

Joint 2D-3D Breast Cancer Classification

Gongbo Liang^{1*}, Xiaoqin Wang², Yu Zhang¹, Xin Xing¹, Hunter Blanton¹, Tawfiq Salem³, Nathan Jacobs¹

¹ Department of Computer Science, University of Kentucky, Lexington, KY, USA

² Department of Radiology, University of Kentucky, Lexington, KY, USA

³ Department of Computer Science and Software Engineering, Miami University, Oxford, Ohio, USA

Email: liang@cs.uky.edu^{*}

Abstract—Breast cancer is the malignant tumor that causes the highest number of cancer deaths in females. Digital mammograms (DM or 2D mammogram) and digital breast tomosynthesis (DBT or 3D mammogram) are the two types of mammography imagery that are used in clinical practice for breast cancer detection and diagnosis. Radiologists usually read both imaging modalities in combination; however, existing computer-aided diagnosis tools are designed using only one imaging modality. Inspired by clinical practice, we propose an innovative convolutional neural network (CNN) architecture for breast cancer classification, which uses both 2D and 3D mammograms, simultaneously. Our experiment shows that the proposed method significantly improves the performance of breast cancer classification. By assembling three CNN classifiers, the proposed model achieves 0.97 AUC, which is 34.72% higher than the methods using only one imaging modality.

Index Terms—Digital mammography, digital breast tomosynthesis, convolutional neural network, clinical inspired

I. INTRODUCTION

Breast cancer is the leading cause of cancer death in over 100 countries [1], [2]. Mammography is the only image screening tool that has been proven to reduce breast cancer mortality [3]. Digital mammography (DM or 2D mammography) and digital breast tomosynthesis (DBT or 3D mammography) are the two types of mammograms that are used in clinical practice [4]. Radiologists usually read both imaging modalities in combination, often looking for changes from slice-to-slice in DBT and comparing that with structures in DM [5]. However, interpreting mammograms is a challenging task, requiring many years of professional training. It is also time-consuming and therefore an expensive process [6]. This is especially problematic given the worldwide shortage of specialized breast radiologists [7].

Deep learning has demonstrated revolutionary potential in medical imaging analysis [8], [9], [10]. Given the need for highly efficient and accurate mammogram analysis, numerous deep learning-based computer-aided diagnosis (CAD) models have been developed [11], [12], [13]. However, the existing models typically focus on using either DM or DBT.

Inspired by clinical practice, we propose a novel breast cancer classification approach using convolutional neural networks (CNN) combined with ensemble strategy. The proposed

This study is sponsored by Grant No. IRG 16-182-28 from the American Cancer Society and Grant No. IIS-1553116 from the National Science Foundation.

network simultaneously reads DM and DBT as what radiologists would do in their daily practice. One key challenge of this work is how to use DBT effectively. The data size of DBT is large and with varying depths (on average, each DBT has $1024 \times 1024 \times 82$ voxels in this study). Training a 3D CNN model for such large data is extremely costly in terms of computation and memory, and may potentially lead to overfitting. We innovatively extract a fixed-size slice representation for each DBT, which captures the changes between DBT slices, and use a 2D CNN for classification. From our experiment, the proposed method has improved the performance significantly. In summary, the proposed method has the following advantages.

- To our best knowledge, this is the first model using whole DM and DBT simultaneously.
- We innovatively extract a fixed-size representation for each DBT. The extracted representation captures the changes between different slices of the same DBT.
- We use a real-world clinical dataset in this study. To our best knowledge, this is the largest breast cancer dataset that contains paired DM and DBT.
- Our method only requires image-level labels for training and significantly improves performance compared to other approaches trained similarly.

