Journal of Biomedical Informatics 113 (2021) 103656

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Comparison of segmentation-free and segmentation-dependent computer-aided diagnosis of breast masses on a public mammography dataset

Rebecca Sawyer Lee [a,1], Jared A. Dunnmon [b,1,*], Ann He [b], Siyi Tang [c], Christopher Ré [b], Daniel L. Rubin [d]

[a] Stanford University Biomedical Informatics Training Program, United States
[b] Stanford University Department of Computer Science, United States
[c] Stanford University Department of Electrical Engineering, United States
[d] Stanford University Departments of Radiology and Biomedical Data Science, United States

ARTICLE INFO

Keywords:
Mammography
Computer assisted diagnosis
Deep learning
Segmentation

ABSTRACT

Purpose: To compare machine learning methods for classifying mass lesions on mammography images that use predefined image features computed over lesion segmentations to those that leverage segmentation-free representation learning on a standard, public evaluation dataset.

Methods: We apply several classification algorithms to the public Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM), in which each image contains a mass lesion. Segmentation-free representation learning techniques for classifying lesions as benign or malignant include both a Bag-of-Visual-Words (BoVW) method and a Convolutional Neural Network (CNN). We compare classification performance of these techniques to that obtained using two different segmentation-dependent approaches from the literature that rely on specific combinations of end classifiers (e.g. linear discriminant analysis, neural networks) and predefined features computed over the lesion segmentation (e.g. spiculation measure, morphological characteristics, intensity metrics).

Results: We report area under the receiver operating characteristic curve (A_Z) values for malignancy classification on CBIS-DDSM for each technique. We find average A_Z values of 0.73 for a segmentation-free BoVW method, 0.86 for a segmentation-free CNN method, 0.75 for a segmentation-dependent linear discriminant analysis of Rubber-Band Straightening Transform features, and 0.58 for a hybrid rule-based neural network classification using a small number of hand-designed features.

Conclusions: We find that malignancy classification performance on the CBIS-DDSM dataset using segmentation-free BoVW features is comparable to that of the best segmentation-dependent methods we study, but also observe that a common segmentation-free CNN model substantially and significantly outperforms each of these ($p < 0.05$). These results reinforce recent findings suggesting that representation learning techniques such as BoVW and CNNs are advantageous for mammogram analysis because they do not require lesion segmentation, the quality and specific characteristics of which can vary substantially across datasets. We further observe that segmentation-dependent methods achieve performance levels on CBIS-DDSM inferior to those achieved on the original evaluation datasets reported in the literature. Each of these findings reinforces the need for standardization of datasets, segmentation techniques, and model implementations in performance assessments of automated classifiers for medical imaging.

* Corresponding author at: 353 Jane Stanford Way, Stanford, CA, 94305, United States.
E-mail address: jdunnmon@cs.stanford.edu (J.A. Dunnmon).
[1] Equal contributions.

https://doi.org/10.1016/j.jbi.2020.103656
Received 5 July 2020; Received in revised form 9 November 2020; Accepted 7 December 2020
Available online 11 December 2020

1. Introduction

Breast cancer is the most deadly cancer for women in developing countries and the second most deadly cancer for those in developed nations [1]. Mammograms are an essential component in early detection of breast cancer, and interpretation sensitivity greatly affects patient survival rates [2]. On the other hand, imperfect specificity of breast lesion diagnosis by mammography causes physical and psychological discomfort to false positive patients who are subjected to further, possibly invasive tests [3,4]. As a result, a variety of Computer Aided Diagnosis (CADx) systems designed to provide quantitative, objective mass classification have been developed [5].

Existing CADx techniques for mass classification generally fall into two categories: *segmentation-dependent* methods, which require a detailed outline of the lesion upon which to compute features for classification, and *segmentation-free* methods, which do not. Segmentation-dependent methods tend to rely on predefined sets of features, while segmentation-free approaches leverage representation learning techniques to learn explanatory features directly from the available data. Segmentation-free approaches could provide substantial value in clinical practice by obviating the need for detailed segmentations, but have only recently been able to achieve results similar to those of segmentation-dependent methods [6–9]. Though recent studies of segmentation-free approaches have shown promising results, CADx systems for mass classification have rarely been evaluated on the same datasets, making the type of comparative performance analysis that should precede clinical deployment difficult to perform [8]. Major limitations to the evaluation of different mammography CADx systems on common datasets have ranged from insufficient descriptions of model implementation in the original literature to challenges with data availability and provenance.

In this work, we leverage the recent Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) to compare segmentation-free and segmentation-dependent approaches to automated mass classification on a standard, public dataset. To perform this comparison, we implement four different techniques from the literature: a segmentation-free Bag-of-Visual-Words (BoVW) mass classification algorithm inspired by traditional computer vision [10], a segmentation-free Convolutional Neural Network (CNN) trained using commodity deep learning software [11], a segmentation-dependent algorithm based on the Rubber-Band Straightening Transform (RBST) of Sahiner et al. [6], and the segmentation-dependent approach of Huo et al. that combines predefined features with an artificial neural network [7]. Evaluating the performance of each of these techniques on the CBIS-DDSM test set after training and tuning was performed using the standard CBIS-DDSM training set yields several useful conclusions. First, we observe that malignancy classification performance on the CBIS-DDSM dataset obtained using the segmentation-free BoVW method is comparable to that of the best segmentation-dependent methods, but also find that the segmentation-free CNN model substantially and significantly ($p < 0.05$) outperforms each of these by 11 points of area under the receiver operating characteristic curve (A_Z). These results reinforce recent findings suggesting that representation learning techniques such as BoVW and CNN can obviate the need for precise or method-specific lesion segmentation while maintaining high levels of performance. Second, we find that our re-implementations of existing segmentation-free methods yield performance levels on CBIS-DDSM inferior to those achieved on the original evaluation datasets reported in the literature. We propose that these discrepancies result from some combination of differences in the segmentation techniques used, parameter tuning on small datasets in the original work, and implementation choices.

2. Background and related work

Due to the importance of characteristics of the lesion margin in differentiating benign and malignant tumors, many existing CADx methods have been based on obtaining mathematical descriptions of the tumor outline [7,12–20]. Such *segmentation-dependent* techniques require accurate segmentation of the lesion margin in order to extract image features. Methods that require hand-drawn segmentation of lesions, a process not usually performed in clinical practice, can make resultant CADx systems inefficient for clinical use. CADx systems utilizing automated lesion segmentation have, however, been studied with promising results. For example, Mudigonda et al. obtained an A_Z of 0.85 for binary classification of breast masses using hand-drawn Regions-of-Interest (ROIs) as a basis for their automated segmentation method [13]. Likewise, Sahiner et al. [14] and Huo et al. [7] developed CADx systems using segmentation methods requiring only a general bounded region identified by the radiologist and achieved A_Z results of 0.91 and 0.94, respectively. An extensive analysis of existing methods for automated mass segmentation methods that support segmentation-dependent CADx can be found in the review of Oliver et al. [21].

More recent *segmentation-free* CADx approaches have attempted to attain high levels of diagnostic performance without any lesion segmentation requirements. For instance, multiple workers such as Jamieson et al. [22] ($A_Z = 0.71$), Liu and Jiang [23] ($A_Z = 0.72$), and Li et al. [24] ($A_Z = 0.796$) have demonstrated the promise of segmentation-free algorithms, most of which have been based on various types of learned features. A particularly large body of work has arisen that applies deep learning algorithms such as Convolutional Neural Networks (CNNs) to automated mass analysis. The recent review of deep learning techniques in mammography by Hamidinekoo et al. found that nine major mass classification studies performed between 1996 and 2017 using deep learning approaches yielded A_Z values between 0.71 and 0.97 [9]. However, because each of these studies used different datasets, several of which are not public, it is difficult to fairly compare the performance of these different algorithms.

