Exploring nonlinear feature space dimension reduction and data representation in breast CADx with Laplacian eigenmaps and *t*-SNE

Andrew R. Jamieson,^{a)} Maryellen L. Giger, Karen Drukker, Hui Li, Yading Yuan, and Neha Bhooshan

Department of Radiology, University of Chicago, Chicago, Illinois 60637

(Received 26 May 2009; revised 9 October 2009; accepted for publication 2 November 2009; published 22 December 2009)

Purpose: In this preliminary study, recently developed unsupervised nonlinear dimension reduction (DR) and data representation techniques were applied to computer-extracted breast lesion feature spaces across three separate imaging modalities: Ultrasound (U.S.) with 1126 cases, dynamic contrast enhanced magnetic resonance imaging with 356 cases, and full-field digital mammography with 245 cases. Two methods for nonlinear DR were explored: Laplacian eigenmaps [M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural Comput.* **15**, 1373–1396 (2003)] and *t*-distributed stochastic neighbor embedding (*t*-SNE) [L. van der Maaten and G. Hinton, “Visualizing data using *t*-SNE,” *J. Mach. Learn. Res.* **9**, 2579–2605 (2008)].

Methods: These methods attempt to map originally high dimensional feature spaces to more human interpretable lower dimensional spaces while preserving both local and global information. The properties of these methods as applied to breast computer-aided diagnosis (CADx) were evaluated in the context of malignancy classification performance as well as in the visual inspection of the sparseness within the two-dimensional and three-dimensional mappings. Classification performance was estimated by using the reduced dimension mapped feature output as input into both linear and nonlinear classifiers: Markov chain Monte Carlo based Bayesian artificial neural network (MCMC-BANN) and linear discriminant analysis. The new techniques were compared to previously developed breast CADx methodologies, including automatic relevance determination and linear stepwise (LSW) feature selection, as well as a linear DR method based on principal component analysis. Using ROC analysis and 0.632+ bootstrap validation, 95% empirical confidence intervals were computed for the each classifier’s AUC performance.

Results: In the large U.S. data set, sample high performance results include, $AUC_{0.632+}=0.88$ with 95% empirical bootstrap interval [0.787;0.895] for 13 ARD selected features and $AUC_{0.632+}=0.87$ with interval [0.817;0.906] for four LSW selected features compared to 4D *t*-SNE mapping (from the original 81D feature space) giving $AUC_{0.632+}=0.90$ with interval [0.847;0.919], all using the MCMC-BANN.

Conclusions: Preliminary results appear to indicate capability for the new methods to match or exceed classification performance of current advanced breast lesion CADx algorithms. While not appropriate as a complete replacement of feature selection in CADx problems, DR techniques offer a complementary approach, which can aid elucidation of additional properties associated with the data. Specifically, the new techniques were shown to possess the added benefit of delivering sparse lower dimensional representations for visual interpretation, revealing intricate data structure of the feature space. © 2010 American Association of Physicists in Medicine. [DOI: 10.1118/1.3267037]

Key words: nonlinear dimension reduction, computer-aided diagnosis, breast cancer, Laplacian eigenmaps, *t*-SNE

I. INTRODUCTION

Radiologic image interpretation is a complex task. A radiologist’s expertise, developed only with exhaustive training and experience, rests in their ability for extracting and meaningfully synthesizing relevant information from a medical image. However, even under idealized image acquisition conditions, precise conclusions may not be possible for certain radiologic tasks. Thus, computer-aided diagnosis (CADx) systems have been introduced in a number of contexts in an attempt to assist human interpretation of medical images.³ A relatively well-developed clinical application, for which

computerized efforts in radiological image analysis have been studied, is the use of CADx in the task of detecting and diagnosing breast cancer.^{4–10} Similar to the radiologist’s task, a computer algorithm is designed to make use of the highly complicated breast image input data, attempting to intelligently reduce image information into more interpretable and ultimately clinically actionable output structures, such as an estimate of the probability of malignancy. Understanding how to optimally make use of the enormity of the initial image information input and best arrive at the succinct conceptual notion of “diagnosis” is a formidable challenge. Al-

though there may be any number of various operations/transformations involved in arriving at this high-level end output, whether in the human brain or *in silico*, two common critical pursuits are proper data representation and reduction. The current study aims to explore the potential enhancements offered to breast mass lesion CADx algorithms through the application of two recently developed dimensionality reduction and data representation techniques, Laplacian eigenmaps and *t*-distributed stochastic neighbor embedding (*t*-SNE).^{1,2}