II. BACKGROUND

A. Existing Deep Learning Models

Ribli et al. used an r-CNN-based [14] approach to classify the 2D mammograms and that achieved 0.95 AUC (area under the receiver operating characteristic curve) for breast tumor classification [11]. This work won 2nd place at the Digital Mammography DREAM Challenge [15]. Shen et al. [16] designed a fully convolutional network for mammogram classification, which achieved 0.94 AUC. Though the performances for these two models are impressive, both works were trained using bounding boxes (BBs), which are usually not available on clinical data due to the high obtaining cost for medical images. More importantly, these methods only designed for 2D mammograms. None of them works on DBT.

Mendel et al. [13] proposed a model using a pre-trained VGG19 [17] network as the feature extractor and using support vector machine (SVM) as the classifier to separately evaluate

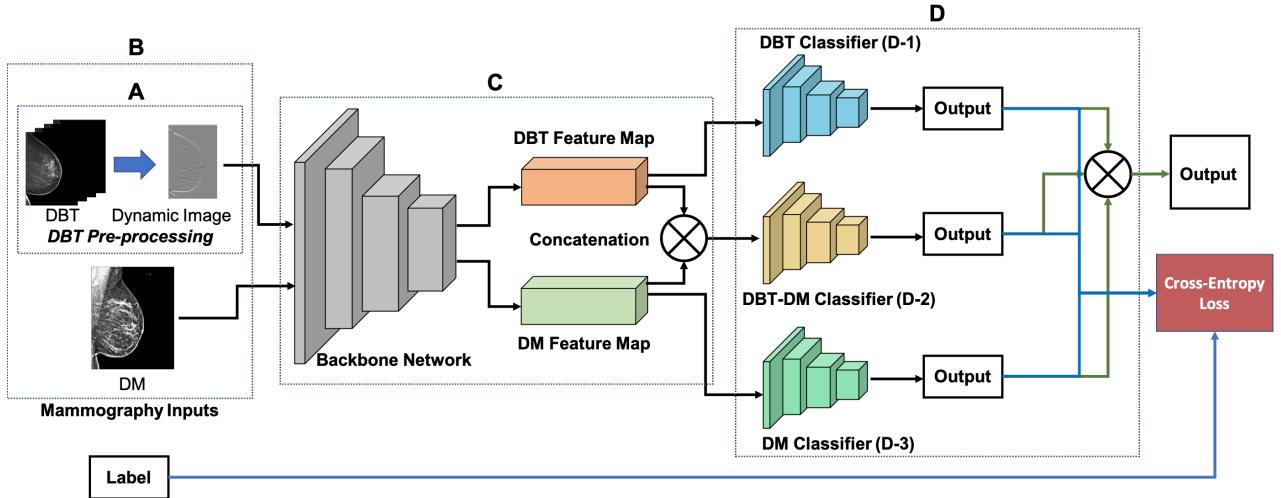


Fig. 1: Overview of the proposed model. A) DBT data pre-processing. B) DM/DBT pairs as the input of the model. C) Feature map extraction using a backbone network. D) Ensemble outputs of different CNN classifiers for testing. Blue lines, model training stage; Green lines, model testing stage; DBT, digital breast tomosynthesis; DM, digital mammogram.

breast lesions in DM and DBT. They reported 0.81 and 0.89 AUC on DM and DBT, respectively. However, the proposed work has several major limitations. For instance, a keyframe of each DBT needs to be selected by a trained radiologist during the data pre-processing step. This human involvement does not only increase the cost of using the method but also introduces bias into the proposed model. More importantly, this human pre-selection step treats DBT as an additional DM, which omits using the most important information of DBT—the slice-to-slice changes. In addition, this method also requires BBs, which are usually not available to clinical practice.

Zhang et al. [12] proposed an end-to-end breast cancer classification method using AlexNet [18] as the backbone. Their method does not require BBs for training, but it still needs to use two different models to evaluate DM and DBT separately. Though their model has some advantages over the previous ones, the model performs poorly on DBT due to the high computational cost of 3D CNN model. They only reported a 0.66 AUC on normal vs. malignant classification.

