Health and Technology
https://doi.org/10.1007/s12553-025-01039-6

ORIGINAL ARTICLE

DBT_DCNN: a new convolutional neural network for mass detection in digital breast tomosynthesis

Gianfranco Paternò[1] · Matteo Bianchini[2] · Paolo Cardarelli[1] · Angelo Taibi[1,2] · Roberta Ricciardi[3,4] · Paolo Russo[3,4] · Giovanni Mettivier[3,4]

Received: 23 January 2025 / Accepted: 20 December 2025

Abstract

Purpose We developed DeepLook, a CAD based on a CNN for the analysis of Digital Breast Tomosynthesis images to classify and locate possible breast mass lesions.

Methods This CAD takes advantage of a pre-processing algorithm to remove pectoral muscle and skin, before entering the lesion classification step and the use of a CNN architecture. Evaluation of the architecture performance was realized using a dataset coming from Duke University, one of the largest databases publicly available online.

Results DeepLook can classify individual DBT images into two classes: healthy (Negative) and presence of a mass (Positive) or in three classes: healthy, presence of a malignant or benignant masses. Accuracy and sensitivity on the binary classification of test data of the most suitable considered dataset reached 82% and 68%, respectively, with an AUC of 0.90.

Conclusions We developed DeepLook, a CAD based on a CNN for the analysis of DBT images to classify and locate possible breast mass lesions. The reported performance was obtained with a CNN architecture that is significantly simpler than the ones proposed in the literature so far and is promising for future clinical implementation in routine DBT diagnosis, so as to suggest to radiologists the slices that are worth paying attention to better. As a further step, a statistical study of the mass locations in images generated through the GradCam algorithm is under development.

Keywords Breast cancer · Digital Breast Tomosynthesis · Convolutional Neural Networks · Deep Learning · Computed Aided Diagnosis · Image processing

1 Introduction

Breast cancer is a type of cancer where the cells in the breast grow abnormally. There are two types of breast cancer: benign and malignant tumors. Benign tumors are not cancerous, but they can increase the risk of breast cancer as they consist of out-of-control cells that cause a lump. Malignant tumors, on the other hand, grow rapidly and can spread to other parts of the body. Breast cancer affects both genders, but it is more common in women and is the second-largest cancer worldwide. About one in eight women in USA will be diagnosed with it at some point in her life [1]. All worldwide health organizations recommend screening x-ray mammography as the gold standard exam to achieve earlier breast cancer detection to decrease breast cancer mortality (about 20% [2]). However, since mammography is a 2D imaging modality, it results in the projection of the internal tissue of the breast onto a single plane, yielding tissue superposition. This may lead to a moderate sensitivity and specificity, especially with dense breasts. The limitations of mammography screening also include over-diagnosis [3], overtreatment and false-positive rates associated with negative psychological impact [4, 5] and unnecessary costs and biopsies [6]. Interpretation of mammograms varies with experience [7], is subjective [8], and prone to error due to the heterogeneous presentation of breast cancer and the masking effect with dense breast tissue.

✉ Giovanni Mettivier
mettivier@na.infn.it

[1] INFN Sezione di Ferrara, Ferrara I-44122, Italy

[2] Dipartimento di Fisica e Scienze della Terra, Università degli Studi di Ferrara, Ferrara I-44122, Italy

[3] Dipartimento di Fisica "Ettore Pancini", Università di Napoli Federico II, Napoli I-80126, Italy

[4] INFN Sezione di Napoli, Napoli I-80126, Italy

Published online: 19 January 2026

In the last years, new techniques were introduced on the market in order to improve the detection rate and reduce recall. Among these techniques, we have digital breast tomosynthesis (DBT) [9] and computed tomography dedicated to the breast (BCT) [10–12]. Digital breast tomosynthesis is a three-dimensional imaging technique for breast cancer screening that instead of projection images delivers multiple cross-sectional slices for each breast and offers better performance [9]. The reduction of tissue overlap provides increased lesion conspicuity particularly in dense breasts compared to full-field digital mammography (DM) and has been shown to improve the detection sensitivity of invasive breast cancer while reducing the recall rate. On the other hand, radiologist interpretation time of DBT studies takes more time, in previous studies, for two-view DBT it was 38–76% longer than for DM [13–17], limiting the ability to screen as many patients as possible due to reader fatigue. Apart from a potential increase in the radiation dose, this represents the major limiting factor for the widespread of DBT modality among the recommended examinations in breast cancer screening programs.

In recent years, Deep Learning (DL) and Convolutional Neural Networks (CNNs) have been used in many applications for the segmentation and classification of images, including the medical field. These kinds of architectures can automatically extract some meaningful descriptors from the input images, thus eliminating the need of using codes for the extraction of hand-crafted features necessary to traditional feature-based classifiers.

Deep-learning CAD systems have proven to be useful in reducing reading times of DM without significantly affecting sensitivity, specificity, producing reductions in reading times from 14 to 40% [18–20]. If AI could ease the burden of reading DBT, this could open possibilities for the broad introduction of DBT in population-based screening programs.

A few studies have reported successful application of Deep Convolutional Neural Networks (DCNN) in the detection and classification of masses in DBT. For instance, Samala et al. [21], Fotin et al. [22], Kim et al. [23, 24] have all demonstrated positive outcomes. Samala et al. [21] developed a serve-layer DCNN, which yielded more accurate DBT mass detection and maintained a similar level of false positive rate when compared to a featured-based CAD system. Other researchers came to similar conclusions by using various DCNN architectures [22] or by manipulating DBT image information from bilateral or depth directions [23].

The aim of this study is to present DeepLook, a DL-based CAD developed in Python, which aims to help the radiologists in the reading of DBT examinations. DeepLook is conceived as an ensemble of tools to perform dedicated tasks. The core module is based on a CNN, which is a refinement on the one presented in [25]. The CNN performs classification of the individual slices of each DBT volume analysed, highlighting those in which there is high probability to find a mass. Most of the studies carried out regarded a binary classification task, namely aimed at identifying slides where at least a mass (malignant or benign) was present (Positive cases), but also the three classes problem was investigated, and the result here presented. Care was spent on the selection of the training and test datasets and a series of actions were taken to avoid overfitting.

