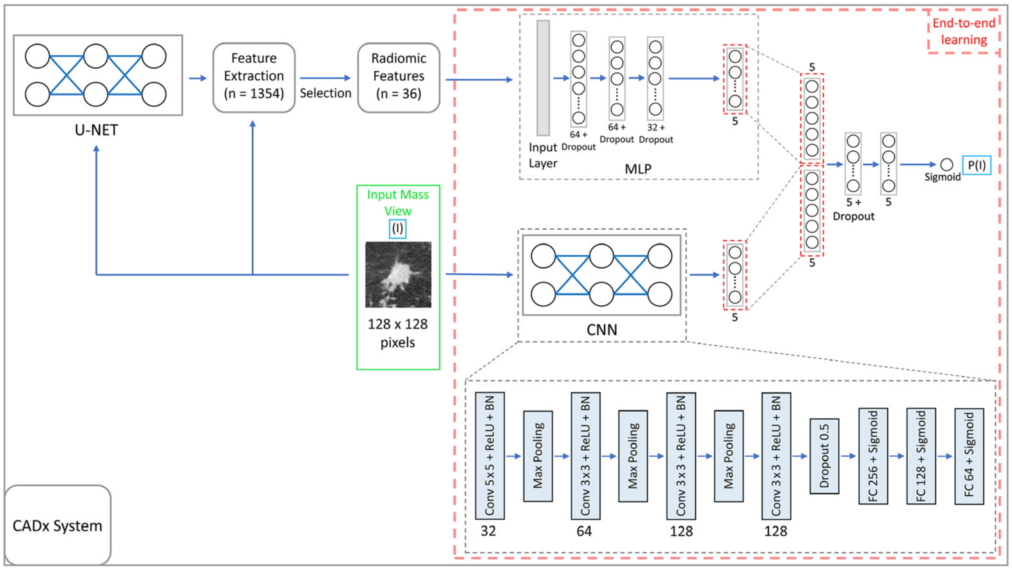

Fig. 3 Architecture of the CADx system. 2D patches are used as the input to the network and are processed in parallel by a handcrafted branch (U-Net for automatic segmentation, radiomic feature extraction, selection, and an MLP network), and a convolutional branch including a deep CNN. The last FC layers from the two branches are concatenated, further processed by two additional FC layers, and a final logistic unit that outputs the patch malignancy probability. In this double-input architecture, the learning occurs simultaneously for both branches, with the errors backpropagated throughout the whole architecture in an end-to-end fashion.

5 units), with each output propagated to the subsequent layer through ReLU activations. Dropout regularization (probability of 0.5) was used in each layer to prevent overfitting. All weights of the network were randomly initialized (random uniform initialization between −0.05 and 0.05) and biases set to zero.

The convolutional-based branch, instead, involves a CNN that directly processes the input patch and converts the image feature space into the corresponding feature vector. The CNN is composed of four convolutional layers (number of filters: 32, 64, 128, and 128) with ReLU activations, alternating as many max pooling layers (all with a 2×2 kernel size, and a stride equal to 2), and four final FC layers (layer size: 256, 128, 64, and 5 units). Batch normalization[49] (momentum term: 0.99; epsilon: 10^{-3}) and dropout (probability of 0.5) were used as regularization (before each pooling layer, and before the first FC layer, respectively).

To complete the CADx architecture, the two branches were merged by concatenating the last FC layers from the MLP and the CNN, which are further processed by an additional two FC layers (five units each). The output of the whole network is given by a single logistic unit that performs the binary classification in terms of malignancy probability through a sigmoid activation function.

This architecture incorporates handcrafted and convolutional radiomic features in a single model, where learning can occur simultaneously by backpropagating the error throughout both branches at each training iteration. The model was trained for 250 epochs using the Adam optimizer, with a batch size of 128 elements, binary cross-entropy as loss function, and a starting learning rate of 10^{-3}. Architecture and hyperparameters were chosen and tuned using the validation set and were not further adjusted in the subsequent testing phase. Prior to training, to speed up convergence, both handcrafted radiomic features (for the MLP branch) and image patches (for the CNN branch) were normalized on the training set using min-max normalization.

2.6 *Comparison with Additional Architectures*

To compare the performance of the previously described CADx system over more traditional approaches, four additional models were developed and are schematized in Fig. 4. These include: (i) the use of only the MLP branch [Fig. 4(b)]; (ii) the use of only the CNN branch [Fig. 4(c)];