II. BACKGROUND

II.A. Current CADx feature representation

Restricted by limited sample data sets, computational power, and lack of complete theoretical formalism, image-based pattern recognition and classification techniques often tackle the objective task at hand by substantially simplifying the problem. Traditionally, breast CADx systems employ a two pronged approach, first, image preprocessing and feature extraction, and second, classification in the feature space, either by unsupervised methods, supervised methods, or both. A review of past and present CADx methods employed can be found in referenced articles referenced.^{3,11} Often, instead of attempting to make use of the complete image,¹² CADx typically condenses image information down to a vector of numerical values, each representative of some attribute of the image or lesion present in the image. One can consider this first data reduction step as “perceptual” processing, meaning that at this stage the algorithm’s goal is to isolate and “perceive” only the most relevant components of the original image that will contribute toward distinguishing between the target classes (e.g., malignant or benign). One of the steps in eliminating unnecessary image information is lesion margin segmentation.^{5,13} Typically, features such as those extracted from the segmented lesion are heuristic in nature and mimic important human identified aspects of the lesion. However more mathematical and abstract feature quantities may also be calculated that may represent information visually imperceptible to the unaided eye. While the use of data from a segmented lesion introduces bias into the algorithm’s task as a whole, this “informed” bias allows for the efficient removal of much unnecessary image data, for instance, normal background breast tissue. From here the second main component of the CADx algorithm falls usually into the context of the well-formalized canonical problem found in statistical pattern recognition for classification.^{14,15}

After the first CADx phase of feature extraction, each high dimensional image in the sample set is now reduced to a single vector in a lower dimensional feature space. However, due to the finite size of image sample data, if too many features are examined simultaneously, regions containing a low density of points in the feature space will exist, resulting in statistically inconclusive classification ability. This dilemma is affectionately termed the curse of dimensionality.¹⁶ Thus, a further reduction in the full feature space is required for a practically useful data representation. This aspect is a major concern of the second component of traditional CADx

schemes, and is succinctly known as “feature selection.” Much literature has been generated on this subject matter in the explicit context of improving CADx performance.^{17–19} Some CADx schemes may employ only four to five features maximum, in which case feature selection may not be necessary, since the data set sample size, even for relatively smaller sizes, may be sufficiently large to avoid overtraining classifiers. However, it is reasonable to imagine CADx researchers interested in testing hundreds of potential features. In either case, when appropriately coupled with a well-regularized supervised classification method, the ultimate objective of features selection is to discover the “optimal” data representation, or subset of features for robustly maximizing the desired diagnostic task performance. That is, the method attempts both to mimic and to maximize the theoretical upper bound or ideal observer performance possible over the sampled joint probability distribution of the selected features. While this step is critical, finding such a subset is nontrivial and may also be highly dependent on the specific characteristics of the sample data. Developed techniques in feature selection for CADx range from simpler linear methods, such as those based on linear discriminant analysis (LDA), to nonlinear and more sophisticated Bayesian-based, such as the use of Bayesian artificial neural networks (BANN) and automatic relevance determination (ARD), to random search stochastic methods such as genetic algorithms as well as information theoretic techniques.^{17,19–21}

The most striking quality of the methods mentioned above, in the context of CADx, is that during feature selection, some features are completely removed from the final classification scheme, and hence image information is either explicitly or implicitly discarded altogether. However, while removing all the information associated with a specific feature not selected, by selecting a smaller subset of individual features, what is gained is greater immediate human interpretability. Specifically, the isolated groups of features may have clear physical or radiological meanings and thus may be of interest to investigators or radiologists for understanding how these characteristics relate to the ability to distinguish class categories (malignant, benign, cyst, etc.). To this end, in order to interpret the nature of the feature space and attempt to identify characteristic trends, one may visually inspect plots displaying single features or attempt to capture synergistic qualities between two or among three features simultaneously. Above three dimensions, as it becomes nontrivial to interpret the structure of the feature space, often instead, the use of a metrics such as the ROC curve and/or AUC based on output from the decision variable of a trained merged feature classifier are used to interrogate the quality of the higher dimensional feature spaces.