The focus of this work is on analyzing the performance of multiple different techniques on the relatively recent CBIS-DDSM dataset. We leverage this dataset because it (a) is provided with standard train/validation splits (b) contains cases verified by pathology (c) contains images curated by expert mammographers from multiple institutions and (d) contains a large amount of data in addition to the raw images – e.g. lesion segmentation, Breast Imaging Reporting And Data System (BI-RADS) descriptors, BI-RADS abnormality ratings, and severity ratings – that make it amenable to analysis via a wide variety of machine learning approaches. Note that BI-RADS describes both a classification system and lexicon for reporting breast imaging results that includes both descriptors of imaging features that were shown to correlate with high predictive values associated with either benign or malignant disease, and a classification system to describe the likelihood that the imaging findings represent malignancy [25].

For the sake of completeness, we provide additional background on recent work in mass classification and the datasets on which these studies were evaluated here. Wang et al. extracted five manually-designed image features and applied logistic regression for classification, achieving A_Z of 0.806 $\pm$ 0.025 on an internal dataset [26]. Zhu et al. developed a deep multi-instance network and achieved A_Z of 0.859 $\pm$ 0.03 on the public INbreast dataset [27]. Arevalo et al. used a combination of CNN-extracted features and hand-crafted features and obtained A_Z of 0.826 on the public BCDR-FM dataset [28]. Kim et al. applied ResNet to a large-scale mammography dataset from 5 institutions, and achieved A_Z of 0.906 [29]. Ribli et al. applied Faster R-CNN to the INbreast database, which achieved A_Z of 0.95 [30]. Al-masni et al. applied a YOLO model for this task and achieved 5-fold cross-validation A_Z of 0.965 on a subset of the DDSM dataset [31]. Lotter et al. developed a curriculum training method with a multi-scale CNN and obtained 0.901 $\pm$ 0.031 on a subset of DDSM; these authors also note that "as full image mammogram classification lacks standardized evaluation framework, it is somewhat difficult to directly compare our results to other work." [32]. In other words, because most studies are evaluated on different, incompletely reported, or private datasets, it is

difficult even for researchers in the field to understand how the performance of their methods compares to that reported by other work. Ting et al. used hierarchical features from a CNN and achieved AUC 0.92 ± 0.02 on a subset of DDSM [33]. Chougrad et al. applied various CNNs including VGG, ResNet-50 and Inception networks, and achieved A_Z of 0.99 on the MIAS dataset and 0.98 on a subset of the DDSM dataset [34].

Many studies have also used the CBIS-DDSM dataset we leverage here, albeit in slightly different ways. Ragab et al. [35] and Li et al. [36] applied convolutional neural networks on segmented ROIs, and achieved A_Z of 0.94 and 0.85 on CBIS-DDSM respectively. Tsochatzidis et al. examined multiple popular CNNs, and achieved A_Z of 0.859 and 0.804 on DDSM-400 and CBIS-DDSM, respectively [37]. Chougrad et al. proposed a multi-label classification setting and fine-tuned a pre-trained CNN, achieving mean A_Z of 0.89 ± 0.08 for 5-fold cross-validation on CBIS-DDSM [38]. Chen et al. [39] and Falconi et al. [40] fine-tuned CNNs and achieved A_Z of 0.86 and 0.844 on CBIS-DDSM respectively. Alkhaleefah et al. fine-tuned VGG-19 and applied data augmentation techniques on their own data splits of CBIS-DDSM, and achieved A_Z of 0.961 [41]. Shu et al. proposed different pooling structures for CNNs and obtained A_Z of 0.838 ± 0.0001 on CBIS-DDSM [42]. Samala et al. aimed to assess the generalization errors of CNNs, and found A_Z of 0.83 ± 0.03 on internal and CBIS-DDSM combined data [43]. Gossmann et al. investigated the performance deterioration of deep neural networks for lesion classification due to distribution shift, and achieved A_Z of 0.833 on CBIS-DDSM [44]. Beltran-Perez et al. developed a three-step pipeline to extract image features using a multiscale generalized radial basis function and the discrete cosine transform, and achieved 93.99% accuracy on CBIS-DDSM [45]. Ansar et al. applied transfer learning of MobileNet and obtained 74.5% accuracy on CBIS-DDSM [46]. De Vriendt et al. proposed an all-in-one graph-based deep semi-supervised learning framework, and obtained A_Z of 0.811 with only 40% of the labeled data on CBIS-DDSM [47].

A small number of existing studies compare multiple computer assisted detection or diagnosis techniques on the same datasets. The work of Oliver et al. implements multiple approaches for mass detection techniques on a single dataset, but does not address mass classification [21]. The recent work of Kooi et al. compares a CNN-based approach to a single reference system based on extracted features, but does so on a non-public dataset [8]. They find that the segmentation-free CNN approach (A_Z = 0.93) slightly outperforms their segmentation-dependent, feature-based method (A_Z = 0.91).

3. Materials and methods

In order to compare multiple segmentation-free representation learning approaches to several segmentation-dependent predefined feature methods using a standard, public dataset, we implemented four high-performing methods from the literature that are both trained and evaluated using standard splits from the CBIS-DDSM mass classification dataset [48]. Segmentation-free techniques analyzed include both a BoVW approach adapted from traditional computer vision as well as a standard CNN-based approach from the field of deep learning [10,11]. Segmentation-dependent techniques analyzed include two high-performing approaches from the literature: linear discriminant analysis of features computed on the lesion margin as described by Sahiner et al. [6] and hybrid rule-based neural network classification of a small number of salient features as described by Huo et al. [7] We obtained existing code for each technique where possible, and re-implemented the remainder as faithfully as possible following descriptions in the literature; specific implementation decisions are described in detail below. We note that it was not possible to obtain code for the exact methods for lesion segmentation implemented in previous studies, and we therefore use the segmentations provided as part of the public CBIS-DDSM dataset. The technique used by Lee et al. [49] to compute these segmentations was based on the Chan-Vese local level set framework [50], where the coarse annotation from the original DDSM dataset was used for initialization. The process of computing and validating these segmentations is described in detail by Lee et al. [49] We describe each method implemented for this study in detail below; implementations are built using MATLAB (v. R2011) and Python 3.6 unless otherwise noted. Note that each method uses only the mammogram image, and no other descriptors included in the CBIS-DDSM dataset.

3.1. Segmentation-Free representation learning methods

Each of the segmentation-free methods described below relies on classification of a set of features learned directly from training data, and does not require lesion segmentation.

3.1.1. Bag-of-visual words

Fig. 1 outlines the procedure for the BoVW method. Each image was first preprocessed using the method developed by Chan et al., which involves filtering a bounded portion of the image around the ROI in order to smooth out structures in the background tissue that may obscure the mass margin [14,51]. Fig. 2 shows an example of an ROI before and after this preprocessing. We then utilize the Scale-Invariant Feature Transform (SIFT) to compute the primitive feature set on which our BoVW method is based using a bounding box around the entire ROI; all SIFT features were computed using the VLFeat open source library in MATLAB (v. R2011) [52]. Subsequent image classification is based on a feature histogram created by counting the number of image patches assigned to each individual visual word by a consensus clustering approach described in detail in the Supplementary Material. Finally, we train an L_1-regularized logistic regression classifier, also known as Least Absolute Shrinkage and Selection Operator (LASSO), utilizing the glmnet (v. 2.0–16) software package[53]. Hyperparameters, such as the number of clusterings upon which to evaluate consensus clustering and the regularization parameter for LASSO, were tuned using 20% of the training data as a held-out validation set.