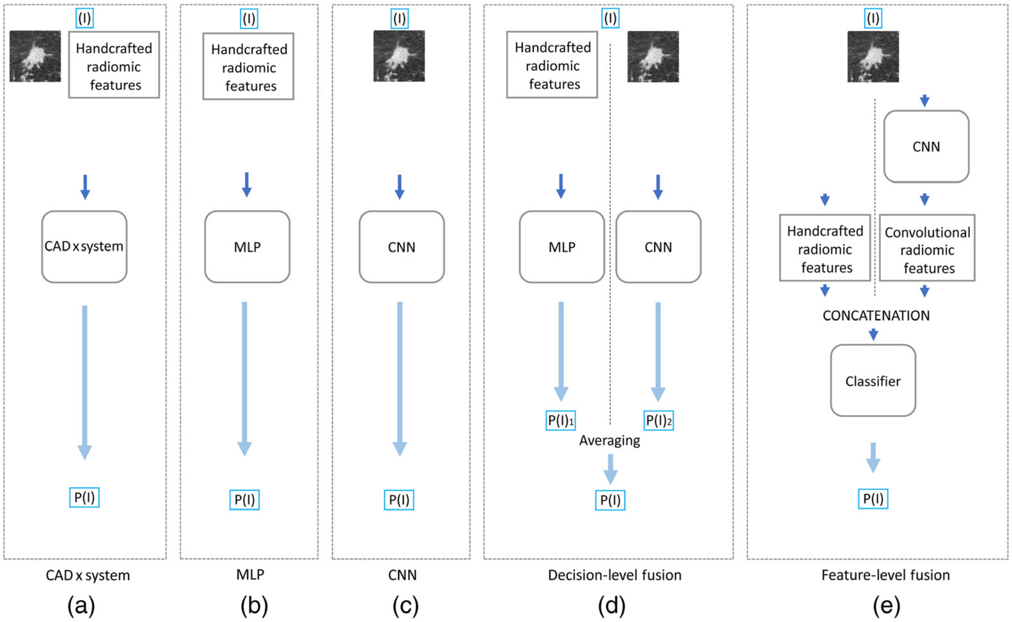

Fig. 4 (b)–(e) Schematics of the additional model architectures developed and used for comparisons with the proposed (a) CADx system. The CADx system architecture is shown in detail in Fig. 3.

(iii) a decision-level fusion, where the outputs from the two separated branches are combined at a late stage through averaging [Fig. 4(d)]; and (iv) a feature-level approach, where handcrafted and convolutional features are extracted separately by the two branches, and then fed to a late-stage classifier for outcome prediction [Fig. 4(e)]. Details about the implementation and training of these additional models are reported in Appendix.

2.7 *Multi-View Output Fusion Strategies*

After development of the patch-based CADx system and comparing its performance with additional architectures, different output fusion strategies were implemented to merge the predicted probabilities extracted from the nine views of each mass, so as to obtain a single malignancy probability score on a per-mass basis. To combine the nine-view mass malignancy probabilities into a single classification score, four output fusion methods were investigated (Fig. 5).

First, the nine predicted probabilities extracted from each mass view were averaged together, and this final probability score was thresholded to result in binary classification decision [Fig. 5(a)].

Second, the averaging operator was simply substituted with a majority voting approach [Fig. 5(b)], in which the nine-view probabilities from each mass were first dichotomized (0 for benign, 1 for malignant), and then the predicted per-mass classification outcome was given by the most frequently represented class.

In the third output fusion strategy, a more advanced approach to combine the malignancy likelihood of the nine views is proposed [Fig. 5(c)]. The goal of this approach is to determine how many views classified as malignant are sufficient to correctly consider the whole mass as malignant. This approach aims to test the hypothesis that only a few views with a high malignancy probability could be sufficient to correctly classify the whole mass as malignant, regardless of the predicted probability values of the other remaining views. For this, first a malignancy likelihood threshold T was defined (between 0 and 1), and then the whole mass was classified as malignant if at least N of the nine probabilities were larger than the selected threshold T. The experiment was repeated for five different values of T (0.1, 0.25, 0.5, 0.75, and 0.9), and

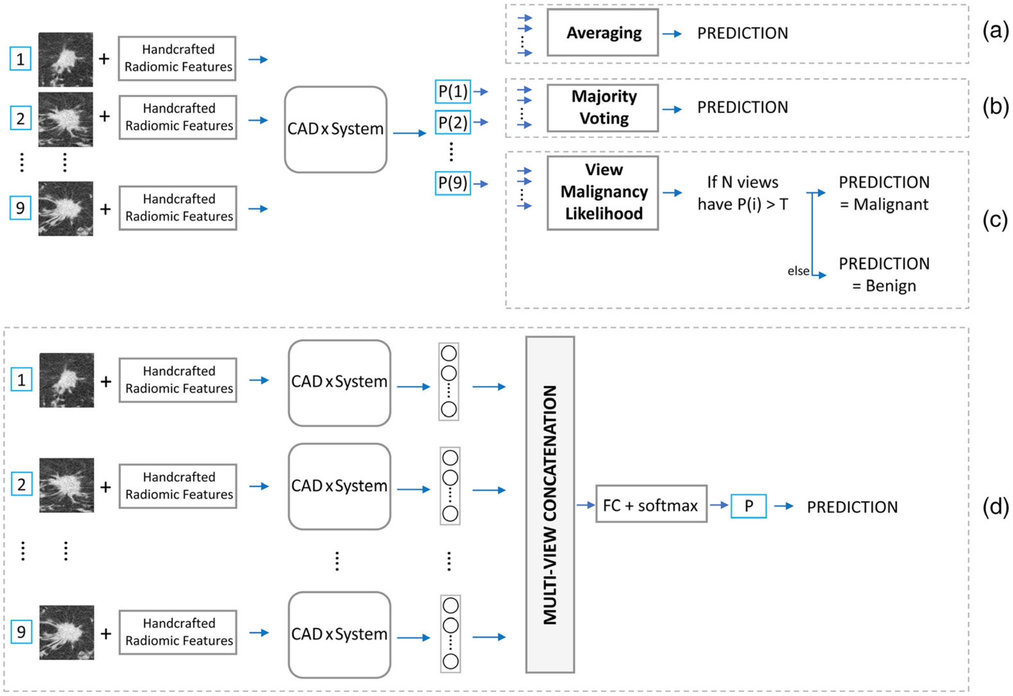

Fig. 5 Schematics of the multi-view output fusion strategies, implemented to merge the 9 predicted probabilities from each view of the mass into a single malignancy score: (a) averaging; (b) majority voting; (c) view malignancy likelihood; (d) multi-view concatenation.

with N ranging between 1 and 9 views for each T investigated. In each condition, the CADx performance was calculated in terms of sensitivity (or true positive rate, TPR) and false positive rate (FPR), and the final optimal pair (T, N) was selected to provide the highest sensitivity associated with the lowest FPR. Further details about this method are reported in Appendix.