2 Materials and methods

2.1 Dataset

The dataset coming from Duke University, that we will refer to as “Duke”, is one of the largest databases publicly available online [26, 27]. It is composed of 13,954 patients out of which a subset of 5610 studies from 5060 patients were annotated by two radiologists; furthermore, about 200 volumes underwent biopsy. DBT acquisitions were performed with a Hologic scanner, in one or more of the following views, left craniocaudal (LCC), right craniocaudal (RCC), left mediolateral oblique (LMLO), right mediolateral oblique (RMLO). Each case comes with information about bounding of lesions boxes for volumes containing any. For this work, only 237 volumes were used: 117 Negative volumes (slices with no masses), and 120 biopsy-validated Positive ones, coming from 103 patients. In fact, the vast majority of the 5610 studies is composed by healthy patients, and using all of them would have led to appreciable class imbalance. Furthermore, the preprocessing step (described later) would have been very time consuming. For these reasons, the number of volumes was limited to 237. It is to be noted that the positive, or tumoral, class contains both malignant and benign tumors, according to biopsy results. Our main goal is to train a neural network to recognize lesions in the same way a radiologist would perform the first volume evaluation step, hence for a first analysis no further splitting of positive class was carried out. The latter was instead performed for a subsequent study regarding the ability of our model to recognize the lesion type, namely for a three classes classification task.

2.2 Slice selection procedure

Due to the nature of DBT modality, contiguous slices of a given DBT volume are very similar to each other. From preliminary tests carried out with standard CNNs we noticed that this fact leads with a high probability to an overfitted model. To avoid this issue and let our DL algorithm gain

generality, we thus decided to consider a limited number of images for each DBT volume. A decimation procedure was implemented to select on average 20 slices equally spaced from each DBT volume. More in particular, five different reduced datasets for training (about 80%) and validation/test (about 20%) were assembled, in order to perform a 5-fold cross-validation during the training of our models. For each reduced dataset, it was ensured that each DBT volume was present in training subset or, alternatively, in the validation/test one. This methodological choice is particularly due to the intention to strengthen the prediction reliability, allowing to conduct tests on volumes of patients not used for training. To compensate for the reduction of the image number due to the decimation, a data augmentation procedure was applied to increment the training subset by 100%. Three of the following transformations, image cropping, flipping, rotation, and scaling, were randomly chosen for each image. The rotation angle (deg) and scaling value were set by randomly choosing them from the sets (−30, −15, −10, 10, 15, 30) and (0.05, 0.04, 0.03, 0.02), respectively. The whole procedure of merging, decimation, subdivision, and augmentation was automatically carried out through a series of dedicated Python scripts, which allowed us to be sure to fulfil the mentioned criteria.

2.3 DeepLook

2.3.1 Pre-processing procedure

All the images used for training/test purpose were pre-processed taking advantage of the tool described in detail in [28]. A schema of the procedure was reported in Fig. 1. In short, the following transformations were applied to all images. First, the contrast was increased, through grey-level histogram stretching of an automatically selected patch inside the breast region, so as to emphasize the glandular tissue and any tumour masses present. Then, the breast skin and nipple were removed from all images through the subtraction of the dilated breast contour (identified through a Canny edge detection filter [29]) from the previous contrast enhanced images. After that, a bilateral filter [30], which replaces the intensity of each pixel with a weighted average of the intensity values of neighbouring pixels, was applied. This reduced the quantum noise of the images but was also useful to homogenize the texture of the pectoral muscle area and preserve its contour in MLO views. This is indeed required for the automatically pectoral muscle removal procedure [31], which was applied to all MLO images. The removal of the breast external border, such the removal of pectoral muscle, is of fundamental importance since otherwise the breasts show bright regions that could be mistaken for lesions by the DL classification algorithm [28].

2.3.2 DBT-DCNN

The deep learning algorithm used for slice classification was based on a CNN that is an improved version of the one described in [25]. It was composed of 5 convolution layers interspersed with Batch normalization and max pooling layers, plus two fully connected output layers, as depicted in Fig. 2. Activation functions were leaky RELU with an angular coefficient for negative values alpha = 0.2.

Since the very first tests were performed with our previous model on the newly assembled datasets, we noticed overfitting issues and a difficulty to generalize. For this reason, in the definition of the new model architecture and training scheme, we adopted a series of actions aiming at reducing overfitting. First, we significantly reduced the numbers of the model parameters by reducing the number of filters, the kernel size of the convolution layers and the number of nodes of the second-last fully connected layer (dln). Then, we adopted dropout connections [32] between the fully connected layers and introduced a L2 regularization term in the loss function, which was the categorical cross-entropy [33]. These strategies act to reduce the model complexity, thus avoiding that the model captures the details

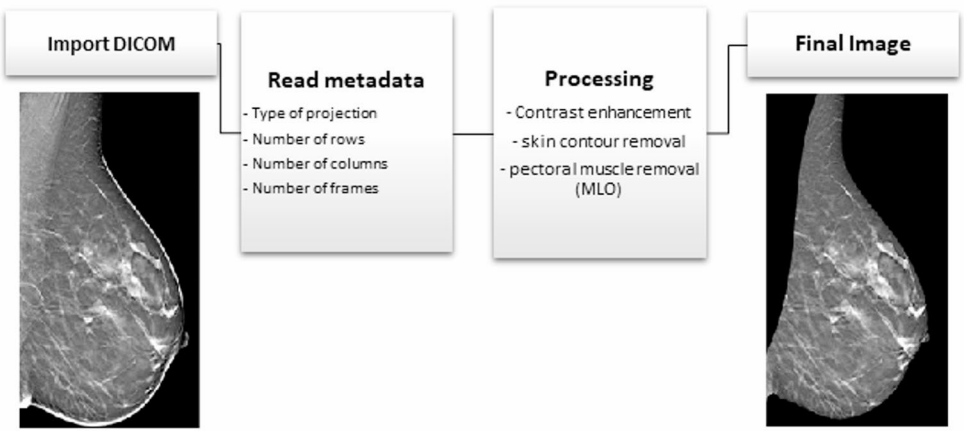

Fig. 1 Schema of the Pre-processing procedure applied on a MLO view [28]