As such, beyond identifying which feature or features appear to hold classification utility, current CADx methods offer little theoretical/formal guidance in a recovering understanding of the inherent data structure represented by the higher dimensional feature spaces.

II.B. Proposed feature space representation and reduction for CADx

Due in part to the ever-growing demand of data driven science, in recent years much interest has emerged in developing techniques for discovering efficient representations of large-scale complex data.²² Conceptually the goal is to discover the intrinsic structure of the data and adequately express this information in a lower dimensional representation. Classically, the problem of dimension reduction (DR) and data representation has been approached by applying linear transformations such as the well-known principal component analysis (PCA) or more general singular value decomposition.^{23,24} Interestingly, despite PCA's age, only recently has this method been considered for the specific application to CADx feature space reduction.²⁵ In this particular breast ultrasound study, while no significant boosts in lesion classification performance were discovered, PCA was found to be a suitable substitute in place of more computationally intensive and cumbersome feature selection methods.²⁵ This efficient lower dimensional PCA data representation, i.e., linear combinations of the original features accounting for the maximum global variance decomposition in the data, proved capable of capturing sufficient information for robust classification. However, PCA is not capable of representing higher order, nonlinear, local structure in the data.

The goal of recently proposed nonlinear data reduction and representation methods focuses on this very problem.^{1,2} The present methods of interest to this study, Laplacian eigenmaps and *t*-SNE, offer two distinct approaches for explicitly addressing the challenge of capturing and efficiently representing the properties of the low dimensional manifold on which the original high dimensional data may lie. Previous studies have investigated other nonlinear DR techniques, including self-organizing maps and graph embedding, for breast cancer in the context of biomedical image signal processing,^{26,27} as well as for a breast cancer BIRADs database clustering.²⁸ To our knowledge, the relationship between breast CADx performance and these nonlinear feature space DR and representation have yet to be properly investigated. These new techniques may contribute two key enhancements to current CADx schemes.

1. A principled alternative to feature selection. Both methods explicitly attempt to preserve as much structure in the original feature space as possible, and thus require no need to assumingly force exclusion of features from the original set, and hence unnecessary loss of image information.
2. A more natural and sparse data representation that immediately lends itself to generating human interpretable visualizations of the inherent structures present in the high dimensional feature data.

It is important to note that by employing DR on CADx feature spaces, one surrenders, to a varying extent, the ability to immediately interpret the physical meaning of the embedded representation. Yet, critically, this is a necessary and fun-

damental trade off, as the conceptual focus is shifted to a more holistic approach, specifically, that of discovering an efficient lower dimensional representation of the intrinsic data structure. The core tenant of such an unsupervised approach is to limit assumptions imposed on the data. This major shift in philosophy regarding the original high dimensional feature space embodies the notion "let the data speak for itself." It seems reasonable to assume that if supervised classifiers are capable of uncovering sufficient data structure in the extracted feature space for producing adequate classification performance, then such principled local geometry preserving reduction mappings should reveal structural evidence corroborating such findings.

II.C. Outline of evaluation for proposed methods

The primary objective of this study is to evaluate the classification performance characteristics of breast lesion CADx schemes employing the Laplacian eigenmap or *t*-SNE DR techniques in place of previously developed feature selection methods. Second, and more qualitatively, we aim to investigate and gain insight into the properties of sample visualizations representative of lower dimensional feature space mappings of high dimensional breast lesion feature data. Additionally, the feasibility and robustness of these nonlinear reduction methods for CADx feature space reduction are tested across three separate imaging modalities: Ultrasound (U.S.), dynamic contrast enhanced MRI (DCE-MRI), and full-field digital mammography (FFDM), having case sets of 1126, 356, and 245 cases, respectively.