Lastly, a fully supervised architecture (multi-view concatenation) was implemented to automatically provide a unique predicted probability directly from the nine input mass views [Fig. 5(d)]. For this, a nine-stream network architecture was implemented, with each CADx stream processing each of the mass views separately. The output merging is performed by concatenating the last FC layers (five neurons each) from each stream, and by connecting them to a final layer with a softmax activation function for outcome prediction. This model was re-trained as explained in Sec. 2.5, with a further data augmentation strategy to account for the larger number of parameters to be adjusted during learning. Given that this multi-input model is nine times bigger than the single CADx patch-based model, and given the isotropic voxel size of the bCT images, training data were further augmented by inputting each of the nine streams with all patches deriving from all nine views. Specifically, for each training mass, the index of each of the nine views was shifted by one for nine times, resulting in nine groups of nine patches each (extracted from the same mass) that are ordered differently for each group. All groups are then used as training examples for the multi-view concatenation model, globally resulting in the training data augmented by a factor of nine.

2.8 *Performance Evaluation*

All developed models were evaluated using the masses from the test set, after extracting the nine patches from each case. Automatic mass segmentation performed by the U-Net (Sec. 2.3) was quantitatively compared against the ground truth manual annotation on all nine views extracted from the test set masses, by calculating the Dice similarity coefficient (DSC), sensitivity and precision metrics. To compare the performance of the patch-based CADx system (explained in Sec. 2.5) against the four more traditional architectures (as explained in Sec. 2.6 and Fig. 4), the $F1$ score and the area under the receiver operating characteristics (ROC) curve (AUC) were used. For this comparison, all nine patches extracted from each mass of the test set were used, and the 95% confidence interval (C.I.) was calculated for each AUC value using bootstrapping with 1000 bootstraps.[50] The difference between the ROC curve of the CADx system and the other four model architectures was also analysed statistically using the Obuchowski non-parametric method.[51] This statistical method was used because it allows for ROC comparisons in datasets containing clusters of inter-correlated data (i.e., the nine patches extracted from each mass). All four multi-view output fusion strategies (Sec. 2.7 and Fig. 5) were compared at two operating points (sensitivity 95% and 90%). This method for comparison was used because the view malignancy likelihood strategy [Fig. 5(c)] works for a defined number of views and a probability threshold, and therefore does not allow for ROC analysis.

Finally, the best-performing model (in terms of the multi-view output fusion approach) was compared to three board-certified breast radiologists, who were asked to provide a malignancy grade for each mass of the test set. For a fair comparison, the radiologists were given the same information as the CADx system, i.e., only the nine views extracted from each mass of the test set. For each set of nine views, the radiologists were asked to provide a number between 1 (definitely benign) and 10 (definitely malignant). Each radiologist's performance, in terms of ROC curve, was used for comparison on a per-mass basis. The 95% CI on the AUC was calculated for each ROC curve (bootstrapping, 1000 bootstraps), and the significance of the difference in AUC values between the radiologists and the CADx system was statistically evaluated with the method proposed by Delong et al.[52] for ROC comparisons.

3 Results

Automatic segmentation achieved a DSC of 0.91 ± 0.05 for the benign and 0.88 ± 0.07 for the malignant masses (Table 2). Some examples of manual and automatic segmentation are shown in Fig. 6.

Table 2 Performance of the U-Net (mean and standard deviation) in the automatic segmentation of the test set masses, expressed in terms of DSC, sensitivity, and precision.

	DSC	Sensitivity	Precision
Benign masses, $n = 405$ patches (45 masses)	0.91 (0.05)	0.95 (0.05)	0.88 (0.09)
Malignant masses, $n = 333$ patches (37 masses)	0.88 (0.07)	0.87 (0.10)	0.91 (0.09)