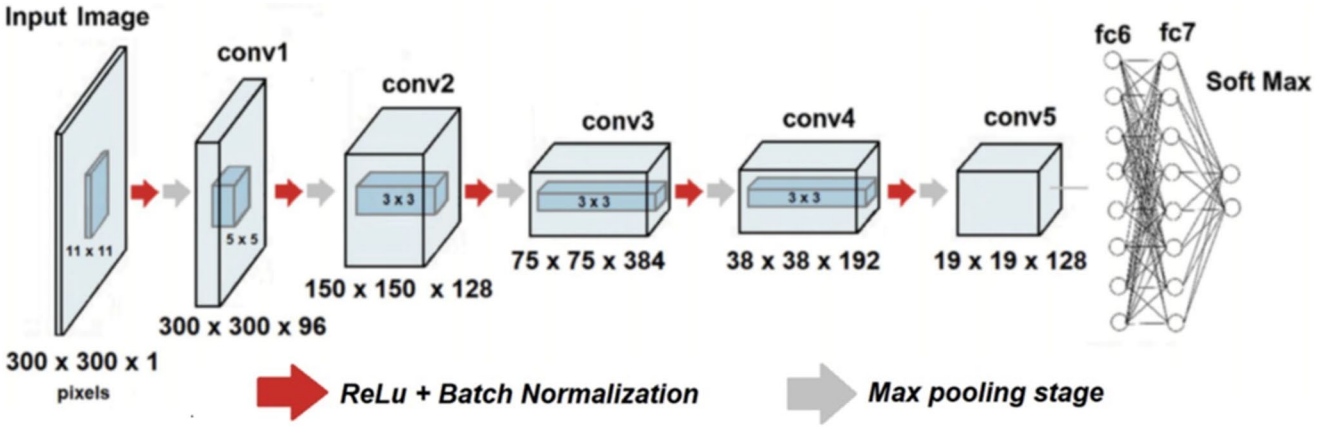

Fig. 2 Block scheme of the CNN proposed in Ricciardi et al. [25]

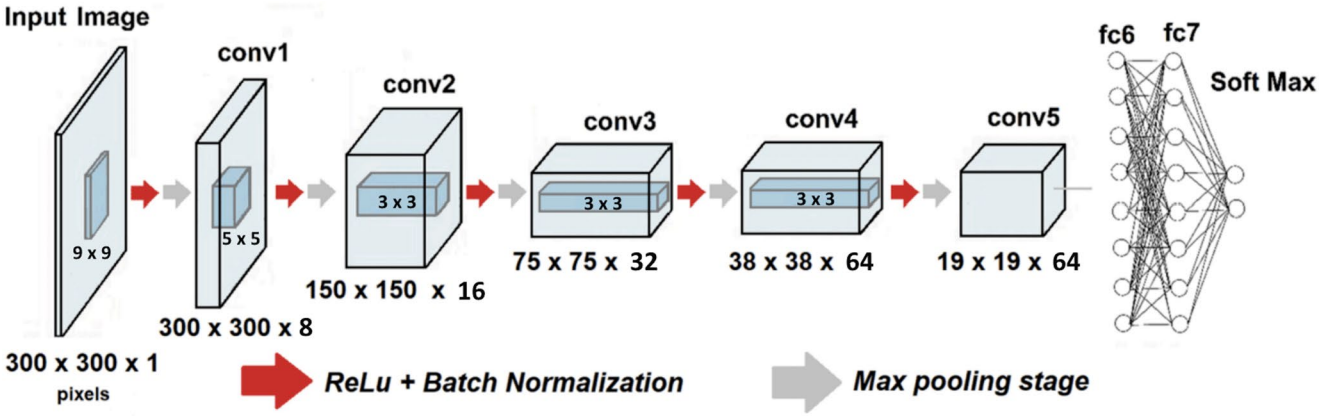

Fig. 3 Block scheme of the new DBT-DCNN algorithm

and noise in training data rather than the general trend. Indeed, if neurons are randomly dropped out of the network during training, other neurons have to change significantly (no co-adaptation) to create the representation required to make predictions. Thus, the number of independent internal representations learned by the network increases and the network itself becomes less sensitive to the specific weights of neurons and more capable of better generalization. On the other hand, if a regularization term is added to the loss function, the higher terms in a model are penalized, thus the model complexity tends to reduce during training. In L2 regularization, the penalty term is the sum of the squared weights times a multiplicative coefficient lambda.

dln, dropout and lambda represented some hyperparameters of the model. Apart from these variables, and the parameter defining the architecture of the CNN, other hyperparameters were the number of training epochs and the batch size. Through a random search, the optimal values for hyperparameters were identified. They are dln = 16, alpha = 0.2, lambda = 0.3, training epochs = 20, and batch size = 16. The block-schema and the number of filters and the kernel size of each convolutional layer of the new-optimized algorithm are shown in Fig. 3. The Adam optimizer was used with step size $\alpha = 10^{-4}$, decay rate for the momentum $\beta_1 = 0.9$ and for second moment $\beta_2 = 0.999$, which are standard values. 50% dropout was also applied to the fully-connected layers in order to improve network's sensitivity to specific weights and reduce overfitting.

Preliminary tests carried out training such CNN models using images of various size showed that convergence and results are actually independent from the shape of the slices. Therefore, to reduce computing time and improve memory usage, images were resized to 300×300 before standardization and augmentation.

The training and test of the model was carried out in a framework developed in Python, taking advantage of Keras, the high-level API (application programming interface) of the TensorFlow platform. The aforementioned framework is quite flexible and allows the user to select easily among different CNN architectures, both custom and standard, and to set all the model hyperparameters and training options, such as loss function, optimizer,

metrics, and callbacks [https://github.com/paternog/DBT_classification_tool]. Also, the code is able to automatically support classification in two (Negative, Positive) or three classes (Negative, Benign and Malign). The developed Python framework can automatically handle the model training on CPU, single GPU and multiple GPUs. For this work, we used a single GPU NVIDIA GeForce RTX 3090 card (10496 CUDA cores, 24 GB GDDR6X video memory). After the training of the model, the developed code automatically plots the learning curves the confusion matrix and calculates the significant metrics described below. Moreover, the framework (CAD) also provides, for each image, a saliency map calculated through a dedicated module based on the GradCAM algorithm [34]. This can help the radiologists the focus on small areas of the breast that the network considered most to perform the classification, and in principle where the presence of a tumour is most likely.

2.4 Evaluation metrics

The common evaluation metrics for classification are accuracy, true positive rate (TPR), true negative rate (TNR), false positive rate (FPR), false negative rate (FNR), precision, recall, F1 score, Receiver Operating Characteristic (ROC) curve, Area under ROC curve (AUC). In the following, we present a brief description of these metrics both in the binary classification and multi-class classification case.