III. METHODS

III.A. Data set

All data characterized in this study consists of clinical breast lesions presented in images acquired at the University of Chicago Medical Center (Chicago, IL). Lesions are labeled according to the truth known by biopsy or radiologic report and collected under HIPAA-compliant IRB protocols. Furthermore, the breast lesion feature data sets were generated from previously developed CADx algorithms at the University of Chicago. For a review of these techniques see Giger, Huo, Kupinski for x-ray mammography and Drukker for U.S., and Chen for DCE-MRI.^{4-11,29}

In each of the modalities, the lesion center is identified manually for the CADx algorithm, which then performs automated seeded segmentation of the lesion margin followed by computerized feature extraction. Table I below summarizes the content of the respective imaging modality databases used, including the total number of initial lesion features extracted. Note that the mammographic imaging modality (FFDM) contains only two lesion class categories, malignant and benign. For ultrasound and DCE-MRI, a more detailed subcategorization is provided, including invasive carcinoma (IDC), ductal carcinoma *in situ* (DCIS), benign solid masses, and benign cystic masses. For clarity, this initial study only considers binary classification performance in the task of distinguishing between the more broad identity of

TABLE I. Feature database characteristics.

Modality	Total number of images	Number of malignant lesions	Number of benign lesions	Total number of lesion features calculated
U.S.	2956	158	968 (401 mass/567 cystic)	81
DCE-MRI	356	223 (151 IDC/72 DCIS)	133	31
FFDM	735	132	113	40

malignant and benign (cancerous versus noncancerous). However, during qualitative inspection of the dimension reduced mappings, it will be of interest to reintroduce these distinctions for visualization purposes.

Geometric, texture, and morphological features, such as margin sharpness, were extracted across all modalities. Also, the DCE-MRI data set includes kinetic features, and the U.S. features include those related to posterior acoustic behavior.^{8,10} All raw extracted feature value data sets were normalized to zero mean and divided by the unit sample standard deviation. Due to page limitations, the details of each feature can be found in Refs. 4–11 and 29.

III.B. Classifiers

In our evaluation of the new DR techniques, we chose two types of classifiers: A relatively simple LDA classifier and a more sophisticated nonlinear, BANN classifier.¹⁵ LDA is a well-known and commonly used linear classification method which will not be reviewed here (for reference and examples in breast lesion CADx see Refs. 4, 30, and 31). The BANN, as the name suggests, follows the usual multilayer perception, neural network design, but additionally employs Bayesian theory as a means of classifier regularization.^{15,32} The BANN has been shown to model the optimal ideal observer for classification given sufficient sample sizes as input for training.³³ The critical technical hurdle in implementing BANNs lies in accurately estimating posterior weight distributions, as analytical calculation is intractable. As such, either approximation or sampling based methods must be deployed in practice.³⁴ Markov Chain Monte Carlo (MCMC) sampling methods can be used to directly sample from the full posterior probability distribution.³² We implemented a MCMC-BANN classifier using the *Netlab* package of Nabney³⁵ for MATLAB. The network architecture $k-(k+1)-1$ was used. That is, k input layer nodes (one for each of the k selected features), a hidden layer with $(k+1)$ nodes, and a single output target as probability of malignancy. For each classifier trained, we generated at least 2000 MCMC samples of the weights' posterior probability distribution. The mean value of the classification prediction (probability of malignancy) output from each of the different 2000 weight samples was used to produce a single classification estimate for new test input cases.

III.C. Explicit supervised feature selection methods

Two previously developed feature selection methods are considered in this paper for comparison, and include linear

stepwise and ARD feature selection. These methods are used to identify a specific set of features for input into the classifier.

III.C.1. Linear stepwise feature selection

Linear stepwise feature selection (LSW-FS) relies on linear discriminant-based functions. Beginning with only a single selected feature, multiple combinations of features are considered one at a time, by exhaustively adding, retaining, or removing each subsequent feature to the potential set of selected features. For each new combination, a metric, the Wilks' lambda is calculated and a selection criterion based on F statistics is used.¹⁷ The “ F -to-enter” and “ F -to-remove” used in this study were automatically adjusted to allow for the specified number of features desired for U.S., DCE-MRI, and FFDM feature selection. For examples of LSW-FS use in breast CADx, references are provided.^{17,25,30}

