Thick Slices for Optimal Digital Breast Tomosynthesis Classication With Deep-Learning

Paul Terrassin, Mickael Tardy, Hassan Alhajj, Nathan Lauzeral, Nicolas Normand

Thick Slices for Optimal Digital Breast Tomosynthesis Classification With Deep-Learning

Paul Terrassin[1,2], Mickael Tardy[1,2], Hassan Alhajj[1], Nathan Lauzeral[1], and Nicolas Normand[2]

[1] Hera-MI, SAS, Saint-Herblain, France
[2] Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

Abstract. Digital breast tomosynthesis (DBT) is a recent medical imaging tool that increases accuracy and interpretability compared to traditional full-field digital mammogram (FFDM). However, DBT interpretation time is estimated to be twice longer than for FFDM, explainable by its 3D nature. Computer-aided diagnosis (CAD) systems can help radiologists in their diagnostic tasks and workload reduction. However, computation times and costs are important for CADs, thus facing the same challenge as health practitioners. This study addresses the problem concerning the processing of DBTs with high cancer detection rates while meeting the constraints of the clinical world. To this end, we propose a method relying on the slabbing approach which generates a set of 2D thick slices "slabs" that summarize a whole DBT volume. We propose a comprehensive benchmark on slabbing exploring several parameters such as slab thickness and overlap between slabs. Our method uses a fully 2D convolutional neural network (CNN) as a binary classifier, trained solely on FFDMs, exploiting the similarity between FFDMs and slabs. We report metrics on the two publicly available datasets containing DBTs: Breast Cancer Screening-DBT (BCS-DBT) and EA1141. This is the first study to explore DBTs of the EA1141 dataset, so we provide data strategy details and make it publicly available on GitHub[3]. We report breast-wise AUC_{ROC} of **0.90** on both BCS-DBT validation and test subsets and **0.97** on EA1141. We achieve competitive specificities at 90% of sensitivity breast-wise with **0.84** and **0.79** on BCS validation and test respectively, while not training on DBTs.

Keywords: Breast Cancer · Digital Breast Tomosynthesis · Slab · Thick slices · Deep Learning · Classification · CNN.

1 Introduction

Breast cancer is the most diagnosed type of cancer among women and the second leading cause of death worldwide [19]. Digital breast tomosynthesis (DBT) is a recent 3D medical imaging tool that can be used in breast cancer screening

[3] https://github.com/racoon-z/dbt-slabbing

programs. Compared to the traditional full-field digital mammography (FFDM), DBT can lead to better accuracy in the diagnosis of breast cancer [11, 12]. Its 3D nature offers better context and interpretability for complex lesions such as micro-calcifications [9]. While overdiagnosis can lead to invasive surgical procedures for patients and increased workload for radiologists, early detection of breast disease remains crucial to patient care [19].

Despite its advantages, DBT analysis results in a significant increase in reading time, with the estimated interpretation time for radiologists using DBT being twice as long as for radiologists using FFDM [15]. Computer-aided diagnosis (CAD) systems aim to help health practitioners increase their cancer detection rates and reduce their workload [18]. The main current and future challenges for DBT concern the reduction of CAD computation time while maintaining high performance. Indeed, end-to-end DBT processing is resource-hungry for deep-learning-based methods, especially for convolutional neural network (CNN) due to the high resolution of the modality[4].

The current literature can be divided into three DBT processing strategies: 1) a single synthesized 2D view (S2D) summarizing a DBT entirely, 2) focusing on 3D patches of the region of interest (ROI), and 3) using the entire DBT in slices or generating thick slices, called "slabs".

The first approach is based on methods that process S2Ds generated by DBT manufacturers, comparable to the processing of FFDMs using a single 2D projection of the breast [6, 20]. It cancels out the advantages of DBT which offers 3D visualization of the lesion, and therefore reverts to FFDM usage. In addition, these views are not systematically provided by manufacturers and S2Ds can generate high-intensity artifacts that could be mistaken for lesions.

The second strategy relies on global model which identifies 3D ROI patches and then exploit neighborhood pixel information, fusing inter-slices features along the z axis of the DBT [21, 26]. These methods take greater account of the DBT nature by extracting multi-dimensional features from patches using both 2D and 3D grouped convolutions. While this approach remains interesting, it does not answer the stated clinical constraints above, as a global model is required to analyze all slices to extract ROIs.