2.4.1 Metrics in binary classification

Calling *TP* (True Positives) the number of examples correctly classified as positive ones, *FP* (False Positives) the number of examples wrongly predicted as positives, *TN* (True Negatives) the number of examples correctly classified as negatives and *FN* (False Negatives) the number of examples wrongly predicted as negatives:

Accuracy: the total number of correct predictions performed by the model on the whole dataset.

$$\frac{TP + TN}{TP + FP + TN + FN}$$

FPR and *FNR*: number of negative examples classified as positives (*False Positives*) out of the negative class, and number of positive examples identified as negatives (*False Negatives*) out of the positive class, respectively. As an example, FPR is

$$\frac{FP}{FP + TN}$$

Precision: Fraction of examples correctly classified as positives out of all the positively predicted ones.

$$\frac{TP}{TP + FP}$$

Recall (or Sensitivty or *True Positive Rate*): Fraction of examples correctly classified as positives over the whole positive class.

$$\frac{TP}{TP + FN}$$

F1-score: Harmonic mean between Precision and Recall. It more accurately describes model performances when classes are unbalanced with respect to Precision and Recall alone.

$$\frac{2 \cdot \ Precision \cdot \ Recall}{Precision + Recall}$$

Specificity (or True Negative Rate): it is the proportion of actual negative cases classified as negative

$$\frac{TN}{(TN + FP)}$$

ROC curve: The TPR plotted against the FPR (1-Specificity) at changing classification thresholds.

AUC: the area under the ROC curve. It represents the probability that the model ranks a random positive example higher than a random negative example, and the closer it gets to 1, the better are the predictions.

2.4.2 Metrics in multi-class classification

In Fig. 5 it is possible to see an example of *Confusion Matrix (CM)* for a three classes classification task.

We adopted the One-Versus-Rest (OVR) approach, so as to converts a multiclass problem into a series of binary tasks for each class in the dataset. According to this approach, for the *i*-th class, with *i* ranging from 1 to 3, True Positives value is the corresponding diagonal term, CM_{ii}, False Positives corresponds to all of the off-diagonal terms in the *i*-th column and False Negatives are all of the off-diagonal terms in the *i*-th row, and True Negatives are all the other *CM* elements [35].

Accuracy for multi-class classification has the same form of the binary case:

$$Accuracy = \sum_i \frac{TP^i + TN^i}{N}$$

where N is the number of predictions performed by the model and i ranges over all the classes.

Meanwhile, Precision and Recall can be computed singularly for each class:

$$Precision^i = \frac{TP^i}{FP^i + TP^i}$$

$$Recall^i = \frac{TP^i}{FN^i + TP^i}$$

Following this methodology, it is possible to also compute FPR, FNR.

In order to characterize the classifier with a single value for each metric M, we computed the weighted average of the metric values over all the classes M^i, taking into account the respective population N^i:

$$M = \sum_i \frac{M^i \cdot N^i}{N}$$

It is worth noting that, in the OVR approach, the weighted Recall is equal to the accuracy. For this reason, in the three classes problem, we do not explicitly report any accuracy value, but only recall values.

Finally, the ROC curve was computed as the macro-average (namely, the arithmetic average) value of class values of the TPR score, TPR_{macro} against the macro average of FPR, FPR_{macro}.

3 Results

3.1 Results of 2 classes classification

Training runs were performed following the 5-fold cross-validation: 5 different subsets were created with a different patient split between test and training, and five models were trained. Final scores are then calculated as the average of all five models results. In Fig. 4a, is reported the Confusion Matrix used to calculate all metrics. The proposed CNN results are compared with scores obtained from other models trained in the same way and built using three popular architectures: *Resnet18* [36], VGG16 [37], DarkNet19 [38]. As shown in Fig. 4b, every architecture performed virtually the same as the others, given the uncertainty associated with the 5-fold averaging. Specifically, accuracy metric reached on average 82% and Precision 68%. ROC AUC reached on average about 90%, meaning that models can make reliable predictions even at greatest threshold, reducing the FPR. Keeping a higher threshold also grants a lower number of examples classified as False Negatives, which are the most dangerous mistakes in a clinical environment.

3.2 Results of 3 classes classification

The Duke database comes with indications about malignancy of tumoral masses for examples proved by biopsy. This way, the original Positive class can be expanded to Malignant and Benign, thus making the task a 3-classes classification problem. Out of the 237 volumes originally selected for the dataset creation, 120 belonged to the positive class composed by 60 benign lesions and 60 malignant ones. This splitting gave rise to an unbalance between those classes and the negative one, which contains 117 volumes total. Training performed using all the negative examples or a limited subset showed that such an unbalance does not influence training in a significant way. Predictions were evaluated using the same metrics previously described but adapted to three classes according to [35]. All metrics are intended as weighted over the number of examples in each class except ROC AUC, which is computed over the macro-average of TPR over FPR (Fig. 5a). Results obtained with the DBT-DCNN model are shown in Fig. 5b and are encouraging, with Recall, Precision and F1 all reaching about 70%. Error and FPR are 30% and 23%, respectively. ROC AUC was the highest-scoring metric, reaching up to 80%.

4 Discussion

DBT has been shown to offer significantly higher sensitivity for breast cancer detection compared to the conventional two-view DM, which relies on mediolateral oblique and craniocaudal views [39–41]. This advantage stems from DBT's ability to capture three-dimensional (3D) imaging data, allowing for better visualization of dense breast tissue and improving the detection of subtle lesions or abnormalities that might otherwise be missed in standard 2D imaging techniques. Despite its benefits, DBT presents certain challenges that limit its widespread adoption. One of the primary issues is the sheer volume of images generated by the multiple slices in a single examination, which significantly increases the time required for radiologists to review and interpret the data. Studies have reported that the interpretation time for DBT can be 50% to 200% longer than that required for traditional DM, creating a substantial burden on radiologists and potentially affecting the efficiency of breast cancer screening programs [18, 42].