III.C.2. Automatic relevance determination

A consequence of the BANNs is the possibility for joint feature selection and classification using ARD.^{15,32,34,35} ARD works by placing Bayesian hyper priors, also known as hierarchical priors, over the initial prior distributions already imposed on the network weights connected to the input nodes. The “relevant” features are then discovered as estimates for the hyper parameters, which characterize the prior distributions over the respective input layer weights, are updated via Gibbs sampling giving the posterior hyper parameter estimate. The magnitudes of the final, converged on hyper parameters are then used to indicate the relative utility of the respective feature input layer weights toward accomplishing the classification task. Thus, by way of the Bayesian regularization, ARD allows for one-shot feature selection and classifier design. Furthermore, a key advantage of ARD feature selection is its ability to identify important nonlinear features coupled to the classification objective, due to the inherent nonlinear nature of BANN.¹⁹ Due to these qualities, ARD-MCMC-BANN classifiers were also included for comparison in our study.

In this study, we extend MCMC-BANN to incorporate ARD following the implementation of Nabney.³⁵ This methodology was previously investigated for breast feature selection and classification in DCE-MRI CADx.¹⁹ In our study, 1000 samples were calculated for the hyper parameters beginning with a gamma hyper prior distribution of mean parameter value equal to 3 and a shape parameter equal to 4.

III.D. Unsupervised dimension reduction feature mappings

In comparison to the supervised feature selection methods, three unsupervised DR methods were evaluated here; the latter two nonlinear methods are offered as a novel application to the field of breast image CADx. The general problem of dimensionality reduction can be described mathematically as provided an initial set $x_1, \dots, x_k$ of k points in R^l , discover a set $y_1, \dots, y_k$ in R^m such that y_i sufficiently describes or “represents” the qualities of interest found in the original set x_i . In the context of breast lesion CADx feature extraction, the ideally lower dimensional mappings should aim to preserve and represent as much relevant structural information toward the task of malignancy estimation. It should be noted that DR still requires, in some sense, feature selection, meaning one must specify the number of mapped dimensions to retain for the subsequent classification step. Ideally, methods designed to estimate intrinsic dimensionality of the data structure could be used to direct this choice.³⁶ However, proper evaluation of the integrity of such methods in this context is beyond the scope of this research effort. Thus, in approaching the problem from a more naïve perspective, as done here, focus is centered on gaining a general intuition for the overall major trends encountered.

III.D.1. Linear feature reduction: PCA

Mathematically, PCA is linear transformation which maps the original feature space onto new orthogonal coordinates. The new coordinates, or principal components (PC), represent ordered orthogonal data projections capturing the maximum variance possible, with the first PC corresponding to the highest global variance.^{23,24} Drukker *et al.*²⁵ used PCA as an alternative to feature selection for breast U.S. CADx.

III.D.2. Nonlinear feature dimension reduction

As discussed in Secs. I and II, the following two recently proposed DR and data representation methods are nonlinear in nature, and specifically designed to address the problem of local data structure preservation. Laplacian eigenmaps and t -SNE offer highly distinct solutions to this problem.

III.D.2.a. Laplacian eigenmaps. Drawing on familiar concepts found in spectral graph theory, Laplacian eigenmaps, proposed by Belkin and Niyogi¹ in 2002, use the notion of a graph Laplacian applied to a weighted neighborhood adjacency graph containing the original data set information. This weighted neighborhood graph is regarded geometrically as a manifold characterizing the structure of the data. The eigenvalues and eigenvectors are computed for the graph Laplacian, which are in turn utilized for embedding a lower dimensional mapping representative of the original manifold. Acting as an approximation to the Laplace Beltrami operator, the weighted graph Laplacian transformation can be shown, in a certain sense, to optimally preserve local neighborhood information.³⁷ Thus, the feature data considered in the reduced dimensional space mapping is essentially a discrete approximate representation of the natural geometry of the

original continuous manifold.

As Belkin and Niyogi¹ note, the algorithm is relatively simple and straightforward to implement. Additionally, the algorithm is not computationally intensive. For our largest data set the mappings were computed within a few seconds using MATLAB code. Algorithm details as well as explanation of necessary input parameters for the implementation used here are provided below in Appendix, Sec. 1.