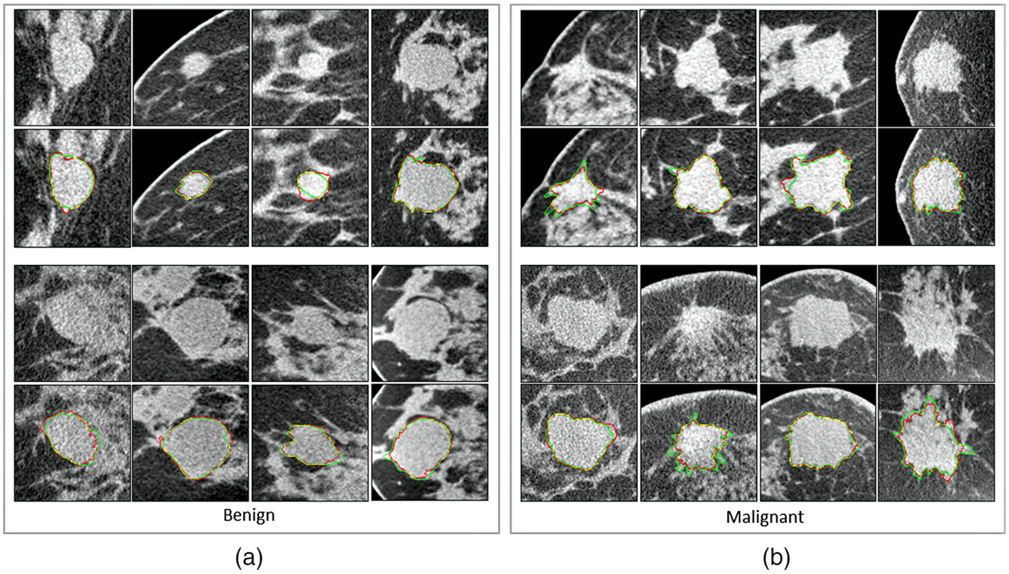

Fig. 6 Examples of (a) benign and (b) malignant mass patches, respective manual ground truth annotations (green contours), and automatic U-Net-based segmentation (red contours). Yellow parts of the contour indicate a perfect overlay between automatic and manual segmentation.

Results of the comparison of the CADx system with the other four additional architectures (calculated on a patch-basis on all mass views extracted from the test set masses) are shown in Fig. 7 and Table 3. Overall, the CADx system performed better than the other models, with an AUC of 0.876. When the two blocks were treated separately [Figs. 4(b)–4(c)], the CNN outperformed the MLP (AUC of 0.835 and 0.794, respectively). The models with decision-level and feature-level fusion [Figs. 4(d)–4(e)] achieved similar performance to the CNN alone (AUC of 0.836 for the decision-level fusion, 0.843 for the feature-level fusion approach).

Regarding the multi-view output fusion strategies performed on the CADx system to merge the 9 patch-based predicted probabilities (Table 4), the multi-view concatenation outperformed the other models, with an FPR of 22% and F1 score of 82%, for a sensitivity of 95%. The malignancy likelihood model reached the highest performance when only one view with a high associated malignancy probability (T = 0.9) was considered to classify a mass as malignant (further details are provided in the Appendix), with an FPR of 31% and $F1$ score of 74%, for a sensitivity of 95%. The fusion approach based on probability averaging led to similar results (FPR 29% and $F1$ score of 76%, for a 95% sensitivity), whereas the majority voting performed worse (FPR of 78% and $F1$ score of 69% for a 95% sensitivity).

Finally, when compared to the three radiologists, the best performing CADx model (multi-view concatenation) performed better in the classification of the 82 test set masses (Fig. 8, Table 5), with an AUC of 0.947, against an average radiologist AUC of 0.859 ± 0.044.

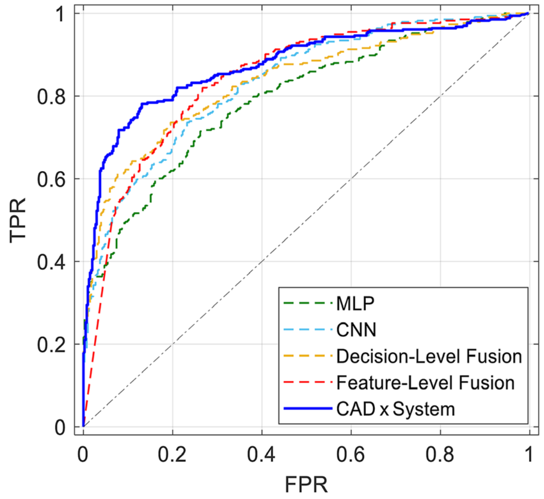

Fig. 7 ROC curve showing the performance of the proposed CADx system and the additional four architectures evaluated for comparison. The five different model architectures were tested on a mass patch-basis ($n = 405$ benign patches from 45 masses, and $n = 333$ malignant patches from 37 masses).

Table 3 Performance of the CADx system, compared with the performance of the additional architectures evaluated, calculated on all test set mass patches (from 45 benign and 37 malignant masses). The *p*-values reported refer to the statistical comparison of the ROC curve of the CADx system to that of each other model architecture evaluated using the Obuchowski non-parametric method.

	AUC	95% CI	*P*-value	*F*1 score
CADx system	0.876	0.846 to 0.90	NA	0.820
MLP	0.794	0.759 to 0.825	0.002	0.702
CNN	0.835	0.801 to 0.861	0.079	0.738
Decision-level fusion	0.836	0.804 to 0.864	0.083	0.754
Feature-level fusion	0.843	0.813 to 0.870	0.218	0.758

Table 4 Results of the different multi-view output fusion strategies performed on the CADx system to merge the nine probabilities extracted from each of the nine mass views using two operating points with high sensitivity.

	Sensitivity (%)	FPR (%)	*F*1 score (%)
Averaging	95	29	76
	90	27	83
Majority voting	95	78	69
	90	57	79
Malignancy likelihood	95	31	74
	90	20	82
Multi-view concatenation	95	22	82
	90	20	87

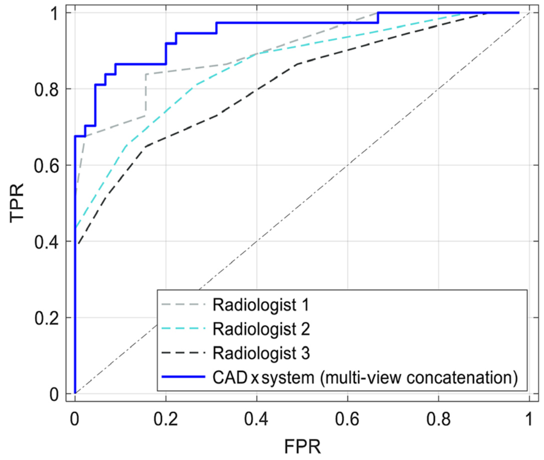

Fig. 8 Results of the best-performing CADx system (with output fusion through multi-view concatenation) and three board-certified breast radiologists in the classification of the 82 test set breast masses.