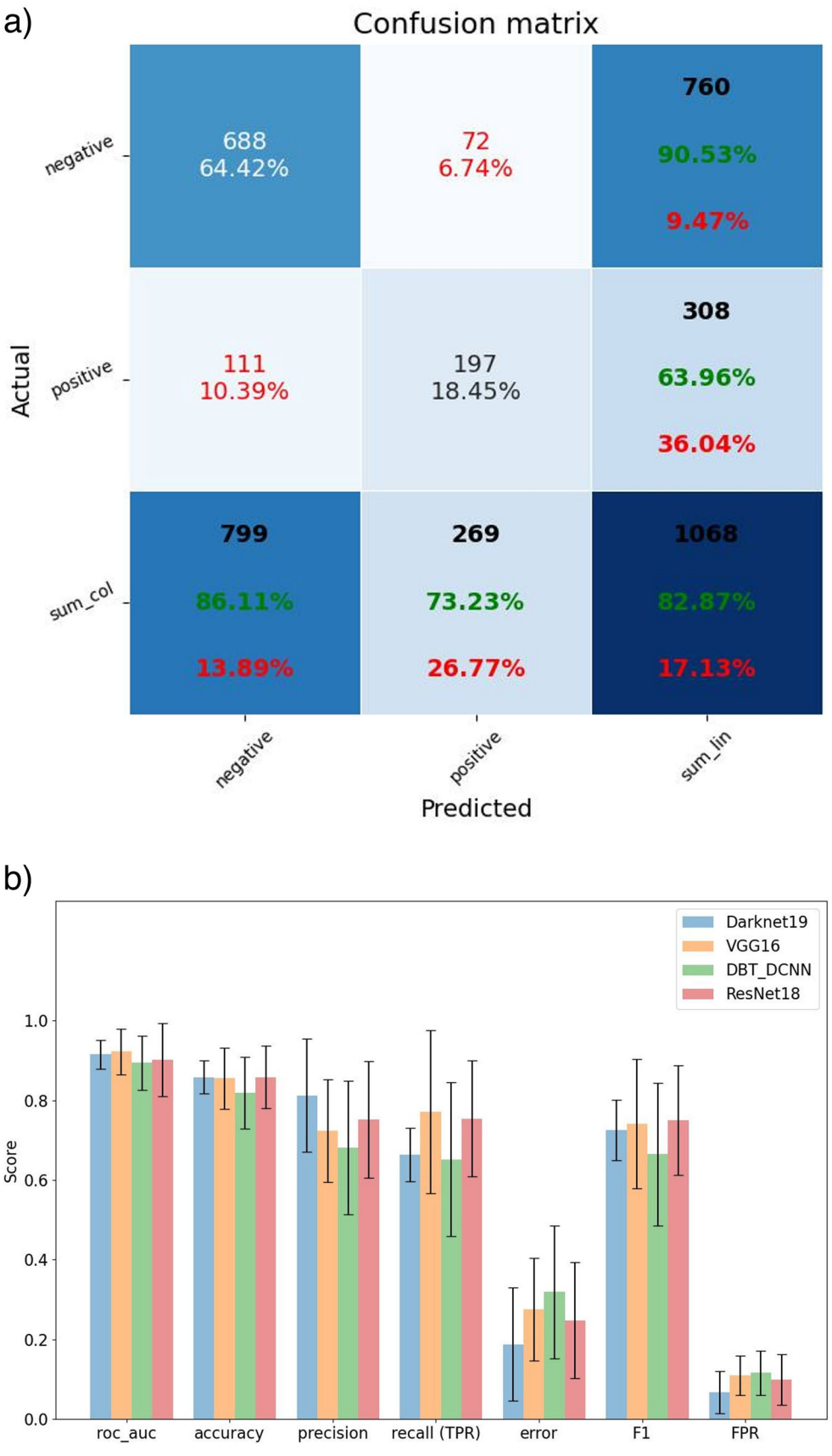

Fig. 4 (**a**) Confusion Matrix obtained by DBT-DCNN 2-class-model for one of the tranining Duke subsets and (**b**) Metrics scores obtained by DBT-DCNN, Darknet19, VGG19,6 and ResNet18 models trained and tested on the Duke dataset. Results are the average of five model predictions and errors have been calculated as the standard deviation. *Error* must be intended as 1-Precision

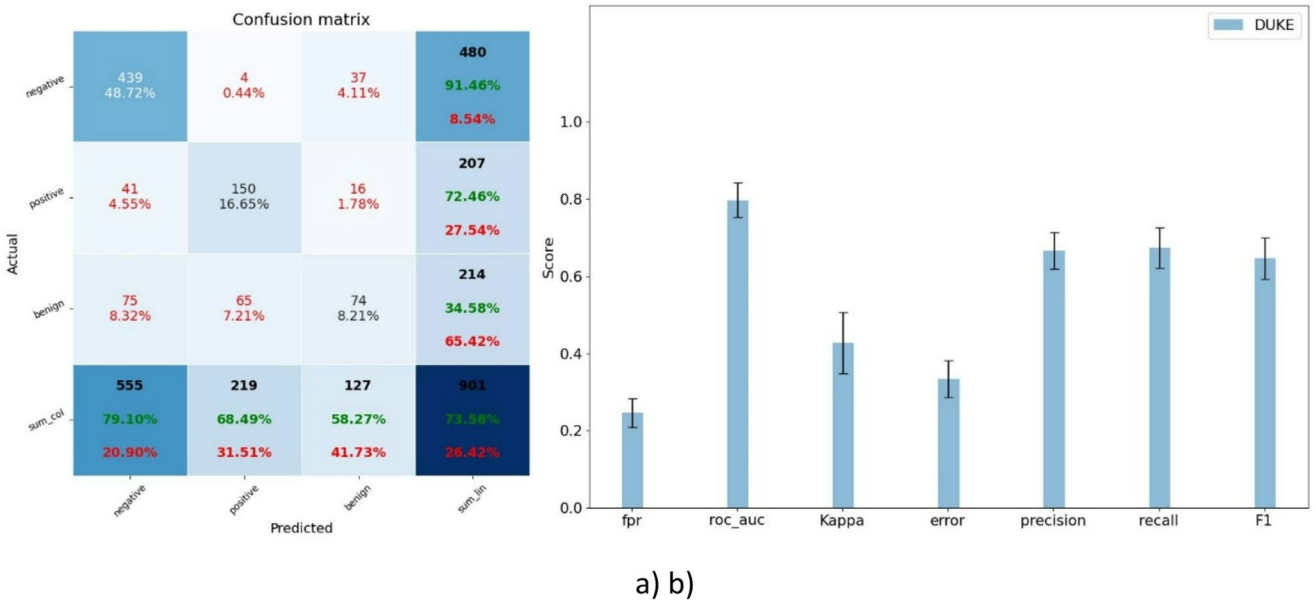

Fig. 5 (**a**) Confusion Matrix for one of the tranining Duke subsets and (**b**) Metrics scores obtained by DBT-DCNN 3-class-model trained and tested on the Duke dataset. Results are the average of five model predictions and errors have been calculated as the standard deviation. *Error* must be intended as 1-Precision