It is important to note that there is no theoretical justification for how to choose the needed parameters for the algorithm. Thus, an array of parameter choices was evaluated in this study. Lastly, parts of the MATLAB code, related only to the implementation of the Laplacian eigenmap, were modified from the publicly available dimension reduction toolbox provided by Laurens van der Maaten of Maastricht University (Maastricht, Netherlands).³⁸

III.D.2.b. t -SNE. The other nonlinear mapping technique considered, t -SNE of van der Maaten and Hinton,² approaches the dimension reduction and data reduction problem by employing entirely different mechanisms to the Laplacian eigenmaps. t -SNE attacks DR from a stochastic and probabilistic-based framework. While requiring orders of magnitude more computational effort, such statistically oriented approaches, provided they are well-conditioned, may potentially offer greater flexibility in certain contexts due in part by the lessening of potentially restrictive theoretical mathematical formalism. For these reasons, the t -SNE method was considered as an interesting comparison alongside the Laplacian eigenmap.

t -SNE is an improved variation in the original stochastic neighbor embedding (SNE) of Hinton and Rowies.³⁹ The basic idea behind SNE is to minimize the difference between specially defined conditional probability distributions that represent similarities, calculated for the data points in both the high and low dimensional representations. In particular, SNE begins by first computing the conditional probability $p_{j|i}$ given by

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|/2\sigma_i^2)}$$

and

$$q_{j|i} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)} \quad (1)$$

and $q_{j|i}$ in the lower dimensional space, with $p_{i|i}$ and $q_{i|i}$ set to zero. These similarities express the probability that x_i (y_i) would select x_j (y_j) as its neighbor, resulting in high values for nearby points and lower values for distantly separated ones. The central assumption in SNE is that if the low dimensional mapped points in Y space correctly model the similarity structure of its higher dimensional counterparts in X , then the conditional probabilities will be equal. The summed Kullback–Leibler (KL) divergence is used to gauge how well $q_{j|i}$ models $p_{j|i}$. Using gradient descent methods, SNE minimizes a KL based cost function. Sampled points from an isotropic Gaussian with small variance centered at the origin are used to initialize the gradient decent. Updates

are made to the mapped space Y for each iteration. Additionally, the parameter σ_i of Eq. (1) must be selected. σ_i is the variance in the Gaussian centered on the high dimensional point x_i . Because of the difficulty in determining if an optimal σ_i exists, a user defined property called perplexity is used to facilitate its selection, defined by $Perp(P_i) = 2^{H(P_i)}$. Calculated in bits, $H(P_i)$ is the Shannon entropy over P_i

$$H(P_i) = - \sum_j p_{j|i} \log_2 p_{j|i}. \quad (2)$$

During SNE, a binary search is performed to find the value of σ_i that produces a P_i with the user specified perplexity. Suggested typical settings range between 5 and 50.²

t -SNE introduces two critical improvements to SNE.² First, the gradient as well as cost function optimization is simplified by using symmetrized conditional probabilities to define the joint probabilities on P and Q [e.g., $p_{ij} = (p_{j|i} + p_{i|j})/2n$] and the minimizing cost over a single KL divergence as opposed to a sum

$$\begin{aligned} C &= \sum_i KL(P_i \| Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \Rightarrow C' = KL(P \| Q) \\ &= \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}. \end{aligned} \quad (3)$$

Second, the distributional form of the low dimensional joint probabilities is changed from a Gaussian, to the heavier tailed Student's t distribution with one degree of freedom. Roughly, this promotes a greater probability for moderately distanced data points in high dimensional space to be expressed by a larger distance in the low dimensional map, thus more “faithfully” representing the original distance structure, and avoiding the “crowding problem.”² The new q_{ij} is defined as

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|y_k - y_j\|^2)^{-1}}. \quad (4)$$

After incorporating the altered q_{ij} , the final gradient for the cost function is given by

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1}. \quad (5)$$

A step by step algorithm outline for t -SNE is provided in Appendix, Sec. 2.

As recommended by Hinton and van der Maaten,² PCA is first applied to the high dimensional input data in order to expedite the computation of the pairwise distances. Lastly, as t -SNE was developed primarily for 2D and 3D data representation and visualization, it is important to note that the authors warn performance of t -SNE is not well understood for the general purpose of DR.² By applying t -SNE to the CADx feature reduction problem we hope to offer at least some empirical insight toward understanding its properties in such contexts. We used van der Maaten's,⁴⁰ publicly available t -SNE MATLAB code and Intel processor optimized “fast_tsne” to generate the present data mappings.