3.1.2. Convolutional neural network

We evaluate the performance of a standard Convolutional Neural Network (CNN) architecture implemented using the Keras (v. 2.2.0) Python software package with a Tensorflow backend [54–56]. Due to its superior performance on a variety of image classification tasks both inside and outside of medicine, we utilize a 121-layer Densely Connected CNN (DenseNet-121) of Huang et al. for this comparison [11,57,58]. The DenseNet-121 model was trained on a random 90% sample of the DDSM-CBIS training set and validated on the remaining

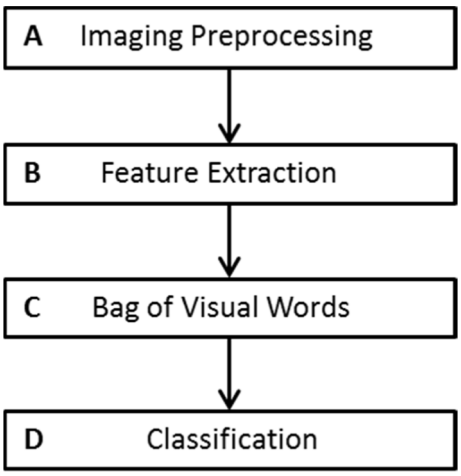

Fig. 1. Flowchart of a Bag-of-Visual-Words (BoVW) method, which relies on sequential steps including image processing, feature extraction, formation of visual words dictionary, and ultimate classification.

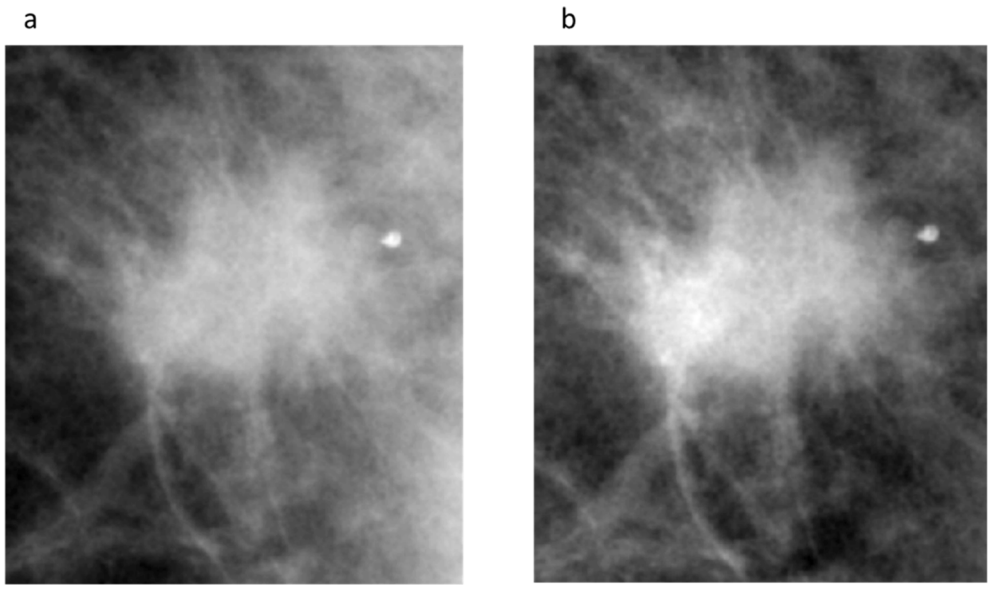

Fig. 2. Example of a Region-of-Interest (ROI) before (a) and after (b) preprocessing for the Bag-of-Visual-Words (BoVW) classifier.

10% for 100 epochs using a single Tesla P100 GPU with a batch size of 32 images, a dropout rate of 0.2, a learning rate of 0.001, and the Adam optimizer. Hyperparameter values were determined using coarse grid search in the vicinity of default parameters. The network was initialized using weights from a model pre-trained on the ImageNet dataset, all parameters were assumed learnable, and the learning rate was decreased by a factor of $\sqrt{0.1}$when validation accuracy had not increased for more than 10 epochs [57]. All images were mean-standard-deviation-normalized, cropped to 750 × 750 pixels around the segmentation centroid, and further downsampled to 224 × 224 pixels before entering into the CNN. Data augmentation was achieved via random application of mild zooms (in the range of [0.8, 1.2]), horizontal flips, and random rotations. While such augmentations are important in most state-of-the art image classification results, note that due to the physical particulars of mammography, augmentations such as contrast and brightness enhancements were specifically not applied [59].

3.2. Segmentation-dependent predefined feature methods

Each of the segmentation-dependent methods described below is drawn from the CADx literature, and relies on classification of a set of features defined *a priori* over a provided lesion segmentation.

3.2.1. Linear discriminant analysis of Rubber-Band Straightening Transform features (LDA-RBST)

The CADx system of Sahiner et al. relies on linear discriminant analysis performed on a variety of predefined features, including those derived using the Rubber-Band Straightening Transform (RBST) for which Sahiner has generously provided the code [6]. The RBST converts the margin of the image into a straight line as shown in Fig. 3, and is accomplished by determining the normal direction to each ROI edge pixel and taking 40 pixels along that direction to compose each line of the RBST. Several texture features were computed from the RBST image, including features from the gray-level co-occurrence matrix (GLCM) at ten pixel differences ($d = 1,2,3,4,6,8,10,12,16,20$) and four directions ($\theta = 0^{\circ}, 45^{\circ}, 90^{\circ}, 135^{\circ}$) as well as run-length statistics (RLS) features in four directions ($\theta = 0^{\circ}, 45^{\circ}, 90^{\circ}, 135^{\circ}$). Additionally, various morphological features were obtained from the original ROI image. Each set of features is listed in Table 1.

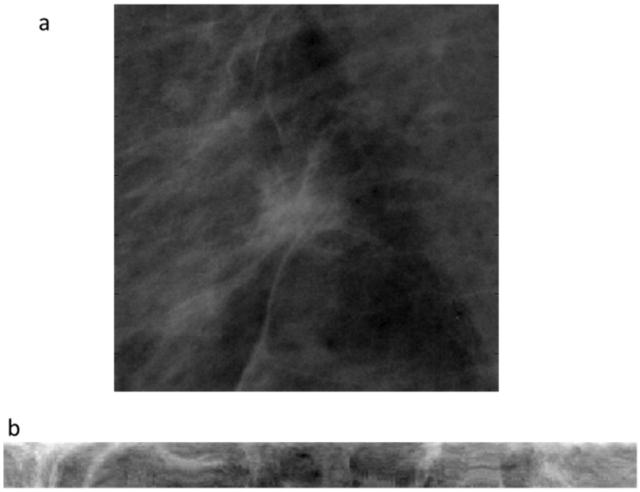

Fig. 3. Example Region-of-Interest (ROI) and resulting Rubber-Band Straightening Transform (RBST) image.

Table 1
List of features employed by method of Sahiner et al.

Morphological	GLCM	RLS
Fourier descriptor	Difference average	Long run emphasis
Convexity	Difference entropy	Run percentage
Rectangularity	Inverse difference moment	Gray level non-uniformity
Perimeter	Difference variance	Run length non-uniformity
NRL mean	Inertia	Short run emphasis
Contrast	Correlation	
NRL entropy	Inf. Measure of correlation 1	
Circularity	Inf. Measure of correlation 2	
NRL area ratio	Energy	
NRL standard deviation	Entropy	
NRL zero-crossing count	Sum variance	
Perimeter-to-area ratio	Sum entropy	
Area	Sum average	

3.2.2. Neural Network Classification of Hand-Designed Features (NN-HDF)

Finally, we re-implemented the feature extraction method from Huo et al. that supports a hybrid rule-based neural network classifier [7]. Five features were included in this method: spiculation measure, sharpness, average gray level, contrast, and texture. The spiculation measure was the key feature for the original work of Huo et al. [7]. It is found using radial edge gradient analysis over the four different neighborhoods shown in Fig. 4, where Sobel filters were used to obtain the gradient magnitude and orientation of the ROI [60]. The orientation was then normalized based on the radial direction, and a gradient-magnitude-weighted histogram of the normalized orientation was found. The spiculation measure is the Full-Width at Half Maximum (FWHM) of this histogram. As Huo et al. do not describe the exact method for determining the FWHM required for their method, we utilize the following straightforward computation. We first smooth the histogram with an averaging filter of length two. We then find the first and last bins at which the histogram exceeded the half maximum. Finally, we use linear interpolation to determine the exact bin position and convert to degrees. We use 24 bins, allowing for a bin size of 15° as in the original paper. We note here that Huo et al. report that only an approximately correct outline of the mass lesion was required for the purposes of this analysis.