Given these challenges, there is an urgent need for advanced computer-aided detection (CAD) systems and diagnostic tools that can optimize the reading process and reduce the time commitment for radiologists while maintaining diagnostic accuracy. Innovative solutions are emerging to address this need. For example, iCAD's PowerLook Tomo Detection system, one of the first CAD platforms designed specifically for DBT, received FDA clearance in 2017 (https://www.icadmed.com). This system aims to assist radiologists by highlighting areas of interest in DBT images, thereby streamlining the interpretation process. Another player in this domain is ScreenPoint Medical, a Netherlands-based company that has developed Transpara, a deep learning-based software solution designed for multimodality applications (https://www.screenpoint-medical.com). Transpara leverages the full 3D information from DBT scans to analyze soft-tissue lesions and calcifications quantitatively. By marking relevant sections of the imaging data, Transpara is designed to improve radiologists' efficiency and reduce the time required to review complex DBT cases.

These activities are accompanied by numerous research activities carried out by different groups. Some examples are shown in the Table 1. The Table 1 shows other more or less recent examples of CAD based on DL and ML with their characteristics.

Looking at the latest results, in Buda et al. [26], the authors experimented with various loss functions, including binary cross-entropy, weighted binary cross-entropy, focal loss, and a reduced focal loss using the same dataset used in this work. Among these, the model demonstrated superior performance using focal loss, achieving a sensitivity of 60% with 2 false positives (FPs) per digital breast tomosynthesis (DBT) volume, highlighting the importance of tailored loss functions in improving the model's effectiveness in handling imbalanced datasets.

Table 1 Characterisics and performance of some research mass DBT classification system reported in literature

	# Patients	Input type	AUC	Sensitivity
Samala et al. [21]	94	ROI	0.9	
Kim et al. [24]	185	ROI	0.92	
Fotin et al. [22]	344	ROI		0.89
Buda et al. [26]	5610	slice		0.65
Zhang et al. [43]	1400	slice	0.85	
Li et al. [44]	261	slice		0.8
Samala et al. [45]	324	ROI	0.85	
Yousefi et al. [46]	87	slice	0.87	
Ricciardi et al. [25]	100	slice	0.89	
DBT-DCNN (this work)	237	slice	0.90	0.68

The study, reported in Li et al. 2021, incorporated prior information and individual DBT slices into a Faster R-CNN model to generate 2D candidates for each slice. Subsequently, these 2D candidates were aggregated to produce 3D candidates for the entire DBT volume using a sophisticated 3D aggregation scheme. This scheme relied on density-based spatial clustering to group 2D candidates into clusters. Within each cluster, the corresponding 2D binary candidate masks were concatenated, effectively combining spatially related predictions into cohesive 3D structures. The evaluation of this combined approach revealed a notable

Table 2 Number of free parameters of the different architectures implemented in this work

Architecture	Parameters
Ricciardi	10^8
DBT-DCNN	10^5
ResNet18	8×10^5
VGG16	7×10^5
Darknet19	10^7

improvement in detection performance. The proposed method achieved a sensitivity of 80% with 1.95 false positives per volume, demonstrating its potential to enhance DBT-based cancer detection by leveraging both 2D and 3D spatial information.

With the same purpose our group in Ricciardi et al. 2021 developed a DL-based model, called DBT-DCNN. The DBT-DCNN network developed in this work showed an accuracy and a sensitivity of (90% ± 4%) and (96% ± 3%), respectively, with an AUC as good as 0.89 ± 0.04. A *k*-fold cross validation test (with k = 4) showed an accuracy of 94.0% ± 0.2%, and a F1-score test provided a value as good as 0.93 ± 0.03. But this model presented some problems such as the high number of free parameters (1.9×10^8), a high training time and a high memory usage (3.5 GB per model).

In order to optimize our model, we re-write the network using python language and all the reported problems have been overcome. In particular, the number of free parameters has been reduced by three orders of magnitude. The obtained results in terms of AUC, precision, recall, etc. are in line with other networks (Fig. 4b) but as can be seen from the Table 2, it has a lower number of free parameters and therefore a shorter training time. In addition, the code is able to automatically support classification in three classes (Negative, Benign and Malign).

5 Conclusion

A new and optimized DBT-CNN classification network was realized in order to optimize the performance of our previously network. This takes advantage of a pre-processing algorithm to remove pectoral muscle and skin, before entering the lesion classification step. DBT-CNN can classify individual DBT images into two or in three classes: healthy, presence of a malignant or benignant masses. Accuracy and sensitivity on the binary classification of test data of the most suitable considered dataset reached 82% and 68%, respectively, with an AUC of 0.90. This performance was obtained with a CNN architecture that is significantly simpler than the ones proposed in the literature so far and is promising for future clinical implementation in routine DBT diagnosis. In the case of three classes classification the ROC AUC was the highest-scoring metric, reaching up to 80%.

Author contributions Conceptualization was done by G. Mettivier and G. Paternò; Supervision was done by G. Mettivier and G. Paternò; Software was carried out by M. Bianchini and Gianfranco Paternò; Formal analysis was carried out by G. Mettivier, G. Paternò, M. Bianchini, R. Ricciardi; Writing–Review and Editing was carried out by G. Mettivier, G. Paternò, R. Ricciardi, P. Russo, P. Cardarelli and A. Taibi.

Funding This research was funded by Istituto Nazionale di Fisica Nucleare (INFN) with the grant name "next_AIM".

Data availability Upon request to the corresponding author.

Code Availability Upon request to the corresponding author.

Declarations

Ethics approval It wasn't necessary.

Consent to participate Upon request to the corresponding author.