Table 5 Results of the comparison between the best-performing, mass-based CADx system (with output fusion through multi-view concatenation) and three board-certified breast radiologists in the classification of the 82 test set breast masses. The *p*-values reported refer to the statistical comparison of the performance of the system to that of each radiologist.

	AUC	95% CI	*P*-value
CADx (multi-view concatenation)	0.947	0.881 to 0.980	NA
Radiologist 1	0.902	0.821 to 0.953	0.27
Radiologist 2	0.861	0.758 to 0.926	0.07
Radiologist 3	0.814	0.699 to 0.891	0.01

4 Discussion

In this study, a computerized system for the diagnosis of masses acquired with independent dedicated bCT devices was developed and validated. The system works by merging the probabilities predicted from multiple 2D patches, extracted from a single 3D mass along multiple planar orientations. It is based on both handcrafted and convolutional radiomic features, which are embedded into a single multi-view architecture, processed, and finally merged to provide the mass-based classification outcome.

When the computerized diagnostic task was performed either by only using a CNN, or using automatic mass segmentation and a handcrafted feature-based MLP model, the CNN outperformed the MLP. This finding was consistent with the results from previously reported studies that demonstrated the higher performance of CNNs over approaches based on handcrafted descriptors, due to the CNN architecture's ability to automatically extract the most appropriate feature set from an image, and use them for an optimized learning.[15] Furthermore, although convolutional models are usually more complex than shallow networks (and therefore require further hyperparameter tuning and optimization), their use obviates the need for the design, extraction, and selection of radiomic features. Moreover, CNNs can be used to extract relevant features directly from the input images, avoiding the segmentation step. This could be beneficial to avoid any potential bias in the radiomic analysis introduced by specific segmentation results, deriving from either expert human readers or specific algorithms. Although this bias can be reduced by discarding the features that are sensitive to intra-case variations in mass contour delineation,[53] as performed in this study, a poor segmentation could still lead to a suboptimal mass characterization, thereby negatively affecting the predicted outcome.

In our previous work,[29] an AUC in breast mass classification of 0.90 was achieved using only handcrafted radiomic features (combined with principal component analysis) deriving from manual segmentation. To evaluate the effect of automatic segmentation on the predicted diagnostic performance in our handcrafted pipeline, the MLP model of this work was tested on a mass-basis (using the average operator as output fusion approach, as performed in Ref. 29), and an AUC of 0.87 was obtained. This confirmed, on the one hand, the strong effect of segmentation on clinical outcome prediction in radiomic analyses. On the other hand, results from this present study obtained with the CADx system (AUC 0.95), and based on automatic segmentation, pointed to the possibility to lower the segmentation effect by increasing the complexity of the classification model. Therefore, while further research is needed to better investigate these results with larger datasets, hybrid approaches based on both handcrafted and deep learning radiomics seem to have potential to improve radiomic pipelines.

When the convolutional and the handcrafted approaches were combined at an output-level (i.e., when the predicted probability determined from the CNN was averaged with the predicted probability from the MLP) or combined at a feature-level (i.e., when convolutional and handcrafted descriptors were concatenated and fed to an independent late-stage classifier), the AUC changed slightly compared to when using only the CNN. Instead, the proposed CADx system performed better than all four of the other models evaluated. This finding suggests that embedding the two approaches in the same architecture could provide advantages compared to treating the two branches separately, or merging the feature vectors at a late stage. These advantages could be due to the learning occurring in parallel for both branches, which facilitated the optimization of all network parameters, accounting for both the handcrafted and the convolutional inputs simultaneously. However, comparisons among the different investigated architectures should be further evaluated in future, with a larger test set, to assess whether the proposed CADx architecture can achieve a statistically significant superior performance. Furthermore, with a larger test set, comparisons of the developed multi-view architectures against radiomic analyses based on 3D CNNs should also be assessed. This may result in an increase in diagnostic performance, thanks to leveraging the fully tomographic nature of bCT images, and due to the advantage of avoiding any output fusion strategy.

Regarding the output fusion methods, the multi-view concatenation approach provided the best per-mass classification performance, likely attributed to its ability to implicitly account for 3D information by incorporating multiple mass views into a single deep learning model. In fact, the multi-view concatenation approach merged the patch-based output of each prediction stream into a common classification layer. This strategy therefore allowed for the ability to better integrate the 3D characteristics of the mass by comparing the output of multiple networks in parallel, in a configuration where the parameters of the layers from multiple streams are shared. The other output fusion strategies considered, instead, the input 3D mass as a stack of independent patches, potentially reducing the amount of 3D information retrieved from multiple views. However, as previously mentioned, approaches that use the full volumetric input information (i.e., not patched-based) may also be worth considering in the future to evaluate the potential gain in performance prediction.

Of the other output fusion strategies, the majority voting resulted in the highest FPR, for the same values of sensitivity used to test the performance of all models. This can be attributed to the fact that dichotomizing the nine view probabilities prior to any fusion operations results in a loss of information, as it does not leverage the continuous nature of the estimated malignancy probability of each of the nine views.