The last identified strategy consists in DBT slices [5, 13, 14, 24] or thick slices processing [4, 22]. El-Shazli et al. [5] and Mota et al. [13] propose a per-slice DBT processing approach solely based on 2D convolutions, without using inter-slice information. Moreover, they both do a harsh slice resizing ($\approx 224\times224$ and 512×512 respectively) which induces a significant loss of information on the xy planes. Park et al. [14] and Wang et al. [24] process DBT in an end-to-end manner and use both 2D and 3D information. However, these methods require large-scale computing infrastructures and significant costs which are not meeting clinical resources limitations. Lastly, Doganay et al. [4] and Tardy et

[4] Standard slice thickness is approximately $1mm$, and the in-plane resolution is around $50\mu m$ to $140\mu m$. For example, for a DBT with dimensions of $80 \times 2500 \times 2000$, the total number of pixels is approximately 400 million.

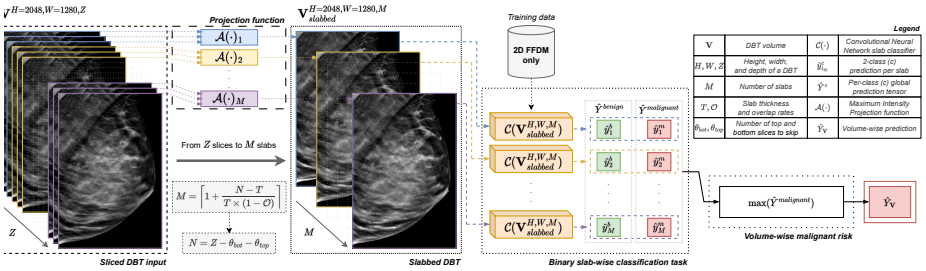

Fig. 1. Schema of the proposed method for an end-to-end classification of DBT volumes. First, we summarize a DBT volume $\mathbf{V}$ into M slabs. Then, a CNN architecture process the slabbed DBT to assess the risk of malignancy of the entire volume.

al. [22] summarize DBT pixel information along the z axis generating thick slices (slabs). However, they present relatively low performances on limited datasets.

Given the above constraints and our analysis of the state of the art, we propose a method leveraging the anisotropic property of DBT. While in-plane pixel information is crucial to detect small lesions such as micro-calcifications, we believe dimensionality reduction can be done along the z axis. A CNN-based classifier aims to assess the risk of malignancy for a slabbed DBT, summarizing all slices into a subset of slabs. In addition, our CNN is trained only on FFDMs to solve the problem of the lack of well-annotated DBT datasets. Finally, our contributions can be summarized as follows:

1. A comprehensive benchmark on the impact of slabbing for a CNN-based method to classify DBT volumes according to their risk of malignancy;
2. Large-scale metrics on all state-of-the-art clinically relevant DBT public datasets: Breast Cancer Screening DBT (BCS-DBT) validation and test sets [1], and EA1141 [2];
3. To the best of our knowledge, we are the first study to specifically use DBTs from EA1141, paving the way for the community to use a new dataset by providing details on data management and free access to our code.

2 Method

The specific details of our approach are: 1) the generation of slabs to summarize DBT volumes (slabbing) in Sec. 2.1 and 2) the use of a 2D CNN-based classifier trained solely on FFDMs in Sec. 2.2.

2.1 Slabbing

Slabbing is an effective approach for radiologists to reduce the interpretation time while having similar diagnostic accuracy [16, 17]. Indeed, DBT pixel information is not equally distributed along the volume as the in-plane (xy) resolution is

higher than in the z axis, proving the anisotropic nature of DBT. Our method relies on this property to generate a sequence of slabs from the input volume. This reduces the dimensionality, and therefore the computation time and cost, required to perform the whole volume classification task.