3.3. Evaluation

Each method was trained (where appropriate) and tested using the CBIS-DDSM data set with the provided train and test splits [48]. The data set includes 691 training cases (355 benign, 336 malignant) and 200 test cases (117 benign, 83 malignant). We assess the performance of each method by analyzing A_Z values and associated 95% confidence intervals [61].

3.4. Statistical techniques

Model performance was assessed using A_Z on a held-out test set, computed using either the scikit-learn (v 0.19) Python library or MATLAB (v. R2011). A statistical test of non-inferiority implemented in the rocNIT (v. 1.0) R library was used to compare different classifiers characterized by similar performance levels. The method of Hanley and McNeil was used to compute 95% confidence intervals on A_Z values, and p-values less than 0.05 were considered statistically significant throughout the analysis [62]. Statistical computations performed by J.A. D., A.H., and R.S.L..

4. Results

Table 2 contains the results for each method described above on

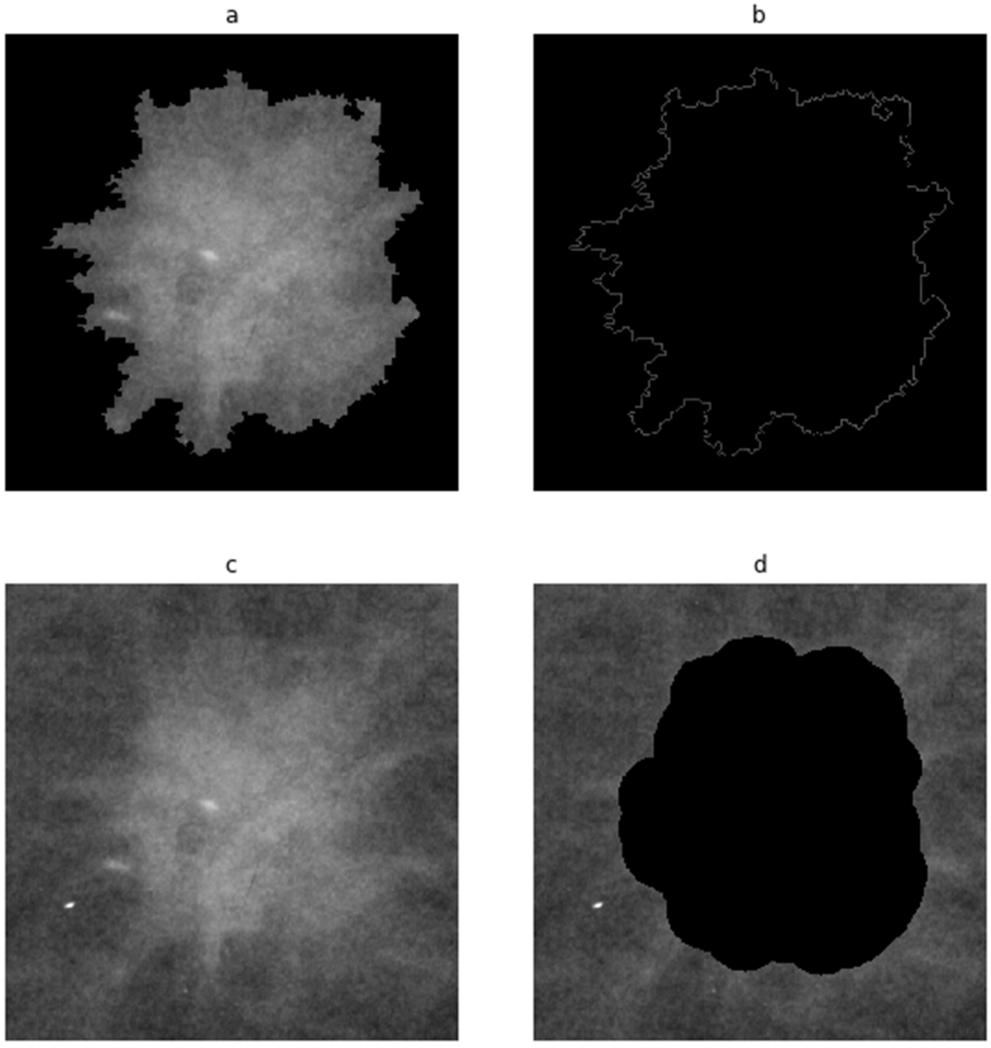

Fig. 4. Neighborhoods used in NN-HDF method (excluded areas in black). Panel (a) represents the segmented mass, panel (b) represents the mass margin, panel (c) represents the mass plus the surrounding periphery, and panel (d) represents the surrounding periphery.

Table 2
Results for different classification methods on CBIS-DDSM dataset.

Method	A_Z [95% Confidence Interval]
BoVW	0.73 [0.66, 0.79]
CNN	**0.86 [0.83, 0.89]**
LDA-RBST	0.75 [0.69, 0.81]
NN-HDF	0.58 [0.51, 0.65]

CBIS-DDSM. Most noticeably, the CNN substantially and significantly outperforms all other approaches with an A_Z of 0.86 [0.83, 0.89] ($p < 0.05$); note that these results are on par with the best segmentation-free results of which we are aware on the standard CBIS-DDSM dataset. Amongst the remaining methods, the LDA-RBST technique of Sahiner et al. yielded the best classification performance results with an A_Z of 0.75 [0.69, 0.81], followed closely by the BoVW method with an A_Z of 0.73 [0.66, 0.79]. A non-inferiority test performed on the results from these two techniques results in a p-value of 0.01 for a δ of 0.15 and an α of 0.05, demonstrating significant non-inferiority of the BoVW method with respect to that LDA-RBST [63]. The NN-HDF method performed significantly worse than all other methods on CBIS-DDSM, with an A_Z value of 0.58 [0.51, 0.65] ($p < 0.05$). We present additional results specific to each technique in detail below.

4.1. Segmentation-free representation learning methods

4.1.1. Bag-of-visual-words

As described in the Supplementary Material [65–68], the BoVW method requires parameter optimization with respect to the number of clusterings used in consensus clustering and the regularization parameter in LASSO. Based on the empirical findings shown in Fig. 5, we chose 10 clusterings and a regularization parameter value of 0.014. With these optimized parameters, we achieved A_Z of 0.73 using the BoVW method.

4.1.2. Convolutional neural network

The fundamental difference between the DenseNet-121 deep learning approach and other methods in this paper is the fact that deep neural networks are able to learn their own feature maps in a manner that best explains the data available. Thus, the trained neural network is itself a feature extractor, and the combination of a fully connected linear layer and a softmax operator is responsible for classification. Performing the training procedure described above with ten different random seeds yielded best-case test set A_Z of 0.88 [0.85, 0.90], worst-case test set A_Z of 0.85 [0.82, 0.89], and median A_Z of 0.86 [0.83 0.89], which represents the best performance of any method described in this manuscript.

In order to ensure that this classification performance is not caused by anomalies within the data or training process, we compute class activation maps (CAMs) to assess whether the CNN classifications are leveraging appropriate spatial regions of the image [64]. Pertinent visualizations can be found in Fig. 6, where we observe that correct classifications result when the network activations are directly over the lesion, while errors in both directions occur when weights for the correct class are high in spatial areas outside of the mass itself.