Regarding the comparison of the final mass-based CADx system to the performance of breast radiologists, the system performed better (although statistical significance was reached only in one case), indicating its potential for automated breast mass diagnosis that outperforms visual perception. However, for these comparisons, the radiologists were provided the same information as the CADx system, i.e., a set of nine views for each test set mass. While, in a real clinical setting, radiologists perform diagnostic decisions on a full image-basis (in addition to any other relevant clinical information), this strategy was adopted for a fair comparison between the CADx system and expert human visual perception. However, such an analysis should be repeated in future studies, possibly with a larger cohort of patient data sets, by evaluating the possible human performance improvements when the radiologists are provided with the entire bCT volume data set.

While encouraging, the results reported herein were derived from a still limited test set (a fact that, at the moment, greatly limits any relevant, clinically applicable conclusion), and should therefore be confirmed in future studies using a much larger number of test cases. Among the major limitations, the relatively limited available dataset size is to be acknowledged. While clinical trials on dedicated bCT are currently being performed only in a few research centers around the world, the inclusion of additional testing masses could provide further insights about the performance of the developed CADx system, by allowing for improved statistics for each ROC comparison. Moreover, only mass-like lesions were included in this study; while bCT has been shown to provide a greater benefit in mass visualization compared to calcifications[14] (in comparison with mammography), the latter will be included in future work to develop a complete diagnostic system.

Regarding the architecture of the proposed CADx system, future work could also include the embedding of the U-Net for automatic segmentation as an extra channel to the CNN, allowing the CNN to potentially take advantage of automatic segmentation. This might further increase the classification performance, since the CNN would be fed also with the segmented mass patches, in addition to the raw ones. The potential benefit of this additional strategy, compared to the one proposed, should be properly assessed in the future with a much larger cohort of patient data sets, where possible differences in performance can be perceivable.

Finally, future work involves the evaluation and assessment of a fully 3D-based CADx system, with both handcrafted radiomic features and CNNs implemented to process 3D reconstructed volumes. While the 2D-based approach presented in this study inherently increased the dataset size, through the extraction of nine views from each mass, and 3D information was implicitly obtained through the multi-view strategy, a future fully 3D approach might help achieve further insights in CADx of breast masses, especially if larger image datasets become available.

5 Conclusions

A CADx system for images of breast masses acquired with dedicated bCT was developed, which incorporates handcrafted and convolutional radiomic features embedded into a single deep learning model. The system retrieves and concatenates the information acquired from multiple 2D mass views to estimate a single per-mass prediction value, in a configuration where learning occurs simultaneously for each view. The proposed method outperformed other machine and deep learning architectures adopting handcrafted and convolutional approaches, and resulted in a better performance prediction compared to the visual assessment of three board-certified breast radiologists.

While future studies with larger image datasets are needed to confirm the performance of the proposed approach, results demonstrated its potential usefulness for breast cancer diagnosis by improving mass malignancy assessment. Furthermore, our findings point to the potential benefit of combining engineered and convolutional radiomic approaches over using single independent architectures, especially for image datasets still limited in size where the learning of convolutional models may be suboptimal.

6 Appendix

6.1 *A.1. Comparison with Additional Architectures*

To compare the performance of the proposed CADx system over more traditional architectures, 4 additional models were developed (Fig. 4).

First, the performance resulting from using only either the MLP or the CNN branch separately was investigated [Figs. 4(b) and 4(c)]. For this, the two branches were simply disconnected from the CADx architecture, and a final sigmoid-based unit was added to the output of each model to perform the binary classification. While the architectures and learning algorithms were not modified compared to when training the CADx system, the learning rate needed to be fine-tuned using the validation set for the MLP, resulting in the best value being 10^{-4}.

A decision-level model [Fig. 4(d)] was developed by considering the MLP and CNN blocks described above separately, but the final classification score was provided based on the average of the two outputs together (one from the CNN, one from the MLP). This was done to assess whether the two radiomic strategies (handcrafted and convolutional) could be successfully combined only at an output level, and therefore evaluate if the two networks could be treated and trained separately (with the advantage of a reduction in training cost) and still benefit from a simple output probability merging.

The last model [feature-level, Fig. 4(e)] consists of the CNN applied to the image patches as a feature extractor, and the extracted convolutional features concatenated at a late stage with the respective handcrafted features from the same image patch. The resulting concatenated feature vector was then fed to a late-stage classifier for final prediction. For this feature-level model, the last two layers were removed from the CNN architecture described above, and the CNN was applied to extract 64 deep features from the training set mass patches, which were concatenated with the respective 36 handcrafted features. The derived feature vectors were used to train a late-stage model for diagnostic outcome prediction. For this late-stage classifier, the same architecture of the MLP described above and detailed in the main manuscript (Sec. 2.5) was used, and fine tuning of the hyperparameters was performed on the validation set (Adam optimizer, learning rate 10^{-4}).

6.2 *A.2. Multi-View Output Fusion Strategies—View Malignancy Likelihood*

This output fusion approach aims to combine the malignancy likelihood predicted by the CADx system from the nine views extracted from each mass, so as to provide a single classification outcome on a mass-basis. Specifically, this method aims to find the optimal number of views identified as malignant (N) sufficient to correctly consider the whole mass as malignant, and is motivated by the fact that only a few views with a high malignancy probability (T) could be sufficient to correctly classify the whole mass as malignant, regardless the predicted probabilities of the other remaining views. To efficiently implement this approach as an output fusion strategy for the nine view-based computerized diagnostic system, several combinations of number of considered views (N) and probability threshold used to classify the views as malignant (T) were investigated, and the classification performance was reported for each pair. Since a given pair of parameters N and T corresponds to a single ROC operating point, each classification performance is provided in TPR and FPR.