Let a DBT volume $\mathbf{V}^{H,W,Z}$ where H, W, and Z denote slice height, width, and depth respectively. The main aim of slabbing is to divide the volume $\mathbf{V}$ into a set of M slabs, along the z axis. Each slab is a 2D projection of a subset of slices from a DBT. The number of generated slabs is computed according to several parameters: the slab thickness $T \in [1, N]$, the overlapping ratio between slabs $\mathcal{O} \in [0, 1)$[5], and the number of slices $N = Z - \theta_{top} - \theta_{bot}$. $\theta \in \mathbb{N}$ represents the number of skipped slices at the boundaries usually suffering from substantial noise due to the nature of reconstruction. The number of slabs M is defined as in Eq. (1).

$$M = \left\lceil 1 + \frac{N - T}{T \times (1 - \mathcal{O})} \right\rceil \tag{1}$$

Once the new slabbed DBT thickness is defined, we use an aggregation function $\mathcal{A}$ to generate the 2D slabs. The new volume $\mathbf{V}^{H,W,M}_{slabbed} = \mathcal{A}(\mathbf{V}^{H,W,Z})$ is dimensionally reduced along the z axis. In our method, we use the maximum intensity projection (MIP) as $\mathcal{A}$, computing the maximum value of a voxel stack along the z axis [3]. This strategy has the advantage of enhancing the microcalcifications contrasts, however, it may generate high-intensity artifacts [3].

2.2 CNN

The classification is performed using a deep convolutional neural network (CNN) relying on UNet3+ [7] and with architecture modifications as described in [23]. These modifications include the multi-task and multi-scale output strategies designed to improve classification performance. Despite the use of a U-shape architecture recognized for segmentation tasks, this modified version of the UNet3+ improves classification performance compared to other state-of-the-art CNN classification methods [23]. Moreover, architectural changes such as the use of depth-wise separable 2D convolutions, and the reduction of convolution filters by a factor 2 aim to decrease model complexity. This is crucial for processing FFDM or DBT slabs which are high-resolution 2D images.

We introduce further adaptations to fit the needs of the study. We trained our CNN for the binary classification task of full FFDM images assessing the risk of malignancy. Our CNN classifier has been trained solely on 2D mammography images and without using a single DBT exam due to the lack of well annotated datasets. Moreover, our method takes advantage of the similarity between FFDM and DBT slabs allowing us to minimize training resources.

In our method, we propose to use this CNN architecture as a classifier for our slabbed volume $\mathbf{V}_{slabbed}$ as follows. Classification function $\mathcal{C}$ generates a set

[5] $\mathcal{O} = 1$ is excluded as it indicates that each slab fully overlaps the previous one, resulting in a stride of 0, which causes the algorithm to get stuck in an infinite loop.

of M predictions ($\hat{y}$) according to binary classification problem as in Eq. (2). The two classes c predicted are benign ($c = b$) and malignant ($c = m$).

$$\mathcal{C}(\mathbf{V}^{H,W,M}_{slabbed}) = [\{\hat{y}^b_1, \hat{y}^b_2, \cdots, \hat{y}^b_M\}, \{\hat{y}^m_1, \hat{y}^m_2, \cdots, \hat{y}^m_M\}] \tag{2}$$

where $\hat{y}^c_i$ is the generated prediction for the i_{th} slab of the c class. Global tensors are: $\hat{Y}^{benign} = \{\hat{y}^b_1, \cdots, \hat{y}^b_M\}$ and $\hat{Y}^{malignant} = \{\hat{y}^m_1, \cdots, \hat{y}^m_M\}$.

Finally, we aim to obtain a single score for the entire volume $\hat{Y}_V$ to assess the risk of malignancy from the above set of predictions. To that end, we compute the maximum from the malignant class predictions, i.e., $\hat{Y}_{\mathbf{V}} = \max(\hat{Y}^{malignant})$, meaning capturing a malignancy on one slab at least.

3 Experiments

3.1 Datasets

The two public datasets containing DBT exams from clinical practice have been used: Breast Cancer Screening-Digital Breast Tomosynthesis (BCS-DBT) [1], and EA1141 [2]. BCS-DBT is a large-scale and well-annotated public DBT dataset that has become state-of-the-art since its release in 2021 in the context of a DBTex Lesion Detection Challenge [8]. To allow the comparison to the results of the challenge we used the validation and test subsets. These subsets are referred to as BCS_{val} and BCS_{test} respectively. EA1141 is a dataset composed of multi-modal exams, including FFDMs, DBTs, and magnetic resonance images (MRIs). All required information to understand the specific processing applied to sort images according to modalities and identify DBT exams is available on the GitHub mentioned on the first page.