4.2. Segmentation-dependent predefined feature methods

4.2.1. Linear discriminant analysis of rubber-band straightening transform features

Our analysis is similar to that of Sahiner et al. [6], where we choose the top ten features from morphological and texture features separately as well as together for use in linear discriminant analysis. Feature selection was accomplished using Wilks' lambda. The resulting A_Z from each of these scenarios was 0.62, 0.70, and 0.75, respectively. Table 3 lists the top ten features found for each category. Note that Sahiner et al. [6] specifically mention that they expect the selected features to change based on the particular training dataset used, so our procedure represents an intended implementation of this technique.

4.2.2. Neural network classification of hand-designed features

Huo et al. developed a hybrid rules-based neural network classifier in their work [7]. The rule pertains to the spiculation measure, automatically concluding that any mass with a spiculation measure higher than 160° was malignant, and using the rest of the features as input to a neural network to determine the malignancy of the lesions with lower spiculation measure. The results of our re-implementation of this method evaluated on CBIS-DDSM are shown per feature in Table 4 along with the results reported by Huo et al. on their original dataset. The resulting A_Z for the hybrid classifier using the 160° threshold was 0.51. With a threshold of 320° optimized for our dataset, the A_Z was 0.58.

5. Discussion

Our analysis supports several distinct conclusions. First, we find that the two segmentation-free methods are able to classify benign vs. malignant masses as well as or better than segmentation-based methods that use predefined features. In particular, while BoVW performs similarly to the best predefined feature method, the deep learning method ($A_Z = 0.86$) improves A_Z by 11 points over the best competing segmentation-based method ($A_Z = 0.75$). These results support the conclusion that the two segmentation-free mass classification methods that leverage representation learning, BoVW and CNN, can obviate the need for accurate segmentations while improving performance with respect to traditional segmentation-based CADx methods that use predefined features. While our observation that segmentation-free deep learning models can outperform segmentation-based models is consistent with expectations of previous work [8], our study is the first to demonstrate this across multiple different techniques on a standard, public mammography dataset. A robust finding that mammography

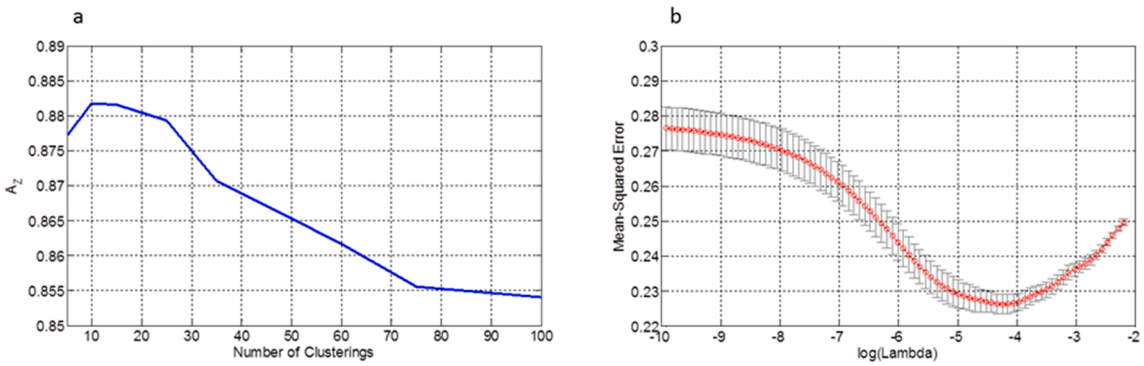

Fig. 5. Tuning rresults per parameter using Bag-of-Visual-Words (BoVW) method: (a) area under the receiver operating characteristic curve (A_Z) versus number of clusterings using Scale Invariant Feature Transform (SIFT) features, (b) mean-squared error versus logarithm of regularization parameter λ.

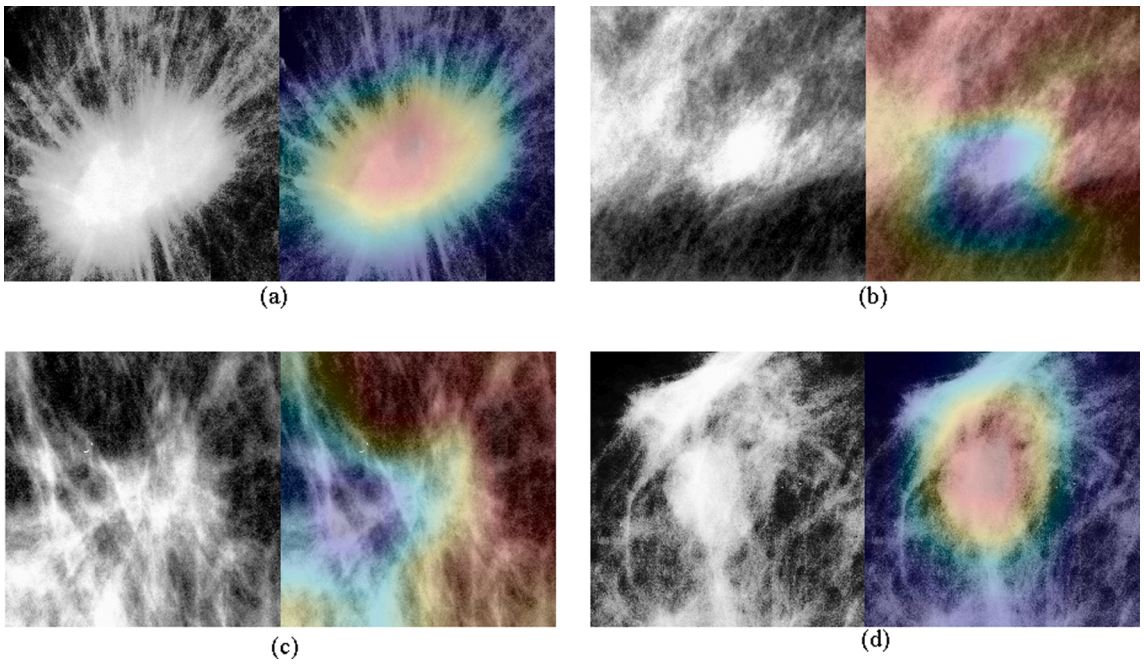

Fig. 6. Images and class activation maps (CAMs) from the convolutional neural network (CNN) model for (a) true positive, (b) false positive, (c) false negative, and (d) true negative examples. CAMs presented depict areas most responsible for the *correct* classification in red and those *least* responsible in blue– i.e. for examples (a) and (c), weights are those for the *malignant* class while in examples (b) and (d) these weights are for the *benign* class. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
List of top features from method of Sahiner et al.

Top 10 Morphological	Top 10 Texture	Top 10 Combined
Perimeter	RLS Long Run Emphasis of RBST, Offset: 45°	Perimeter
Contrast	GLCM Entropy of RBST, Offset: −2, 2, Distance: 2	Contrast
Mean of RDS[a)]	GLCM Entropy of RBST, Offset: −3, −3, Distance: 3	GLCM Difference Entropy of RBST, Offset: −16, −16, Distance: 16
Area Ratio of RDS [a)]	GLCM Difference Variance of RBST, Offset: 0, 12, Distance: 12	GLCM Difference Average of RBST, Offset: −12, 12, Distance: 12
Perimeter to Area Ratio	GLCM Difference Variance of RBST, Offset: −12, −12, Distance: 12	GLCM Difference Variance of RBST, Offset: −4, 0, Distance: 4
Convexity	GLCM Difference Variance of RBST, Offset: −6, −6, Distance: 6	GLCM Difference Variance of RBST, Offset: 0, 8, Distance: 8
Area	GLCM Entropy of RBST, Offset: −8, 0, Distance: 8	GLCM Difference Entropy of RBST, Offset: 0, 16, Distance: 16
Entropy of RDS [a)]	GLCM Energy of RBST, Offset: −16, −16, Distance: 16	GLCM Entropy of RBST, Offset: −2, 0, Distance: 2
Standard Deviation of RDS [a)]	GLCM Difference Average of RBST, Offset: −2, 2, Distance: 2	GLCM Entropy of RBST, Offset: 0, 4, Distance: 4
Zero-crossing Count of RDS [a)]	GLCM Inverse Difference Moment of RBST, Offset: −6, −6, Distance: 6	GLCM Sum Entropy of RBST, Offset: 0, 1, Distance: 1

[a)] Radial Distance Signal.