III.E. Classifier performance estimation and evaluation

The high dimensional feature spaces DR methods were tested across all modalities for a range of lower target dimensions and user defined algorithm parameters. We evaluated the classifier performance using the area under the receiver operating curve ROC curve (AUC) via the nonparametric Wilcoxon–Mann–Whitney statistic, as calculated using the PROPROC software.^{41–43} Statistical uncertainty in classification performance due to finite sample sizes was estimated by implementing 0.632+ bootstrapping methods for training and testing the classifiers.^{44,31} Additionally, we computed the 95% empirical bootstrap confidence intervals on AUC values as estimated by no less than 500 bootstrap case set resamplings. In all values reported, the sampling was conducting on a *by lesion* basis, as there may be multiple images associated with each unique lesion. In this regard, during classifier testing, the set of classifier outputs associated with a unique lesion were averaged to produce a single value. For the supervised feature selection methods (ARD and LSW), feature selection was conducted, up to the specified number of features, on each bootstrapped sample set. Notably, the more general MCMC-BANN was coupled with both the nonlinear ARD and linear-based feature selection methods, while the linear LDA was only with the linear stepwise feature selection. As some of the calculations are computationally intensive, particularly the t -SNE mappings and MCMC-BANN training for the larger U.S. data set, a 256 CPU shared computing resource cluster was employed to accomplish runs in a feasible time frame.

IV. RESULTS

IV.A. Classification performance

MCMC-BANN and LDA classification performance is plotted as a function of the mapped or feature selected input space dimension for the three data sets, U.S., DCE-MRI, and FFDM, using the three different DR techniques, as well as the nonreduced selected features in Figs. 1(a)–1(f). Performance is characterized in terms of the 0.632+ bootstrapped AUC (left axis) and variability as gauged by the width of the empirical 95% bootstrap interval (right axis). The t -SNE perplexity was set to $Perp=30$ and Laplacian eigenmaps were generated with nearest neighbor=45 and $t=1.0$. Overall, the highest classification performance was attained by the largest sample size U.S. feature data set with the DR-MCMC-BANN just slightly eclipsing the LDA, achieving approximately $AUC_{0.632+} \sim 0.90$, while the smaller DCE-MRI and FFDM feature data produced peaks around $AUC_{0.632+} \sim 0.80$. The variability in bootstrapped AUCs is also lowest for the large U.S. data set, hovering near ~ 0.07 as the number of inputs into the classifier is increased.

A few key observations can be made from the results regarding the use of DR. Primarily, the DR techniques, for both linear (PCA) and nonlinear (t -SNE and Laplacian eigenmaps), overall, appear to at least match, or and in some cases exceed, explicit feature selection classification

FIG. 1. The 0.632+ bootstrap area under the ROC curve (AUC) (left axis) and the variation as measured by the width of the 95% empirical bootstrap confidence intervals (right axis) versus the selected feature (PCA, *t*-SNE, Laplacian eigenmap) classifier input space dimension, (a) MCMC-BANN, (b) LDA, classifier performance on the originally 81 dimensional U.S. feature data set. (c) MCMC-BANN, (d) LDA classifier performance on the originally 31 dimensional DCE-MRI feature data set, (e) MCMC-BANN, (f) LDA classifier performance on the originally 40 dimensional FFDM feature data set.

$AUC_{0.632+}$ performance. This is most evident when compared to the ARD-FS coupled with the MCMC-BANN performance across all three imaging modalities [Figs. 1(a), 1(c), and 1(e) (left axis)]. Specifically, in all cases the DR methods exhibited a more rapid rise to peak $AUC_{0.632+}$ performance and remained higher than the ARD-based feature selection for all dimension input sizes. Additionally, compared to the ARD feature selection approach, the DR methods produced less variability in the bootstrap AUC. Figures 1(a), 1(c), and 1(e) (right axis) substantially highlight this phenomenon. In particular, for the U.S. data, the ARD-FS variability, being greater than that of all the DR methods, clearly trends downward as more features are selected for input; gradually approaching the DR variability levels, yet usually remaining higher. By comparison, save for a slight increase at 1D, the DR variability is relatively consistent from 2D to 13D.