We evaluate the method in two scopes: image-wise (IW) and breast-wise (BW). Our goal is to distinguish benign and normal DBTs from malignant ones focusing on biopsy-proven malignant lesions. For the two BCS-DBT subsets, it means that "Cancer" labels only were considered as malignant and "Benign", "Actionable", and "Normal" ground truths were assigned to the benign class (we refer the reader to the original manuscript for the details). The sorting strategy for EA1141 was different as the dataset contains both an MRI and a DBT clinical outcome. Therefore, we propose the two following strategies: 1) samples with malignant biopsy outcome from DBT vs. the rest, excluding malignant exams detected from MRIs (this subset is referred to as EA_{DBT}), and 2) malignant biopsy outcome from DBT and MRI vs. the rest (referred to as EA_{MRI}). The malignancy class was assigned when "DCIS" or "Invasive" words were present in the biopsy outcomes of DBTs and MRIs. Tab. 1 summarizes the distribution between benign and malignant sets.

3.2 Implementation details

The CNN architecture exploited in the study follows the training procedure similar to Terrassin et al [23] except that we use full FFDMs instead of patches

Table 1. IW and BW data distribution between benign and malignant classes for each dataset. BCS_{test} and BCS_{val} denote BCS-DBT Validation and Test subsets while EA_{DBT} and EA_{MRI} refer to the two sets, excluding or including MRI biopsy outcomes.

		BCS_{val}	BCS_{test}	EA_{DBT}	EA_{MRI}
Image-wise	Benign	1126	1661	1425	1803
	Malignant	37	60	4	17
Breast-wise	Benign	565	830	679	861
	Malignant	20	30	2	8

and train on the image-wise binary classification task. The neural network was optimized to be trained on the NVIDIA GeForce RTX 2080 Ti GPU, fitting 11GB RAM, and is capable of inferring on CPU.

We standardized DBT slices before the inference as follows. First, erasing noisy background pixels, using the triangle threshold method [27]. Then, we removed skin borders [1] and cropped slices aiming to suppress irrelevant pixels [25]. DBT slices were resized to 2048×1280 pixels to align with the input expected by the CNN used, and at last, we truncated the histogram from extreme values and normalized pixel intensity in the range $[0, 1]$ as in [23].

We aimed to evaluate several parameters of our method by creating an experimental plan with different slab thicknesses $T = [1, 6, 8, 10, 12, 15, 20, N]$. When $T = 1$, it consists of processing all DBT slices and $T = N$ a single in-plane, similar to S2D. We evaluated $\mathcal{O} = 0.5$ and $\mathcal{O} = 0$ corresponding to 50% and no overlap between slabs, respectively. We also evaluated $\theta = 0$ and $\theta = T \times 0.5$, i.e., keeping noisy slices or removing them.

4 Results

We computed the following classification metrics: Area Under the ROC Curve (AUC_{ROC}), the Area under Precision-Recall curve (AUC_{PR}), and Spec@90%. AUCs allow us to measure the classifier's ability to distinguish between the benign and malignant classes, while sensitivities and specificities assess the true positive and true negative rates. Aiming to improve the detection rate in clinical practice, we also evaluated the specificity (Spec@90%) of the method when setting the sensitivity to above 90% (i.e., higher than an average of human reading of 87.4% [10]). We report IW and BW metrics with 95% confidence intervals (CIs), using the bootstrap approach presented by Buda et al. [1].

Based on our experiments, we found that $T = N$ was the worst slab thickness, with drastically lower AUCs and Spec@90%. It can be explained by the maximum intensity projection method used for slabbing, which generates important noisy artifacts on thick slabs and prevents them from benefitting from the 3D nature of DBT. No clear trend is observed in performances with and without overlap between slabs ($\mathcal{O} = 0$ and $\mathcal{O} = 0.5$) as very comparable AUC_{PR} , AUC_{PR} and Spec@90% are achieved. However, excluding overlapping allows to

Table 2. Metrics table with the ROC AUC (AUC_{ROC}), Precision-Recall AUC (AUC_{PR}), and Spec@90%. IW and BW metrics are reported on the three datasets BCS_{val} ,BCS_{test}, and EA_{DBT} following the ablation on several slab thicknesses: $T \in [6, 8, 10, 15, 20]$ mm.