B. CNN Model for Volumetric Data

Two types of 3D CNNs are widely used for volumetric data classification. One is the fully 3D CNN architecture, such as I3D [19] and 3D-ResNet [20]. The second is to use 2D CNN models in a 3D way, such as [12]. Even though the two approaches work differently, they both suffer from the same limitations. For instance, volumetric data usually have much more extensive data size than a regular image. The average size of ImageNet data is 469×387 pixels, but the average size of DBT used in this study is $1024 \times 1024 \times 82$ voxels. Training a 3D CNN model on such a large data size is extremely computationally costly and may potentially lead to overfitting. To reduce the negative effect, [12] takes only 30 slices of each DBT as the input. However, by doing this,

either a pre-selection step is needed, or we could only hope the slices we decide to feed into the model will represent the whole volumetric data sufficiently. Neither of the scenarios is optimal. Thus, directly training a 3D CNN model for DBT may not be a good option.

III. ARCHITECTURE OVERVIEW

We propose a novel CNN ensemble method for breast tumor classification. The proposed approach consists of three main components: 1) DBT pre-processing approach (Figure 1A), 2) DBT and DM feature extraction and feature map concatenation (Figure 1C), and 3) multiple classifiers and ensemble outputs of each classifier (Figure 1D).

A. DBT Pre-processing

In non-medical domain, a popular method to represent a series of images is to apply a temporal pooling operator to the features extracted at individual images, for instance, temporal templates [21], ranking functions [22] and sub-videos [23], as well as other traditional pooling operators [24]. We adopt the idea of temporal pooling operator to the medical imaging domain. Inspired by Bilen et al., we applied RankSVM [25] directly on DBT data to extract a fixed, one-slice representation of each DBT. Since the extracted fixed representation keeps the dynamic features (i.e., the slice-to-slice changes) of DBT, we call it *dynamic feature image*. See Figure 2 for an example.

One dynamic feature image is a single RGB image, which captures the slice-to-slice changes of a DBT. A ranking function is used to obtain the dynamic feature image for a series of slices I_1, \dots, I_T , temporally. More specifically, let $\psi(I_t) \in \mathbb{R}^d$ be the feature vector extracted from each individual slices I_t in the series. Let $V_t = \frac{1}{t} \sum_{\tau=1}^t \psi(I_\tau)$ be the average time of these features up to time t . The ranking function associates

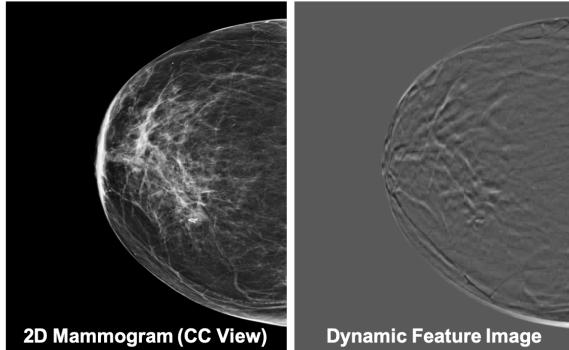


Fig. 2: Example of DM and the corresponding dynamic feature images.

to each time t a score $S(t|\mathbf{d}) = \langle \mathbf{d}, V_t \rangle$, where $\mathbf{d} \in \mathbb{R}^d$ is a vector of parameters. The function parameters \mathbf{d} are learned so that the scores reflect the rank of the slices in the series. Therefore, later times are associated with larger scores, i.e. $q \succ t \Rightarrow S(q|\mathbf{d}) > S(t|\mathbf{d})$. Learning \mathbf{d} is posed as a convex optimization problem using the RankSVM function:

$$\begin{aligned} \mathbf{d}^* &= \rho(I_1, \dots, I_T; \psi) = \operatorname{argmin}_{\mathbf{d}} E(\mathbf{d}), \\ E(\mathbf{d}) &= \frac{\lambda}{2} \|\mathbf{d}\|^2 + \frac{2}{T(T-1)} \times \\ &\quad \sum_{q>t} \max\{0, 1 - S(q|\mathbf{d}) + S(t|\mathbf{d})\}. \end{aligned} \quad (1)$$