The experiment was repeated for five different values of T (0.1, 0.25, 0.5, 0.75, and 0.9), and with N ranging between 1 and 9 for each T investigated. The final optimal pair (T, N) was selected to provide the highest TPR associated with the lowest FPR.

Figure 9 shows the results of this analysis. As expected, the sensitivity was inversely proportional to the number of considered views N (and the same was for the FPR). The best results were achieved when a single view with a high predicted malignancy probability ($T > 0.9$) was considered to classify the whole mass as malignant, regardless the predicted probability of the other views. Results confirmed the hypothesis that considering only a low number of views with a high

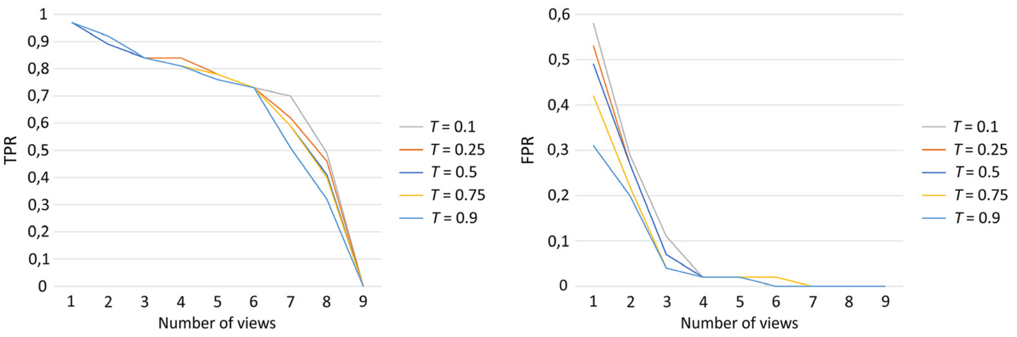

Fig. 9 Results of the view malignancy likelihood analysis.

predicted malignancy probability is sufficient to correctly classify a breast mass as malignant, reaching a sensitivity of 95% (for $N = 1$ view) with respective FPR of 31%.

Disclosures

Authors John M. Boone and Andrew M. Hernandez have patents (pending and issued) pertaining to bCT. Author John M. Boone has prior research support, licensing agreements with, and is a shareholder of Izotropic Imaging Corp. of Canada.

Acknowledgments

This research was supported in part by research under grants Nos. R01CA181171 and R01CA181081 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute, the National Institutes of Health. The authors are grateful to Susanne Lardenoije-Broker and Mechli Imhof-Tas for the help provided in the radiologist ROC study.

References

1. R. J. Gillies, et al., "Radiomics: images are more than pictures, they are data," *Radiology* **278**(2), 563–577 (2013).
2. O. Ozdemir, R. L. Russell, and A. A. Berlin, "A 3D probabilistic deep learning system for detection and diagnosis of lung cancer using low-dose CT scans," *IEEE Trans. Med. Imaging* **39**, 1419–1429 (2019).
3. S. Hussein, et al., "Lung and pancreatic tumor characterization in the deep learning era: novel supervised and unsupervised learning approaches," *IEEE Trans Med Imaging* **38**(8), 1777–1787 (2019).
4. L. Liu, et al., "Multi-task deep model with margin ranking loss for lung nodule analysis," *IEEE Trans Med Imaging* **39**(3), 718–728 (2020).
5. R. Wilson and A. Devaraj, "Radiomics of pulmonary nodules and lung cancer," *Transl. Lung Cancer Res.* **6**(1), 86–91 (2017).
6. W. Dai et al., "Prognostic and predictive value of radiomics signatures in stage I–III colon cancer," *Clin. Transl. Med.* **10**(1), 288–293 (2020).
7. A. Ashraf, et al., "Breast DCE-MRI kinetic heterogeneity tumor markers: preliminary associations with neoadjuvant chemotherapy response," *Transl. Oncol.* **8**(3), 154–162 (2015).
8. J. Ferlay et al., "Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012," *Eur. J. Cancer* **49**, 1374–1403 (2013).
9. M. Heidari et al., "Development and assessment of a new global mammographic image feature analysis scheme to predict likelihood of malignant cases," *IEEE Trans Med Imaging* **39**(4), 1235–1244 (2020).
10. M. U. Dalmis et al., "Artificial intelligence-based classification of breast lesions imaged with a multiparametric breast MRI protocol with ultrafast DCE-MRI, T2, and DWI," *Invest Radiol* **54**(6), 325–332 (2019).
11. A. S. Tagliafico et al., "Breast cancer Ki-67 expression prediction by digital breast tomosynthesis radiomics features," *Eur. Rad. Expr.* **3**, 36 (2019).
12. A. O'Connel, et al., "Cone-beam CT for breast imaging: radiation dose, breast coverage, and image quality," *Am J Roentgenol* **195**(2), 496–509 (2010).
13. A. M. Hernandez et al., "Effects of kV, filtration, dose and object size on soft tissue and iodine contrast in dedicated breast CT," *Med Phys* **47**, 2869–2880 (2020).
14. K. K. Lindfors et al., "Dedicated breast CT: initial clinical experience," *Radiology* **246**(3), 725–733 (2008).
15. P. Afshar et al., "From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities," *IEEE Signal Proc. Mag.* **36**(4), 132–160 (2019).
16. N. Lessmann et al., "Automatic calcium scoring in low-dose chest CT using deep neural networks with dilated convolutions," *IEEE Trans. Med. Imaging* **37**(2), 615–625 (2018).