Dataset	T	Image-Wise (IW)			Breast-Wise (BW)		
		AUC_{ROC}	AUC_{PR}	**Spec@90%**	AUC_{ROC}	AUC_{PR}	**Spec@90%**
BCS_{val}	6	0.86 (0.78-0.93)	0.25 (0.14-0.39)	0.55 (0.52-0.58)	**0.90** (0.79-0.97)	0.28 (0.14-0.51)	**0.87** (0.84-0.90)
	8	0.87 (0.79-0.94)	0.26 (0.15-0.42)	**0.66** (0.64-0.69)	**0.90** (0.79-0.97)	0.31 (0.16-0.54)	0.84 (0.81-0.87)
	10	**0.88** (0.81-0.93)	**0.29** (0.16-0.44)	0.65 (0.63-0.68)	**0.90** (0.80-0.97)	**0.36** (0.17-0.59)	0.80 (0.77-0.83)
	15	0.86 (0.78-0.92)	0.25 (0.13-0.39)	0.50 (0.48-0.53)	0.88 (0.78-0.95)	0.29 (0.12-0.50)	0.79 (0.76-0.83)
	20	0.86 (0.78-0.92)	0.23 (0.13-0.38)	0.61 (0.58-0.64)	0.89 (0.78-0.96)	0.25 (0.12-0.48)	0.81 (0.77-0.84)
BCS_{test}	6	**0.86** (0.81-0.91)	0.25 (0.16-0.37)	**0.58** (0.56-0.60)	0.88 (0.82-0.94)	0.25 (0.14-0.42)	0.73 (0.69-0.76)
	8	**0.86** (0.81-0.91)	0.29 (0.18-0.42)	0.49 (0.46-0.51)	**0.90** (0.84-0.95)	0.32 (0.17-0.51)	**0.79** (0.76-0.81)
	10	**0.86** (0.82-0.91)	0.27 (0.17-0.40)	0.54 (0.52-0.57)	0.89 (0.82-0.94)	0.28 (0.16-0.45)	0.74 (0.71-0.76)
	15	0.85 (0.80-0.90)	0.25 (0.16-0.38)	0.53 (0.51-0.56)	0.85 (0.76-0.92)	0.25 (0.13-0.43)	0.46 (0.43-0.50)
	20	0.82 (0.77-0.88)	**0.32** (0.20-0.45)	0.51 (0.48-0.53)	0.85 (0.77-0.91)	**0.33** (0.16-0.51)	0.49 (0.46-0.53)
EA_{DBT}	6	0.84 (0.48-0.98)	0.02 (0.00-0.07)	0.47 (0.45-0.50)	**0.97** (0.95-0.98)	**0.04** (0.02-0.12)	**0.97** (0.95-0.98)
	8	0.85 (0.48-0.99)	**0.03** (0.00-0.08)	0.48 (0.45-0.50)	**0.97** (0.95-0.98)	**0.04** (0.02-0.12)	0.96 (0.95-0.98)
	10	**0.86** (0.54-0.98)	0.02 (0.00-0.07)	0.53 (0.51-0.56)	0.96 (0.94-0.97)	0.03 (0.01-0.09)	0.96 (0.94-0.97)
	15	0.80 (0.34-0.98)	0.02 (0.00-0.06)	0.33 (0.31-0.36)	0.92 (0.84-0.98)	0.02 (0.00-0.08)	0.86 (0.83-0.88)
	20	0.84 (0.58-0.99)	0.02 (0.00-0.08)	**0.57** (0.55-0.60)	0.86 (0.72-0.99)	0.02 (0.00-0.09)	0.74 (0.71-0.77)

reduce the inference time as fewer slabs are processed. We observed identical metrics varying θ values , meaning the CNN is not influenced by the noise in the volume boundaries. The slab thicknesses ablation is reported in Tab. 2.

Clinically meaningful BW AUC_{ROC} scores are obtained on the BCS_{val} and BCS_{test} sets reaching **0.90** in both cases. Those performances are remarkable as they are achieved using a classifier trained only on FFDMs. Interestingly, it mimics the radiologists' performances on slabbed volumes observed in [16]. The highest IW and BW AUC_{ROC} are generally obtained with 8 and 10-mm slabs on the three datasets. Best AUC_{PR} scores are also achieved with these thicknesses on the BCS_{val} subset, with only the 20-mm slab outperforming on BCS_{test}.