The first term in the objective function is a quadratic regularizer used in SVM. The second term is a hinge-loss that counts how many pairs $q \succ t$ are incorrectly ranked by the scoring function. The optimizer to the RankSVM is written as a function $p(I_1, \dots, I_T; \psi)$ that maps a series of T slices to a single vector \mathbf{d}^* . Since this vector contains enough information to rank all the slices in the series, it aggregates information from all of them and can be used as a descriptor of a series of slices. The process of constructing \mathbf{d}^* is known as rank pooling [26], which can be applied to DBT directly.

B. CNN Architectures

The proposed network contains two kinds of CNNs: the backbone CNN feature extracting network (feature extractor) and the shallow CNN classifier (classifier).

1) *CNN Feature Extractor*: The feature extractor is a fully convolutional network (FCN), which takes a $W \times H \times K$ image as input and output of a $W' \times H' \times K'$ feature map. We use the common CNN classification architecture to build the feature extractor by pre-training it on ImageNet [27] dataset. After the model is well-trained, the fully connected (FC) layers of the model are removed. The pooling layer between the first FC layer and the last convolution (Conv) layer is also removed, if applicable. We use the output of the last convolutional layer of the model as the extracted feature map. All the parameters are frozen during the feature extracting step.

2) *CNN Classifier*: There are three CNN classifiers with two different architectures included in the proposed model.

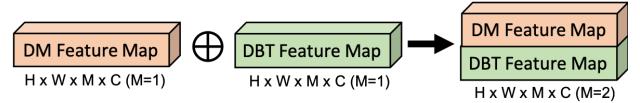


Fig. 3: Concatenate features on *modality* dimension. H =height, W =width, M =modality, C =channel.

TABLE I: Detail of CNN Classifiers.

Classifier	Input Shape	Conv Layer	Conv Type	Pooling	FC1	FC2
DBT or DM	$w \times h \times c$	$c @ 1 \times 1$	2D Conv	2×2	256	128
DM-DBT	$w \times h \times 2 \times c$	$c @ 1 \times 1 \times 2$	3D Conv	$2 \times 2 \times 1$	256	128

The DBT Classifier and DM Classifier (Figure 1D-1 and 1D-3) are used for DBT feature map classification and DM feature map classification, respectively. These two classifiers share the same architecture but with different weights, which was implemented as a 2D Conv layer followed by two FC layers. The DM-DBT Classifier (Figure 1D-2) simultaneously evaluates the DM and DBT by taking the feature maps of the two imaging modality in combination. Since we concatenated the two feature maps on the *modality* dimension, the dimensionality of the feature map is increased by 1 (see Figure 3). We replace the 2D Conv layer in the other classifiers with a 3D Conv layer. The 3D Conv kernels are applied on the *height*, *width*, and *modality* dimensions. Both of the 2D and 3D Conv layer included convolution, batch normalization, leaky ReLU, and max pooling. The batch size is 32. Max pooling has a 2×2 or $2 \times 2 \times 1$ receptive field with stride 1 for 2D or 3D Conv layer, respectively. Cross-entropy loss is used in training. Adam optimizer with a learning rate of 0.0001 is used as the optimizer. Dropout with a rate of 0.5 is applied to the FC layers. See Table I for Classifier architecture detail.

C. Classifier Ensemble

We propose to use the ensemble learning strategy to improve both the model performance and prediction confidence. In order to keep our method intuitive and straightforward, we use the majority voting strategy [28] in this study.

Suppose we have K classifiers, the majority voting can be computed as:

$$C(X) = \operatorname{argmax}_i \sum_{j=1}^K w_j I(h_j(X) = i), \quad (2)$$

where h_i is the classifier, w_i is the weights that sum to 1, and $I(\cdot)$ is an indicator function.