CADx could safely move to segmentation-free methods would benefit busy clinical workflows, as providing precise lesion segmentation or region-of-interest outlines can be laborious and time consuming in practice. Thus, further evaluation studies in larger patient cohorts and more diverse image sets would be useful to confirm the performance improvement we have observed on CBIS-DDSM from segmentation-free representation learning techniques.

Table 4
Per-feature Results of NN-HDF Method on CBIS-DDSM and Literature Datasets.

Feature	Literature A_Z	CBIS-DDSM A_Z
Spiculation Measure	0.88	0.53
Sharpness	0.53	0.53
Average Gray Level	0.65	0.52
Contrast	0.59	0.52
Texture Measure	0.54	0.51

Our second important finding is that our re-implementation of existing segmentation-dependent methods yielded performance levels on CBIS-DDSM inferior to those reported on the original evaluation datasets in the literature. For instance, Huo et al. reported an A_Z of 0.88 using only the spiculation measure feature, while our re-implementation achieved an A_Z of only 0.53 on CBIS-DDSM using that same feature. Additionally, while Sahiner et al. reported an A_Z of 0.91, we find an A_Z of only 0.75 on CBIS-DDSM using our re-implementation of this method. There exist several possible explanations for these discrepancies. First, the technique used to provide mass lesion segmentations for DDSM was not the same as that originally used in any of the segmentation-dependent algorithms, which could affect their efficacy. We propose that further investigating this sensitivity would be a productive direction for future work. Second, as described in the Methods section, the literature does not always describe existing methods in sufficient detail to ensure an exact re-implementation, and the original code is rarely available, meaning that there likely exist differences in implementation between our study and the original work. Third, segmentation-dependent techniques in the literature are often tuned on small datasets, such as 95 images from 68 patients for Huo et al. [7] and 168 mammograms from 72 patients for Sahiner et al. [6] While we do tune salient parameters for these methods as described in Methods, other preprocessing choices made in the literature (e.g. neighborhood size for Huo et al.) could affect these models' ability to transfer to new datasets. These results reinforce the importance of performing algorithm

assessment on large public datasets and exerting consistent effort to publicly release new datasets for evaluation as acquisition hardware and software change.

Our study has several important limitations, some of which have been previously mentioned. First, the comparison we have performed between our CBIS-DDSM results and those reported in the literature is imperfect, as we were not able to acquire the complete code for every method. This being said, we use standard implementations of CNN and BoVW methods [10,11], use code provided by Sahiner for LDA-RBST [6], and had two separate researchers implement NN-HDF to ensure repeatability; code for each method can be made available upon request. Furthermore, our results in Fig. 4 indicate that our implementation for isolating the margin and periphery – two key parameters of NN-HDF – yields appropriate results (cf. the original paper [7]). Another limitation of our work is that we utilize the segmentations provided by CBIS-DDSM because code for methods used in Sahiner et al. and Huo et al. was unavailable; while we believe this to be a reasonable approach, the difference in segmentation methods could affect the performance of these two techniques. A final limitation of our study is that the DDSM data set is itself an old collection comprising scanned film mammography. Modern mammography is digital, and the results of the methods described in this paper could be different if we used a digital mammography collection. Note that this caveat is not confined to segmentation-dependent methods, as segmentation-free deep learning methods in particular carry the potential to focus on features that are semantically nonsensical as a result of data-driven feature learning, and rigorous evaluation procedures should be utilized to ensure that clinically reasonable features and spatial regions are being utilized. Given the scarcity of public, freely available collections of digital mammography images, and because many prior works have used the DDSM for evaluation, we have chosen to use the CBIS-DDSM collection as the basis of the present work. In the future, it would be helpful to evaluate all CADx methods on digital mammography data sets should they become available, and to remain keenly attuned to the potential for confounding variables such as image quality, latent subsets in the data, and label inconsistency to result in flawed assessments of classifier performance.

6. Conclusions

In this work, we use the public CBIS-DDSM dataset to compare the performance of multiple segmentation-free and segmentation-dependent CADx algorithms using a common evaluation standard. We find that segmentation-free representation learning techniques such as BoVW and CNN are able to equal or outperform re-implementations of segmentation-dependent CADx algorithms on CBIS-DDSM. If verified on larger populations, the use of segmentation-free techniques could increase the positive impact of CADx systems on clinical workflows by minimizing the amount of clinician time and precision required to utilize them effectively. We also observe that segmentation-dependent CADx algorithms do not perform as well on CBIS-DDSM as on the original evaluation datasets in the literature, implying that some combination of differences in segmentation approach, variations in implementation, or an underlying lack of generalizability are affecting algorithm performance. It is our hope that this work provides motivation for further study of different mass classification algorithms using public datasets, which would greatly benefit both clinical and scientific communities.

CRediT authorship contribution statement

Rebecca Sawyer Lee: Conceptualization, Methodology, Software, Investigation, Writing - original draft. **Jared A. Dunnmon:** Conceptualization, Methodology, Software, Investigation, Writing - review & editing. **Ann He:** Software, Investigation. **Siyi Tang:** Software, Investigation, Visualization. **Christopher Ré:** Writing - review & editing. **Daniel L. Rubin:** Conceptualization, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We gratefully acknowledge the support of DARPA under Nos. FA86501827865 (SDH) and FA86501827882 (ASED); NIH under No. U54EB020405 (Mobilize) and U01-CA242879, NSF under Nos. CCF1763315 (Beyond Sparsity), CCF1563078 (Volume to Velocity), and 1937301 (RTML); ONR under No. N000141712266 (Unifying Weak Supervision); the National Cancer Institute; the Moore Foundation, NXP, Xilinx, LETI-CEA, Intel, IBM, Microsoft, NEC, Toshiba, TSMC, ARM, Hitachi, BASF, Accenture, Ericsson, Qualcomm, Analog Devices, the Okawa Foundation, American Family Insurance, Google Cloud, Swiss Re, the Intelligence Community Postdoctoral Fellowship, the Stanford Human-Centered Artificial Intelligence Seed Grants Program, and members of the Stanford DAWN project: Teradata, Facebook, Google, Ant Financial, NEC, VMWare, and Infosys. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements, either expressed or implied, of DARPA, NIH, ONR, or the U.S. Government.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2020.103656.