Regarding EA_{DBT} , we achieved an AUC_{ROC} of **0.97** and a Spec@90% of **0.97**, yet noting a very small malignant population (4 malignant DBTs, 2 breasts) as shown by the low AUC_{PR} average to ≈ 0.03. When we consider EA_{MRI}, *i.e.,* including clinical outcomes from MRI, we observe a remarkable drop in performance with the best AUC_{ROC} scores of 0.67 and 0.76 IW, BW respectively. Still, we note the sensitivity of the method to be higher than that of the radiologists when compared to DBT BI-RADS assessments.

We explored the ways to maximize predictions from different thicknesses by aggregating predictions from slabs of different thicknesses, but no improvements have been observed with similar averaged BW AUC_{ROC} of 0.90, 0.88, and 0.97 for BCS_{val} ,BCS_{test}, and EA_{DBT} respectively.

To compare with other state-of-the-art methods, we computed the same metrics using the predictions published from the DBTex phase 2 challenge [8] as shown in Tab. 3. We can see that NYU BTeam [8] and Zedus [8] teams outperformed our method on all but AUC_{PR} metrics. Nevertheless, the performances remain comparable, given the reported CIs. Noteworthy, we achieved these results by learning from FFDMs only, while all top-performing challenges included DBT data in training.

Table 3. State-of-the-art comparison table between our method using a 8-mm slab and methods proposed from the DBTex phase 2 challenge [8]. We report binary classification metrics using the same methodology as mentioned above.

Methods	BCS_{val}			BCS_{test}		
	AUC_{ROC}	AUC_{PR}	Spec@90%	AUC_{ROC}	AUC_{PR}	Spec@90%
NYU BTeam	0.95 (0.93-0.97)	**0.39** (0.20-0.64)	0.89 (0.86-0.93)	**0.93** (0.91-0.95)	0.27 (0.15-0.41)	**0.86** (0.78-0.90)
Vicorob	0.93 (0.89-0.96)	0.33 (0.15-0.55)	0.85 (0.74-0.90)	**0.93** (0.90-0.96)	**0.33** (0.19-0.52)	0.78 (0.68-0.95)
Zedus	**0.96** (0.93-0.98)	0.35 (0.19-0.58)	**0.90** (0.85-0.94)	0.92 (0.88-0.95)	0.25 (0.14-0.40)	**0.86** (0.59-0.90)
Ours (8mm slab)	0.90 (0.79-0.97)	0.31 (0.16-0.54)	0.84 (0.81-0.87)	0.90 (0.84-0.95)	0.32 (0.17-0.51)	0.79 (0.76-0.81)

The processing time decreases when reducing the number of slabs M. We timed the inferences of 2787 volumes (BCS_{val} and BCS_{test}) on 2 CPUs (AMD EPYC9474@3.6GHz 48 cores). The obtained average times per volume are: 74s ($T = 6$), 55s ($T = 8$), 45s ($T = 10$), and 37s ($T = 12$), resulting in an average $\approx 3.72s$ per image. Hence, processing slices separately gives a volume-wise inference time of $\approx 372s$ for DBT with 100 slices, which may not be acceptable in practice. Moreover, no performance gain was observed in none of the dataset when processing slices independently compared to $6 - 10$mm thicknesses.

5 Conclusion

In this study, we evaluate a DBT classification method in the context of breast cancer screening. To align with the lack and heterogeneity of DBT training data, we used a 2D CNN architecture trained solely on FFDMs. To preserve the resolution and reduce computation times and costs, we propose to use a slabbing approach to summarize an entire DBT volume into several 2D slabs. This strategy leverages the anisotropic property of DBT, copes with the scarcity of isolated slices, and mimics the running clinical practices.

We propose a comprehensive benchmark evaluating different sets of parameters to generate slabs. We place the proposed method in the realistic screening scenario (*i.e.,* strongly imbalanced towards benign and normal cases), evaluating on two public screening datasets: BCS-DBT and EA1141. We share the splits used for EA1141 in a GitHub repository to facilitate future works.