IV. EVALUATION

A. Dataset

A private clinical dataset is used in this study. All the DM and DBT data were retrospectively collected from patients seen at the University of Kentucky Medical Center from Jan 2014 to Dec 2017. The dataset contains 415 benign patients and 709 malignant patients. Each patient was reviewed by practicing breast radiologists. Both the benign and malignant cases were proved with a biopsy. All patients had both DM and DBT in either craniocaudal (CC) or mediolateral

oblique (MLO) view or both views. Approximately 1400 paired DM/DBT data were included. To our best knowledge, this is the largest paired DM/DBT breast cancer dataset.

The DM was provided in 12-bit DICOM format at 3328×4096 resolution. The DBT was provided in 8-bit AVI format with a resolution of 1024×1024 . All the frames of each DBT data was saved to a set of 8-bit JPEG images before generating the dynamic feature images. Both the DM and dynamic feature images were down sampled to 832×832 . Data augmentation was also applied to each of the mammography images and dynamic feature image through a combination of reflection and rotation. Each original image was flipped horizontally and rotated by each of 90, 180, and 270 degrees. In total, 6875 paired DM/DBT data were used in this study.

The dataset was randomly partitioned into training and testing datasets with a 4 : 1 ratio on the patient-level. All the images of the same patient will be in either the training set or the test set. The benign and malignant ratio was maintained in both training and testing sets. To minimize the imbalance effect (low benign to malignant ratio), we balanced each mini-batch during training.

B. Implementation and Evaluation Metrics

Four popular CNN networks were used as the backbone feature extractor in this study, namely AlexNet [18], ResNet [29], DenseNet [30], and SqueezeNet [31]. The model was implemented in Pytorch [32], and trained with balanced mini-batches for 100 epochs on a Linux computer server with eight Nvidia GTX 1080 GPU cards.

The classification accuracy (ACC), area under the receiver operating characteristic curve (AUC), precision (Prec), recall (Reca), F1 score, average precision (AP), and average correct predict confident (AC) were used as the evaluation metrics in this study.

C. Baseline Model and Ensemble Approach

We use the 2D-T3-Alex and 3D-T2-Alex models from [12] as the baseline model for DM and DBT, respectively. The 2D-T3-Alex model is a transfer learning 2D CNN model, which uses pre-trained AlexNet to extract features. The 3D-T2-Alex model is a 3D CNN model, which firstly uses the regular AlexNet model to extract feature maps of every slice in a DBT. Then, K feature maps of each DBT are fed into a one-Conv-layer 3D CNN model for classification. $K = 30$ was chosen in their paper.

Our experiment shows the proposed model significantly improves the performance. By only using DBT data (i.e., the dynamic feature images), the performance can be improved from 0.72 AUC to 0.89 AUC (23.61% increasing). When using DM and DBT in combination, a single model can achieve 0.95 AUC. After assembling the three classifiers (DM Classifier, DBT Classifier, and DM-DBT Classifier, which uses DM only, DBT only, and DM and DBT data, respectively), the proposed model can further improve the performance to 0.97 AUC (Table II).

TABLE II: Ensemble results for different backbone networks.

Model	Input Data	Backbone Network	AUC
<i>2D – T3 – Alex</i>	DM only	AlexNet	0.87
<i>3D – T2 – Alex</i>	DBT only	AlexNet	0.72
<i>Ours_{DBT}</i>	DBT only	AlexNet	0.89
<i>Ours_{DM – DBT}</i>	DM & DBT	AlexNet	0.95
<i>Ours_{Ensemble}</i>	DM & DBT	AlexNet	0.97

TABLE III: Comparing with baseline model.

Backbone Network	Input Data	AUC
AlexNet	DM & DBT	0.97
ResNet	DM & DBT	0.96
DenseNet	DM & DBT	0.97
SqueezeNet	DM & DBT	0.97

Table III lists the ensemble result of all different backbone networks. The performance is consistency among the four different feature extractors, which indicates the proposed method is not limited to any specific architecture.