Consent for publication All authors agree with publication in this journal.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. AA.VV. How common is breast cancer? Breast cancer statistics. America Cancer Society. 2020 . https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html
2. Tabar L, Vitak B, Chen THH, Yen AMF, Cohen A, Tot T, Chiu SYH, Chen SLS, Fann JCY, Rosell J, Fohlin H, Smith RA, Duffy SW. Swedish two-country trial: impact of mammographic screening on breast cancer mortality during 3 decades. Radiology. 2011;260(3):658–63. https://doi.org/10.1148/radiol.11110469.
3. Jorgensen KJ, Gotzsche PC. Overdiagnosis in publicly organised mammography screening programmes: systematic review of incidence trends. BMJ. 2009;339:b2587. https://doi.org/10.1136/bmj.b2587.
4. Bond M, Pavey T, Welch K, Cooper C, Garside R, Dean S, Hyde C. Systematic review of the psychological consequences of false-positive screening mammograms. Health Technol Assess. 2013;17(13):1–170. https://doi.org/10.3310/hta17130.
5. Tosteson ANA, Fryback DG, Hammond CS, Hanna LG, Grove MR, Brown M, Wang Q, Lindfors K, Pisano ED. Consequences of false-positive screening mammograms. JAMA Intern Med. 2014;174(6):954–61.
6. Pharoah PDP, Sewell B, Fitzsimmons D, Bennet HS, Pashayan N. Cost effectiveness of the NHS breast screening programme: life table model. BMJ. 2013;346:f2618. https://doi.org/10.1136/bmj.f2618.
7. Elmore JG, Jackson SL, Abraham L, Miglioretti DL, Carney PA, Geller BM, Yankaskas BC, Kerlikowske K, Onega T, Rosenberg RD, Sickles EA, Buist DSM. Variability in interpretive performance at screening mammography and associated with accuracy. Radiology. 2009;253(3):641–51. https://doi.org/10.1148/radiol.2533082308.
8. Miglioretti DL, Smith-Bindman R, Abraham L, Brenner RJ, Carney PA, Bowles EJA, et al. Radiologist characteristics associated with interpretive performance of diagnostic

mammography. J Natl Cancer Inst. 2007;99(24):1854-1863e63. https://doi.org/10.1093/jnci/djm238.
9. Sechopoulos I. A review of breast tomosynthesis. Part I. The image acquisition process. Med Phys. 2013;40(1):014301. https://doi.org/10.1118/1.4770279.
10. Sarno A, Mettivier G, Russo P. Dedicated breast computed tomography. Basic Aspects Med Phys. 2015;42:2786–804. https://doi.org/10.1118/1.4919441.
11. Zhu Y, O'Connell AM, Ma Y, Liu A, Li H, Zhang Y, et al. Dedicated breast CT: state of the art – Part I. Historical evolution and technical aspects. Eur Radiol. 2022;32:1579–89. https://doi.org/10.1007/s00330-021-08179-z.
12. Zhu Y, O'Connell AM, Ma Y, Liu A, Li H, Zhang Y, et al. Dedicated breast CT: state of the art – Part II. Clinical application and future outlook. Eur Radiol. 2022;32:2286–300. https://doi.org/10.1007/s00330-021-08178-0.
13. Aase HS, Holen AS, Pedersen K, Houssami N, Haldorsen IS, Sebuodegard S, et al. A randomized controlled trial of digital breast tomosynthesis versus digital mammography in population-based screening in Bergen: interim analysis of performance indicators from the To-Be trial. Eur Radiol. 2019;29(3):1175–86. https://doi.org/10.1007/s00330-018-5690-x.
14. Clauser P, Nagl G, Helbich TH, Pinker-Domenig K, Weber M, Kapetas P, Bernathova M, Baltzer PA. Diagnostic performance of digital breast tomosynthesis with a wide scan angle compared to full-field digital mammography for the detection and characterization of microcalcifications. Eur J Radiol. 2016;85:2161–8. https://doi.org/10.1016/j.ejrad.2016.10.004.
15. Tagliafico AS, Calabrese M, Bignotti B, Signori A, Fisci E, Rossi F, Valdora F, Houssami N. Accuracy and reading time for six strategies using digital breast tomosynthesis in women withmammographically negative dense breasts. Eur Radiol. 2017;27(12):5179–84. https://doi.org/10.1007/s00330-017-4918-5.
16. Clauser P, Baltzer PAT, Kapetas P, Woitek R, Wweber M, Leone F, et al. One view or two views for wide-angle tomosynthesis with synthetic mammography in the assessment setting? Eur Radiol. 2022;32:661–70. https://doi.org/10.1007/s00330-021-08079-2.
17. Houssami N, Lockie D, Clemson M, Pridmore V, Taylor D, Marr G, Evans J, Macaskill P. Pilot trial of digital breast tomosynthesis (3D mammography) for population-basedscreening in breastscreen Victoria. Med J Aust. 2019;211:357–62. https://doi.org/10.5694/mja2.50320.
18. Chae EY, Kim HH, Jeong JW, Chae SH, Lee S, Choi YW. Decrease in interpretation time for both novice and experienced readers using a concurrent computer-aided detection system for digital breast tomosynthesis. Eur Radiol. 2019;29(5):2518–25. https://doi.org/10.1007/s00330-018-5886-0.
19. Balleyguier C, Arfi-Rouche J, Levy L, Toubiana PR, Cohen-Scali F, Toledano AY, Boyer B. Improving digital breast tomosynthesis reading time: a pilot multi-center, multi-case study using concurrent computer-aided detection (CAD). Eur J Radiol. 2017;97:83–9. https://doi.org/10.1016/j.ejrad.2017.10.014.
20. Benedikt RA, Boatsman JE, Swann CA, Kirkpatrick AD, Toledano AY. Concurrent computer-aided detection improves reading time of digital breast tomosynthesis and maintains interpretaions performance in a multireader multicase study. AJR Am J Roentgenol. 2018;210(3):685–94. https://doi.org/10.2214/AJR.17.18185.
21. Samala RK, Chan HP, Hadjiiski LM, Helvie MA, Wei J, Cha KH. Mass detection in digital breast tomosynthesis: deep convolutional neural network with transfer learning from mammography. Med Phys. 2016;43:6654–66. https://doi.org/10.1118/1.4967345.
22. Fotin SV, Yin Y, Haldankar H, Hoffmeister JW, Periaswamy S. 2016. Detection of soft tissue densities from digital breast tomosynthesis: comparison of conventional and deep learning approaches. In: SPIE medical imaging 2016, San Diego, California, United States, 97850X:1–6.
23. Kim DH, Kim ST, Ro YM. Latent feature representation with 3-D multi-view deep convolutional neural network for bilateral analysis in digital breast tomosynthesis. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), Shanghai, China. 2020; 927–931.
24. Kim DH, Kim ST, Chang JM, Ro YM. Latent feature representation with depth directional long-term recurrent learning for breast masses in digital breast tomosynthesis. Phys Med Biol. 2017;62:1009–31. https://doi.org/10.1088/1361-6560/aa504e.
25. Ricciardi R, Mettivier G, Staffa M, Sarno A, Acampora G, Minelli S, Santoro A, Antignani E, Orientale A, Pilotti IAM, Santangelo V, D'Andria P, Russo P. A deep learning classifier for digital breast tomosynthesis. Physica Med. 2021;83:194–193. https://doi.org/10.1016/j.ejmp.2021.03.021.
26. Buda M, Saha A, Walsh R, Ghate S, Li N, Swiçicki A, Lo JY, Mazurowski MA. Detection of masses and architectural distortions in digital breast tomosynthesis: a publicly available dataset of 5060 patients and a deep learning model. arXiv:2011.07995v4. 2021. https://doi.org/10.48550/arXiv.2011.07995
27. Buda M, Ashiribani S, Walsh R, Gate S, Li N, Swiecicki A, Lo JY, Mazurowski MA. A data set and deep learning algorithm for the detection of masses and architectural distorsions in digital breast tomisynthesis images. JAMA Netw Open. 2021;4(8):e2119100. https://doi.org/10.1001/jemanetworkopen.2021.19100.
28. Esposito D, Paternò G, Ricciardi R, Sarno A, Russo P, Mettivier G. A pre-processing tool to increase performance of deep learning-based CAD in digital breast Tomosynthesis. Health Technol. 2024;14:81–91. https://doi.org/10.1007/s12553-023-00804-9.
29. Li X, Jiao H, Wang Y. Edge detection algorithm of cancer image based on deep learning. Bioengineered. 2020;11:693–707. https://doi.org/10.1080/21655979.2020.1778913.
30. Santos CFGD, Papa JP. Avoiding overfitting: a survey on regularization methods for convolutional neural networks. ACM Comput Surv. 2022;54:1–213. https://doi.org/10.1145/3510413.
31. Andreozzi E, Fratini A, Esposito D, Cesarelli M, Bifulco P. Toward a priori noise characterization for real-time edge-aware denoising in fluoroscopic devices. Biomed Eng Online. 2021. https://doi.org/10.1186/s12938-021-00874-8. 20,.
32. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929–58. https://doi.org/10.5555/2627435.2670313.
33. Murphy KP. Probabilistic Machine Learning: An Introduction, MIT press 2022.
34. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via Gradient-based localization. Proc IEEE Int Conf Comput Vis (ICCV). 2017;618–26. https://doi.org/10.1109/ICCV.2017.74.
35. Grandini M, Bagli E, Visani G. Metrics for multi-class classification: an overview. In: arXiv preprint arXiv:2008.05756. 2020. https://doi.org/10.48550/arXiv.2008.05756
36. Kaiming H, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2016;770–778. https://doi.org/10.48550/arXiv.1512.03385
37. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. ArXiv Preprint arXiv: 1409 1556. 2014. https://doi.org/10.48550/arXiv.1409.1556.
38. Redmon J, Farhadi A. Yolov3: An incremental improvement. In: arXiv preprint arXiv:1804.02767. 2018. https://doi.org/10.48550/arXiv.1804.02767
39. Zackrisson S, Lång K, Rosso A, Johnson K, Dustler M, Fornvik D, Fornvik h, Sartor H, Timberg P, Tingberg A, Andersson I.