17. T. Kooi et al., "Large scale deep learning for computer aided detection of mammographic lesions," *Med. Image Anal.* **35**, 303–312 (2017).
18. R. Paul et al., "Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma," *Tomography* **2**(4), 388–395 (2016).
19. J. Lao et al., "A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme," *Sci. Rep.* **7**(1), 10353 (2017).
20. S. Chen et al., "Automatic scoring of multiple semantic attributes with multi-task feature leverage: a study on pulmonary nodules in CT images," *IEEE Trans. Med. Imaging* **36**(3), 802–814 (2017).
21. P. Prasanna et al., "Radiomics-based convolutional neural network for brain tumor segmentation on multiparametric magnetic resonance imaging," *J. Med. Imaging* **6**(2), 024005 (2019).
22. L. Chen et al., "Combining many-objective radiomics and 3D convolutional neural network through evidential reasoning to predict lymph node metastasis in head and neck cancer," *Phys. Med. Biol.* **64**(7), 075011 (2019).
23. N. Antropova et al., "A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets," *Int. J. Med. Phys. Res. Pract.* **44**(10), 5162–5171 (2017).
24. B. Huynh et al., "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," *J. Med. Imaging* **3**(3), 034501 (2016).
25. S. Liu et al., "Pulmonary nodule classification in lung cancer screening with three-dimensional convolutional neural networks," *J. Med. Imaging* **4**(4), 041408 (2017).
26. J. M. Boone et al., "An x-ray computed tomography/positron emission tomography system designed specifically for breast imaging," *Technol. Cancer Res. Treat.* **9**(1), 29–43 (2010).
27. P. Ghazi et al., "Shading artifact correction in breast CT using an interleaved deep learning segmentation and maximum likelihood polynomial fitting approach," *Med. Phys.* **46**(8), 3414–3430 (2019).
28. I. Sechopoulos, S. S. Feng, and C. J. D'Orsi, "Dosimetric characterization of a dedicated breast computed tomography clinical prototype," *Med. Phys.* **37**(8), 4110–4120 (2010).
29. M. Caballo et al., "Multi-marker quantitative radiomics for mass characterization in dedicated breast CT imaging," *Med. Phys.* **48**, 313–328 (2020).
30. A. A. A. Setio et al., "Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks," *IEEE Trans. Med. Imaging* **35**(5), 1160–1169 (2016).
31. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
32. N. Srivastava et al., "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014).
33. D. P. Kingma and J. L. Ba, "Adam: a method for stochastic optimization," arXiv:1412.6980 (2014).
34. M. Caballo et al., "Breast parenchyma analysis and classification for breast masses detection using texture feature descriptors and neural networks in dedicated breast CT images," *Proc SPIE* **10950**, 09500J1 (2019).
35. R. M. Haralick et al., "Textural features for image classification," *IEEE Trans. Syst. Man Cybern.* **SMC-3**(6), 610–621 (1973).
36. M. M. Galloway, "Texture analysis using gray level run lengths," *Comput. Graphics Image Process.* **4**, 172–179 (1975).
37. T. Ojala et al., "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 971–987 (2002).
38. K. I. Laws, "Rapid texture identification," *Proc. SPIE* **0238**, 376–381 (1980).
39. Y. Iwahoria et al., "Automatic detection of polyp using Hessian filter and HOG features," *Proc. Comput. Sci.* **60**, 730–739 (2015).
40. J. M. Keller et al., "Texture description and segmentation through fractal geometry," *Comput. Vision Graphics Image Process.* **45**(2), 150–166 (1989).

41. M. Haghighat et al., "CloudID: trustworthy cloud-based and cross-enterprise biometric identification," *Expert Syst. Appl.* **42**(21), 7905–7916 (2015).
42. L. Shen et al., "Application of shape analysis to mammographic calcifications," *IEEE Trans. Med. Imaging* **13**(2), 263–274 (1994).
43. J. Kilday et al., "Classifying mammographic lesions using computerized image analysis," *IEEE Trans. Med. Imaging* **12**(4), 664–669 (1993).
44. L. Gupta and M. D. Srinath, "Contour sequence moments for the classification of closed planar shapes," *Pattern Recognit.* **20**(3), 267–272 (1987).
45. Z. Huo et al., "Analysis of spiculation in the computerized classification of mammographic masses," *Med. Phys.* **22**(10), 1569–1579 (1995).
46. Z. Huo et al., "Automated computerized classification of malignant and benign masses on digitized mammograms," *Acad. Radiol.* **5**, 155–168 (1998).
47. J. Wu et al., "Intratumoral spatial heterogeneity at perfusion MR imaging predicts recurrence-free survival in locally advanced breast cancer treated with neoadjuvant chemotherapy," *Radiology* **288**(1), 26–35 (2018).
48. R. J. Urbanowicz et al., "Relief-based feature selection: introduction and review," *J. Biomed. Inf.* **85**, 189–203 (2018).
49. S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, pp. 448–456 (2015).
50. B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, p. 57, CRC Press, Boca Raton, Florida (1994)
51. N. A. Obuchowski, "Nonparametric analysis of clustered ROC curve data," *Biometrics* **53**, 567–578 (1997).
52. E. R. DeLong et al., "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics* **44**(3), 837–845 (1988).
53. M. Caballo et al., "Deep learning-based segmentation and characterization of breast masses in dedicated breast CT imaging: radiomic feature stability and diagnostic performance," *Comput. Biol. Med.* **118**, 103629 (2020).

Biographies of the authors are not available.