D. Single Modality vs. Multiple Modalities

In this section, we evaluate the model performance using a single imaging modality vs. multiple imaging modalities. More specifically, we are comparing the performance of DM Classifier, DBT Classifier, and DM-DBT Classifier. Four different backbone networks were used. In total, 12 models were trained and compared in this experiment.

Table IV reveals when using multiple imaging modalities together, the model performance is significantly better. The DM-DBT Classifier achieves a 0.95 AUC on average. However, the save metric for DM Classifier and DBT Classifier is 0.88 and 0.89, respectively. The table also shows when using DBT data, the model prediction confidence can be improved, especially when using DM and DBT in combination. On average, the prediction confidence of DM Classifier is 0.83, the same metric of DBT Classifier and DM-DBT Classifier is 0.89 and 0.93, respectively. As in the previous section, the performance of all four different backbone networks is consistent. They all achieved a similar result, except the average prediction confidence of single modality classifiers (i.e., DM Classifier and DBT Classifier). Among the four backbone networks, the DenseNet performance is slightly better than others, which achieves the highest scores of 17 out of 21 different metrics for different classifiers.

V. CONCLUSION

We propose a novel deep learning ensemble model for breast lesion classification, which simultaneously uses digital mammograms (DM) and digital breast tomosynthesis (DBT). We innovatively use the RankSVM algorithm on DBT to extract a fixed representation, dynamic feature image, of DBT. Dynamic feature image captures the slice-to-slice difference in DBT, which is the information often looked by radiologists. The experiments show that when using both DM and DBT in combination, the single model performance can be improved nearly 10% on AUC and 23% on the prediction confidence. By applying ensemble strategy on the three classifiers, the best

TABLE IV: Evaluation results of models trained with a single modality vs. models trained with the multiple modalities.

Backbone Network	DM Classifier							DBT Classifier							DM-DBT Classifier						
	ACC	AUC	F1	Prec	Reca	AP	AC	ACC	AUC	F1	Prec	Reca	AP	AC	ACC	AUC	F1	Prec	Reca	AP	AC
AlexNet	0.78	0.87	0.76	0.87	0.75	0.70	0.78	0.81	0.89	0.80	0.84	0.76	0.76	0.83	0.90	0.95	0.89	0.91	0.87	0.86	0.83
ResNet	0.78	0.87	0.79	0.75	0.83	0.71	0.96	0.79	0.88	0.79	0.80	0.78	0.73	0.97	0.87	0.94	0.87	0.82	0.93	0.80	0.96
DenseNet	0.79	0.90	0.80	0.76	0.83	0.72	0.79	0.85	0.91	0.85	0.86	0.84	0.80	0.97	0.91	0.96	0.91	0.93	0.89	0.88	0.96
SqueezeNet	0.78	0.88	0.78	0.78	0.79	0.71	0.80	0.79	0.89	0.78	0.85	0.72	0.75	0.79	0.90	0.96	0.91	0.93	0.88	0.88	0.97
Average	0.78	0.88	0.78	0.79	0.80	0.71	0.83	0.81	0.89	0.81	0.84	0.78	0.76	0.89	0.89	0.95	0.90	0.90	0.90	0.86	0.93

performance can be improved to 0.97 AUC. This improvement indicates that deep learning models, like radiologists, benefit from combining both mammographic image formats. Also, the consistency of better performance across different feature extractors and classifiers suggests that our method is not limited to any specific deep learning architecture. The proposed DBT data representation method and dynamic feature image can also increase the classification performance of using DBT-only data by nearly 24%. In addition, our approach uses only the image-level labels. Due to a large number of incoming data in the daily clinical practice, annotating images with bounding boxes is not practical. However, we believe that with more precise labels, such as bounding boxes, the performance of our model can be further improved. Our model can adapt to bounding boxes labeling with minor changes.