One-view breast tomosynthesis versus two-view mammography in the Malmö breast tomosynthesis screening trial (MBTST): a prospective, population-based, diagnostic accuracy study. Lancet Oncol. 2018;19:1493–503. https://doi.org/10.1016/S1470-2045(18)30521-7.

40. Skaane P, Sebuødegård S, Bandos AI, Gur D, Osteras BH, Gullien R, Hofvind S. Performance of breast cancer screening using digital breast tomosynthesis: results from the prospective population-based Oslo tomosynthesis screening trial. Breast Cancer Res Treat. 2018;169:489–96. https://doi.org/10.1007/s10549-018-4705-2.
41. Bernardi D, Macaskill P, Pellegrini M, Valentini M, Fantò C, Ostillio L, Tuttobene P, Luparia A, Houssami N. Breast cancer screening with tomosynthesis (3Dmammography) with acquired or synthetic 2D mammography compared with 2D mammography alone (STORM-2): a population-based prospective study. Lancet Oncol. 2016;17:1105–13. https://doi.org/10.1016/S1470-2045(16)30101-2.
42. Moger TA, Swanson JO, Holen ÅS, Hanestad B, Hofvind S. Cost differences between digital tomosynthesis and standard digital mammography in a breast cancer screening programme: results from the To-Be trial in Norway. Eur J Health Econ. 2019;20:1261–9. https://doi.org/10.1007/s10198-019-01094-7.
43. Zhang Y, Wang X, Blanton H, Liang G, Xing X, Jacobs N. 2D Convolutional Neural Networks for 3D Digital Breast Tomosynthesis classification. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, San Diego, CA, USA, 2019. 2020;1013–1017. https://doi.org/10.1109/BIBM47256.2019.8983097
44. Li Y, He Z, Lu Y, Ma X, Guo Y, Xie Z, et al. Deep learning of mammary gland distribution for architectural distortion detection in digital breast tomosynthesis. Phys Med Biol. 2021;66:035028. https://doi.org/10.1088/1361-6560/ab98d0.
45. Samala RK, Chan H, Hadjiiski L, Helvie MA, Richter CD, Cha KH. Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural Nets. IEEE Trans Med Imaging. 2019;38:686–96. https://doi.org/10.1109/TMI.2018.2870343.
46. Yousefi M, Krzyzak A, Suen CY. Mass detection in digital breast tomosynthesis data using convolutional neural networks and multiple instance learning. Comput Biol Med. 2018;96:283–93. https://doi.org/10.1016/j.compbiomed.2018.04.004.