References

[1] International Agency for Research on Cancer, World Health Organization. Breast Cancer Estimated Incidence, Mortality and Prevalence Worldwide in 2012.
[2] D.A. Berry, K.A. Cronin, S.K. Plevritis, et al., Effect of Screening and adjuvant therapy on mortality from breast cancer, N. Engl. J. Med. 353 (17) (2005) 1784–1792, https://doi.org/10.1056/NEJMoa050518.
[3] American Cancer Society. Breast Cancer: Facts and Figures 2015-2016.
[4] M.S. Fuller, C.I. Lee, J.G. Elmore, Breast cancer screening: an evidence-based update, Med. Clin. North Am. 99 (3) (2016) 451–468, https://doi.org/10.1016/j.mcna.2015.01.002.Breast.
[5] E.L. Henriksen, J.F. Carlsen, I. Mm Vejborg, M.B. Nielsen, C.A. Lauridsen, The efficacy of using computer-aided detection (CAD) for detection of breast cancer in mammography screening: a systematic review, doi:10.1177/0284185118770917.
[6] B. Sahiner, H.P. Chan, N. Petrick, M.a. Helvie, M.M. Goodsitt, Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis, Med. Phys. 25(4) (1998) 516–526. doi:10.1118/1.598228.
[7] Z. Huo, M.L. Giger, C.J. Vyborny, D.E. Wolverton, R.A. Schmidt, K. Doi, Automated computerized classification of malignant and benign masses on digitized mammograms, Acad. Radiol. 5 (3) (1998) 155–168, https://doi.org/10.1016/S1076-6332(98)80278-X.
[8] T. Kooi, G. Litjens, B. van Ginneken, et al., Large scale deep learning for computer aided detection of mammographic lesions, Med. Image Anal. 35 (2017) 303–312, https://doi.org/10.1016/J.MEDIA.2016.07.007.
[9] A. Hamidinekoo, E. Denton, A. Rampun, K. Honnor, R. Zwiggelaar, Deep learning in mammography and breast histology, an overview and future trends, Med. Image Anal. 47 (2018) 45–67, https://doi.org/10.1016/j.media.2018.03.006.
[10] A. Vedaldi, A. Zisserman, Efficient additive kernels via explicit feature maps, IEEE Trans. Pattern Anal. Mach. Intell. 34(3) (2012) 480–492. https://www.robots.ox.ac.uk/vgg/publications/2011/Vedaldi11/vedaldi11.pdf. Accessed June 4, 2018.
[11] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, IEEE CVPR 2 (2017) 2261–2269, https://doi.org/10.1109/CVPR.2017.243.
[12] R.M. Rangayyan, N.R. Mudigonda, J.E. Desautels, Boundary modelling and shape analysis methods for classification of mammographic masses, Med. Biol. Eng. Comput. 38 (5) (2000) 487–496, https://doi.org/10.1007/BF02345742.
[13] N.R. Mudigonda, R.M. Rangayyan, J.E. Desautels, Gradient and texture analysis for the classification of mammographic masses, IEEE Trans. Med. Imaging 19 (10) (2000) 1032–1043, https://doi.org/10.1109/42.887618.