REFERENCES

- [1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.
- [2] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2019,” *CA: a cancer journal for clinicians*, vol. 69, no. 1, pp. 7–34, 2019.
- [3] P. Henrot, A. Leroux, C. Barlier, and P. Génin, “Breast microcalcifications: the lesions in anatomical pathology,” *Diagnostic and interventional imaging*, vol. 95, no. 2, pp. 141–152, 2014.
- [4] “Breast cancer screening guidelines.” [Online]. Available: <https://www.cancer.org/health-care-professionals/american-cancer-society-prevention-early-detection-guidelines/breast-cancer-screening-guidelines.html>
- [5] S. M. Friedewald, E. A. Rafferty, S. L. Rose, M. A. Durand, D. M. Plecha, J. S. Greenberg, M. K. Hayes, D. S. Copit, K. L. Carlson, T. M. Cink *et al.*, “Breast cancer screening using tomosynthesis in combination with digital mammography,” *Jama*, vol. 311, no. 24, pp. 2499–2507, 2014.
- [6] M. L. Giger, N. Karssemeijer, and J. A. Schnabel, “Breast image analysis for risk assessment, detection, diagnosis, and treatment of cancer,” *Annual review of biomedical engineering*, vol. 15, pp. 327–357, 2013.
- [7] P. Wing and M. H. Langteller, “Workforce shortages in breast imaging: impact on mammography utilization,” *American Journal of Roentgenology*, vol. 192, no. 2, pp. 370–378, 2009.
- [8] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, p. 115, 2017.
- [9] R. P. Mihail, G. Liang, and N. Jacobs, “Automatic hand skeletal shape estimation from radiographs,” *IEEE Transactions on NanoBioscience*, vol. 18, no. 3, pp. 296–305, July 2019.
- [10] G. Liang, S. Fouladvand, J. Zhang, M. A. Brooks, N. Jacobs, and J. Chen, “Ganai: Standardizing ct images using generative adversarial network with alternative improvement,” *bioRxiv*, p. 460188, 2018.
- [11] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, “Detecting and classifying lesions in mammograms with deep learning,” *Scientific reports*, vol. 8, no. 1, p. 4165, 2018.
- [12] X. Zhang, Y. Zhang, E. Y. Han, N. Jacobs, X. Han, and J. Liu, “Classification of whole mammogram and tomosynthesis images using deep convolutional neural networks,” *IEEE transactions on nanobioscience*, vol. 17, no. 3, pp. 237–242, 2018.
- [13] K. Mendel, H. Li, D. Sheth, and M. Giger, “Transfer learning from convolutional neural networks for computer-aided diagnosis: a comparison of digital breast tomosynthesis and full-field digital mammography,” *Academic radiology*, vol. 26, no. 6, pp. 735–743, 2019.
- [14] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, 2015, pp. 91–99.
- [15] “The digital mammography dream challenge,” <https://www.synapse.org/#/Synapse:syn4224222/wiki/401743>, Nov. 2016.
- [16] L. Shen, “End-to-end training for whole image breast cancer diagnosis using an all convolutional design,” *arXiv preprint arXiv:1708.09427*, 2017.
- [17] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [19] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.
- [20] K. Hara, H. Kataoka, and Y. Satoh, “Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.
- [21] A. F. Bobick and J. W. Davis, “The recognition of human movement using temporal templates,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 3, pp. 257–267, 2001.
- [22] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars, “Modeling video evolution for action recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5378–5387.
- [23] M. Hoai and A. Zisserman, “Improving human action recognition using score distribution and ranking,” in *Asian conference on computer vision*. Springer, 2014, pp. 3–20.
- [24] M. S. Ryoo, B. Rothrock, and L. Matthies, “Pooled motion features for first-person videos,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 896–904.
- [25] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, “Dynamic image networks for action recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3034–3042.
- [26] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars, “Rank pooling for action recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 773–787, 2016.
- [27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.
- [28] G. James, “Majority vote classifiers: theory and applications,” 1998.
- [29] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [31] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and $\frac{1}{10}$ mb model size,” *arXiv preprint arXiv:1602.07360*, 2016.
- [32] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” 2017.