Our method achieves high performances with breast-wise AUC_{ROC} of **0.90** on BCS-DBT validation and test subsets and **0.97** on EA_{DBT}. To evaluate the clinical viability, we show specificity at 90% of sensitivity on the three datasets obtaining **0.87**, **0.79**, and **0.97** for BCS_{val}, BCS_{test}, and EA_{DBT} respectively. Future works will focus on the integration of DBTs in training data, and the design of a method that fuses both 2D and 3D features to exploit inter-slices information to improve both classification and detection performance.

Acknowledgements

This research is supported by the CIFRE program granted by the French ANRT organism under contract no. 2022/155. Computational resources were provided by the CPER Pays de la Loire Datacenter et Calcul Scientifique (DaCaS) project and GLiCID cluster.

References

1. Buda, M., Saha, A., Walsh, R., Ghate, S., Li, N., Święcicki, A., Lo, J.Y., Mazurowski, M.A.: A data set and deep learning algorithm for the detection of masses and architectural distortions in digital breast tomosynthesis images. JAMA network open **4**(8), e2119100–e2119100 (2021)
2. Comstock, C.E., Gatsonis, C., Newstead, G.M., Snyder, B.S., Gareen, I.F., Bergin, J.T., Rahbar, H., Sung, J.S., Jacobs, C., Harvey, J.A., Nicholson, M.H., Ward, R.C., Holt, J., Prather, A., Miller, K.D., Schnall, M.D., Kuhl, C.K.: Abbreviated breast mri and digital tomosynthesis mammography in screening women with dense breasts (ea1141). The Cancer Imaging Archive (2023)
3. Diekmann, F., Meyer, H., Diekmann, S., Puong, S., Muller, S., Bick, U., Rogalla, P.: Thick slices from tomosynthesis data sets: phantom study for the evaluation of different algorithms. Journal of digital imaging **22**, 519–526 (2009)
4. Doganay, E., Li, P., Luo, Y., Chai, R., Guo, Y., Wu, S.: Breast cancer classification from digital breast tomosynthesis using 3d multi-subvolume approach. In: Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications. vol. 11318, pp. 103–109. SPIE (2020)
5. El-Shazli, A.M.A., Youssef, S.M., Soliman, A.H.: Intelligent computer-aided model for efficient diagnosis of digital breast tomosynthesis 3d imaging using deep learning. Applied Sciences **12**(11), 5736 (2022)
6. Hassan, L., Saleh, A., Singh, V.K., Puig, D., Abdel-Nasser, M.: Detecting breast tumors in tomosynthesis images utilizing deep learning-based dynamic ensemble approach. Computers **12**(11), 220 (2023)
7. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 1055–1059. IEEE (2020)
8. Konz, N., Buda, M., Gu, H., Saha, A., Yang, J., Chłędowski, J., Park, J., Witowski, J., Geras, K.J., Shoshan, Y., et al.: A competition, benchmark, code, and data for using artificial intelligence to detect lesions in digital breast tomosynthesis. JAMA network open **6**(2), e230524–e230524 (2023)
9. Kopans, D., Gavenonis, S., Halpern, E., Moore, R.: Calcifications in the breast and digital breast tomosynthesis. The breast journal **17**(6), 638–644 (2011)
10. Lee, C.I., Abraham, L., Miglioretti, D.L., Onega, T., Kerlikowske, K., Lee, J.M., Sprague, B.L., Tosteson, A.N., Rauscher, G.H., Bowles, E.J., et al.: National performance benchmarks for screening digital breast tomosynthesis: update from the breast cancer surveillance consortium. Radiology **307**(4), e222499 (2023)
11. McDonald, E.S., McCarthy, A.M., Akhtar, A.L., Synnestvedt, M.B., Schnall, M., Conant, E.F.: Baseline screening mammography: performance of full-field digital mammography versus digital breast tomosynthesis. American Journal of Roentgenology **205**(5), 1143–1148 (2015)
12. Michell, M., Iqbal, A., Wasan, R., Evans, D., Peacock, C., Lawinski, C., Douiri, A., Wilson, R., Whelehan, P.: A comparison of the accuracy of film-screen mammography, full-field digital mammography, and digital breast tomosynthesis. Clinical radiology **67**(10), 976–981 (2012)
13. Mota, A.M., Clarkson, M.J., Almeida, P., Matela, N.: Automatic classification of simulated breast tomosynthesis whole images for the presence of microcalcification clusters using deep cnns. Journal of Imaging **8**(9), 231 (2022)