[14] B. Sahiner, H.P. Chan, N. Petrick, M.A. Helvie, L.M. Hadjiiski, Improvement of mammographic mass characterization using spiculation meausures and morphological features, Med. Phys. 28 (7) (2001) 1455–1465, https://doi.org/10.1118/1.1381548.
[15] J. Bozek, M. Kallenberg, M. Grgic, N. Karssemeijer, Use of volumetric features for temporal comparison of mass lesions in full field digital mammograms, Med. Phys. 41 (2) (2014) 021902, https://doi.org/10.1118/1.4860956.
[16] P. Görgel, A. Sertbas, O.N. Uçan, Computer-aided classification of breast masses in mammogram images based on spherical wavelet transform and support vector machines, Expert Syst. 32 (1) (2015) 155–164, https://doi.org/10.1111/exsy.12073.
[17] D. Brzakovic, X.M. Luo, U. Brzakovic, An approach to automated detection of tumors in mammograms, IEEE Trans. Med. Imaging 9 (3) (1990) 233–241, https://doi.org/10.1109/42.57760.
[18] S. Timp, C. Varela, N. Karssemeijer, Temporal change analysis for characterization of mass lesions in mammography, IEEE Trans. Med. Imaging 26 (7) (2007) 945–953, https://doi.org/10.1109/TMI.2007.897392.
[19] K. Ganesan, U.R. Acharya, C.K. Chua, L.C. Min, T.K. Abraham, Automated diagnosis of mammogram images of breast cancer using discrete wavelet transform and spherical wavelet transform features: a comparative study, Technol. Cancer Res. Treat. 13 (6) (2014) 605–615, https://doi.org/10.7785/tcrtexpress.2013.600262.
[20] J.Y. Choi, D.H. Kim, K.N. Plataniotis, Y.M. Ro, Classifier ensemble generation and selection with multiple feature representations for classification applications in computer-aided detection and diagnosis on mammography, Expert Syst. Appl. 46 (2016) 106–121, https://doi.org/10.1016/j.eswa.2015.10.014.
[21] A. Oliver, J. Freixenet, J. Martí, et al., A review of automatic mass detection and segmentation in mammographic images, 2009. doi:10.1016/j.media.2009.12.005.
[22] A.R. Jamieson, K. Drukker, M.L. Giger, Breast image feature learning with adaptive deconvolutional networks, 2012, 831506. doi:10.1117/12.910710.
[23] B. Liu, Y. Jiang, A multitarget training method for artificial neural network with application to computer-aided diagnosis, Med. Phys. 40 (1) (2013) 011908, https://doi.org/10.1118/1.4772021.
[24] X.Z. Li, S. Williams, G. Lee, M. Deng, Computer-aided mammography classification of malignant mass regions and normal regions based on novel texton features, in: 2012 12th Int Conf Control Autom Robot Vision, ICARCV 2012, 2012, 2012 (December), pp. 1431–1436. doi:10.1109/ICARCV.2012.6485399.
[25] S.J. Magny, R. Shikhman, A.L. Keppke, Breast Imaging Reporting and Data System. StatPearls Publishing; 2020. http://www.ncbi.nlm.nih.gov/pubmed/29083600. Accessed October 31, 2020.
[26] Y. Wang, F. Aghaei, A. Zarafshani, Y. Qiu, W. Qian, B. Zheng, Computer-aided classification of mammographic masses using visually sensitive image features, J. Xray Sci. Technol. 25 (1) (2017) 171–186, https://doi.org/10.3233/XST-16212.
[27] W. Zhu, Q. Lou, Y.S. Vang, X. Xie, Deep multi-instance networks with sparse label assignment for whole mammogram classification, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10435, LNCS. Springer Verlag, 2017, pp. 603–611. doi:10.1007/978-3-319-66179-7_69.
[28] J. Arevalo, F.A. González, R. Ramos-Pollán, J.L. Oliveira, M.A. Guevara Lopez, Representation learning for mammography mass lesion classification with convolutional neural networks, Comput. Meth. Prog. Biomed. 127 (2016) 248–257, https://doi.org/10.1016/j.cmpb.2015.12.014.
[29] E.K. Kim, H.E. Kim, K. Han, et al., Applying data-driven imaging biomarker in mammography for breast cancer screening: preliminary study, Sci. Rep. 8 (1) (2018) 1–8, https://doi.org/10.1038/s41598-018-21215-1.
[30] D. Ribli, A. Horváth, Z. Unger, P. Pollner, I. Csabai, Detecting and classifying lesions in mammograms with Deep Learning, Sci. Rep. 8 (1) (2018) 1–7, https://doi.org/10.1038/s41598-018-22437-z.
[31] M.A. Al-masni, M.A. Al-antari, J.M. Park, et al., Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system, Comput. Meth. Prog. Biomed. 157 (2018) 85–94, https://doi.org/10.1016/j.cmpb.2018.01.017.
[32] W. Lotter, G. Sorensen, D. Cox, A multi-scale CNN and curriculum learning strategy for mammogram classification, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol 10553 LNCS. Springer Verlag, 2017, pp. 169–177. doi:10.1007/978-3-319-67558-9_20.
[33] F.F. Ting, Y.J. Tan, K.S. Sim, Convolutional neural network improvement for breast cancer classification, Expert Syst. Appl. 120 (2019) 103–115, https://doi.org/10.1016/j.eswa.2018.11.008.
[34] H. Chougrad, H. Zouaki, O. Alheyane, Deep Convolutional Neural Networks for breast cancer screening, Comput. Meth. Prog. Biomed. 157 (2018) 19–30, https://doi.org/10.1016/j.cmpb.2018.01.011.
[35] D.A. Ragab, M. Sharkas, S. Marshall, J. Ren, Breast cancer detection using deep convolutional neural networks and support vector machines, PeerJ 2019 (1) (2019), e6201, https://doi.org/10.7717/peerj.6201.
[36] H. Li, D. Chen, W.H. Nailon, M.E. Davies, D. Laurenson, Dual Convolutional Neural Networks for Breast Mass Segmentation and Diagnosis in Mammography. August 2020. http://arxiv.org/abs/2008.02957. Accessed October 31, 2020.
[37] L. Tsochatzidis, L. Costaridou, I. Pratikakis, Deep learning for breast cancer diagnosis from mammograms—a comparative study, J. Imaging 5 (3) (2019) 37, https://doi.org/10.3390/jimaging5030037.
[38] H. Chougrad, H. Zouaki, O. Alheyane, Multi-label transfer learning for the early diagnosis of breast cancer, Neurocomputing 392 (2020) 168–180, https://doi.org/10.1016/j.neucom.2019.01.112.
[39] Y. Chen, Q. Zhang, Y. Wu, B. Liu, M. Wang, Y. Lin, Fine-tuning ResNet for breast cancer classification from mammography, in: Lecture Notes in Electrical Engineering, vol 536. Springer Verlag, 2019, pp. 83–96. doi:10.1007/978-981-13-6837-0_7.
[40] L.G. Falconi, M. Perez, W.G. Aguilar, A. Conci, Transfer learning and fine tuning in breast mammogram abnormalities classification on CBIS-DDSM database, Adv. Sci. Technol. Eng. Syst. 5 (2) (2020) 154–165, https://doi.org/10.25046/aj050220.
[41] M. Alkhaleefah, P. Kumar Chittem, V.P. Achhannagari, S.C. Ma, Y.L. Chang, The influence of image augmentation on breast lesion classification using transfer learning, in: 2020 International Conference on Artificial Intelligence and Signal Processing, AISP 2020. Institute of Electrical and Electronics Engineers Inc., 2020. doi:10.1109/AISP48273.2020.9073516.
[42] X. Shu, L. Zhang, Z. Wang, Q. Lv, Z. Yi, Deep neural networks with region-based pooling structures for mammographic image classification, IEEE Trans. Med. Imaging 39 (6) (2020) 2246–2255, https://doi.org/10.1109/TMI.2020.2968397.
[43] R.K. Samala, H.P. Chan, L.M. Hadjiiski, M.A. Helvie, C.D. Richter, Generalization error analysis for deep convolutional neural network with transfer learning in breast cancer diagnosis, Phys. Med. Biol. 65 (10) (2020) 105002, https://doi.org/10.1088/1361-6560/ab82e8.
[44] A. Gossmann, K.H. Cha, X. Sun, Performance deterioration of deep neural networks for lesion classification in mammography due to distribution shift: an analysis based on artificially created distribution shift, in: H.K. Hahn, M.A. Mazurowski (eds.), Medical Imaging 2020: Computer-Aided Diagnosis. Vol 11314. SPIE; 2020, p. 3. doi:10.1117/12.2551346.
[45] C. Beltran-Perez, H.L. Wei, A. Rubio-Solis, Generalized multiscale RBF networks and the DCT for breast cancer detection, Int. J. Automat. Comput. 17 (1) (2020) 55–70, https://doi.org/10.1007/s11633-019-1210-y.
[46] W. Ansar, A.R. Shahid, B. Raza, A.H. Dar, Breast cancer detection and localization using mobilenet based transfer learning for mammograms, in: Communications in Computer and Information Science, vol. 1187 CCIS. Springer; 2020, pp. 11–21. doi:10.1007/978-3-030-43364-2_2.
[47] M. de Vriendt, P. Sellars, A.I. Aviles-Rivero, The GraphNet zoo: an all-in-one graph based deep semi-supervised framework for medical image classification, in: LNCS. vol 12443, Springer, Cham, 2020, pp. 187–197. doi:10.1007/978-3-030-60365-6_18.
[48] R.S. Lee, F. Gimenez, A. Hoogi, K.K. Miyake, M. Gorovoy, D.L. Rubin, The Curated Breast Imaging Subset of the Digital Database for Screening Mammography, 2015. doi:https://doi.org/10.7937/K9/TCIA.2016.7O02S9CY.
[49] R.S. Lee, F. Gimenez, A. Hoogi, K.K. Miyake, M. Gorovoy, D.L. Rubin, A curated mammography data set for use in computer-aided detection and diagnosis research, Sci. Data 4 (2017) 170177, https://doi.org/10.1038/sdata.2017.177.
[50] C. Newton-Cheh, T. Johnson, V. Gateva, et al., Genome-wide association study identifies eight loci associated with blood pressure, Nat. Genetics 41 (6) (2009) 666–676, https://doi.org/10.1038/ng.361.
[51] H. Chan, D. Wei, M.A. Helvie, et al., Computer-aided classification of mammographic masses and normal tissue: linear discriminant analysis in texture feature space, Phys. Med. Biol. 40 (5) (1995) 857–876, https://doi.org/10.1088/0031-9155/40/5/010.
[52] A. Vedaldi, B. Fulkerson, VLFeat: An Open and Portable Library of Computer Vision Algorithms, 2008.
[53] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (1) (2010), https://doi.org/10.18637/jss.v033.i01.
[54] V. Gulshan, L. Peng, M. Coram, et al., Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, JAMA 316 (22) (2016) 2402, https://doi.org/10.1001/jama.2016.17216.
[55] A. Esteva, B. Kuprel, R.A. Novoa, et al., Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks, 2017. doi:10.1038/nature21056.
[56] H.-C. Shin, H.R. Roth, M. Gao, et al., Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning, IEEE Trans. Med. Imaging 35 (5) (2016) 1285–1298, https://doi.org/10.1109/TMI.2016.2528162.
[57] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database, in: IEEE CVPR. June 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.
[58] J.A. Dunnmon, D. Yi, C.P. Langlotz, C. Ré, D.L. Rubin, M.P. Lungren, Assessment of convolutional neural networks for automated classification of chest radiographs, Radiology 290 (2) (2019) 537–544, https://doi.org/10.1148/radiol.2018181422.
[59] A.J. Ratner, H.R. Ehrenberg, Z. Hussain, J. Dunnmon, C. Ré, Learning to Compose Domain-Specific Transformations for Data Augmentation. September 2017. http://arxiv.org/abs/1709.01643. Accessed October 2, 2017.
[60] Z. Huo, M.L. Giger, C.J. Vyborny, et al., Analysis of spiculation in the computerized classification of mammographic masses, Med. Phys. 22 (10) (1995) 1569–1579, https://doi.org/10.1118/1.597626.
[61] J. Liu, M. Ma, C. Wu, J. Tai, Tests of equivalence and non-inferiority for diagnostic accuracy based on the paired areas under ROC curves, Stat. Med. 25 (7) (2006) 1219–1238, https://doi.org/10.1002/sim.2358.
[62] J.A. Hanley, B.J. McNeil, A method of comparing the areas under receiver operating characteristic curves derived from the same cases, Radiology 148(3) (1983) 839–843. https://pubs.rsna.org/doi/pdf/10.1148/radiology.148.3.6878708. Accessed March 25, 2018.
[63] Z. Du, Y. Hao, rocNIT: Non-Inferiority Test for Paired ROC Curves, 2016. doi:10.1002/sim.2358.
[64] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: IEEE CVPR, 2016, pp. 2921–2929. https://www.cv

-foundation.org/openaccess/content_cvpr_2016/papers/Zhou_Learning_Deep_Features_CVPR_2016_paper.pdf. Accessed March 30, 2018.
[65] D.G. Lowe, Object recognition from local scale-invariant features, Proc Seventh IEEE Int Conf Comput Vis. 2 (1999), https://doi.org/10.1109/ICCV.1999.790410.
[66] N. Nguyen, R. Caruana, Consensus clusterings, in: Seventh IEEE International Conference on Data Mining, 2007:, pp. 607–612, https://doi.org/10.1109/ICDM.2007.73.
[67] A. Strehl, J. Ghosh, Cluster ensembles – a knowledge reuse framework for combining multiple partitions, J. Mach. Learn. Res. 3 (2002) 583–617.
[68] Giecold G. Cluster Ensembles.