14. Park, J., Chłędowski, J., Jastrzębski, S., Witowski, J., Xu, Y., Du, L., Gaddam, S., Kim, E., Lewin, A., Parikh, U., et al.: An efficient deep neural network to classify large 3d images with small objects. IEEE Transactions on Medical Imaging (2023)
15. Partridge, G.J.W., Darker, I., James, J.J., Satchithananda, K., Sharma, N., Valencia, A., Teh, W., Khan, H., Muscat, E., Michell, M.J., et al.: How long does it take to read a mammogram? investigating the reading time of digital breast tomosynthesis and digital mammography. European Journal of Radiology p. 111535 (2024)
16. Pujara, A.C., Joe, A.I., Patterson, S.K., Neal, C.H., Noroozian, M., Ma, T., Chan, H.P., Helvie, M.A., Maturen, K.E.: Digital breast tomosynthesis slab thickness: impact on reader performance and interpretation time. Radiology **297**(3), 534–542 (2020)
17. Sauer, S.T., Christner, S.A., Kuhl, P.J., Kunz, A.S., Huflage, H., Luetkens, K.S., Schlaiß, T., Bley, T.A., Grunz, J.P.: Artificial-intelligence-enhanced synthetic thick slabs versus standard slices in digital breast tomosynthesis. The British Journal of Radiology **96**(1145), 20220967 (2023)
18. Shoshan, Y., Bakalo, R., Gilboa-Solomon, F., Ratner, V., Barkan, E., Ozery-Flato, M., Amit, M., Khapun, D., Ambinder, E.B., Oluyemi, E.T., et al.: Artificial intelligence for reducing workload in breast cancer screening with digital breast tomosynthesis. Radiology **303**(1), 69–77 (2022)
19. Siegel, R.L., Miller, K.D.e.a.: Cancer statistics, 2023. CA: A Cancer Journal for Clinicians **73**(1), 17–48. https://doi.org/10.3322/caac.21763
20. Singh, S., Matthews, T.P., Shah, M., Mombourquette, B., Tsue, T., Long, A., Almohsen, R., Pedemonte, S., Su, J.: Adaptation of a deep learning malignancy model from full-field digital mammography to digital breast tomosynthesis. In: Medical Imaging: Computer-Aided Diagnosis. vol. 11314, pp. 25–32. SPIE (2020)
21. Sun, H., Wu, S., Chen, X., Li, M., Kong, L., Yang, X., Meng, Y., Chen, S., Zheng, J.: Sah-net: Structure-aware hierarchical network for clustered microcalcification classification in digital breast tomosynthesis. IEEE Trans. on Cybernetics (2022)
22. Tardy, M., Mateus, D.: Trainable summarization to improve breast tomosynthesis classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 140–149. Springer (2021)
23. Terrassin, P., Tardy, M., Lauzeral, N., Normand, N.: Annotation-free deep-learning framework for microcalcifications detection on mammograms. In: Medical Imaging 2024: Computer-Aided Diagnosis. vol. 12927, pp. 208–217. SPIE (2024)
24. Wang, J., Sun, H., Jiang, K., Cao, W., Chen, S., Zhu, J., Yang, X., Zheng, J.: Capnet: Context attention pyramid network for computer-aided detection of microcalcification clusters in digital breast tomosynthesis. Computer Methods and Programs in Biomedicine **242**, 107831 (2023)
25. Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzębski, S., Févry, T., Katsnelson, J., Kim, E., et al.: Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans. on Medical Imaging **39**(4), 1184–1194 (2019)
26. Xiao, B., Sun, H., Meng, Y., Peng, Y., Yang, X., Chen, S., Yan, Z., Zheng, J.: Classification of microcalcification clusters in digital breast tomosynthesis using ensemble convolutional neural network. BioMedical Engineering OnLine **20**, 1–20 (2021)
27. Zack, G.W., Rogers, W.E., Latt, S.A.: Automatic measurement of sister chromatid exchange frequency. J. of Histochemistry & Cytochemistry **25**(7), 741–753 (1977)