However, when coupled with the LSW feature selection, the MCMC-BANN produced more competitive results against the DR performance. For example, for this MRI data set, except for 10D and 11D, the LSW-MCMC-BANN edged above all the DR-based methods. Likewise, the use of the LSW feature selection with the MCMC-BANN resulted in substantially reduced variation in classifier performance compared to the ARD-FS. The LSW-MCMC-BANN variation nearly matched the DR output for both the U.S. and MRI across all input dimensions. For the FFDM data, except for 2D-5D, the LSW-MCMC-BANN held close to the DR variation level.

The less complex yet more stable LDA classifier [Figs. 1(b), 1(d), and 1(f) (left axis)] produced different characteristic results. In all cases the LSW feature selection performance was initially higher, however, as the dimension input space was increased, the DR methods became comparable. Expectedly, when coupled with the linear LDA, the highly nonlinear stochastic based t -SNE DR consistently underperformed. Turning to variation in the LDA [Figs. 1(b), 1(d), and 1(f) (right axis)] the LSW-FS again exhibited different behavior from ARD-FS, in that except for the smaller case sized FFDM data, variability does not considerably fluctuate moving from 1D to 13D for both the LSW-FS and DR methods.

One manner by which to concisely analyze the performance characteristics of dimension reduction/feature selection and classifiers designs for a particular data set is to plot the bootstrap cross-validation AUC against the variability. An example is provided for the U.S. feature data set in Fig. 2, with each point representing a different number of input dimensions. Data points located in the upper left corner indicate the most preferred performance qualities, i.e., higher classification performance and lower expected variability. Also provided in Fig. 3 is a plot displaying classification results for both MCMC-BANN and LDA, in terms of the bootstrap AUC for the U.S. data. Included within this plot are the empirical 95% confidence intervals to aid in gauging statistical significance for differences between estimated AUC values.

FIG. 2. Summary of the classification performance on the 81 dimensional U.S. feature data set. The 0.632+ bootstrapped area under the ROC curve versus variability as gauged by the width of the 95% empirical bootstrap confidence intervals. Each point corresponds to a different input space dimension size. Points located in the upper left corner represent the highest expected AUC as well as least expected variation in performance due to sampling.

IV.B. 2D and 3D visual representations of mappings

Due to the large sample size of the U.S. feature data, a high density of points is produced (and hence the clearest delineation of structures) in the reduced dimension mapping representations. Figures 4(a)–4(f) provides visual representations of the entire originally 81 dimensional U.S. feature data mapped into 2D and 3D Euclidean space by the unsupervised PCA, t -SNE, and Laplacian eigenmaps. The data points were subsequently colored to reflect the distribution of the lesions types (malignant tumor, benign lesion, cyst) with the reduced space.

Two key aspects are considered regarding the respective mappings: Natural class separability and overall geometric

FIG. 3. The 0.632+ bootstrapped area under the ROC curve is shown for MCMC-BANN (vertical axis) versus LDA (horizontal axis) with 95% empirical bootstrap confidence intervals included, for the originally 81 dimensional U.S. feature data set dimension reduced input or with LSW selected features.

FIG. 4. 2D and 3D visualizations of the unsupervised reduced dimension representations of the entire originally 81 dimensional breast lesion ultrasound feature data set; green data points signifying benign lesions, red: Malignant, and yellow: Benign-cystic. Visualization of linear reduction using (a) PCA, first two principal components, (b) first three principal components, 3D PCA. (c) 2D and (d) 3D visualization of the nonlinear reduction mapping using t-SNE. (e) 2D and (f) 3D visualization of the nonlinear mapping using Laplacian eigenmaps.

FIG. 5. 3D visualization of the unsupervised local structure preserving nonlinear dimension reduction representation using Laplacian eigenmaps on breast lesion feature data, (a) 3D visualization of the entire originally 31 dimensional DCE-MRI feature data, green data points signify benign lesions, red: Malignant IDC, and blue: Malignant DCIS. (b) 3D visualization of the entire originally 40 dimensional feature data, green points for benign and red for malignant lesions.

traits characteristic of the represented structures, such as smoothness and sparsity. PCA is shown in Figs. 4(a) and 4(b). Certain regions are potentially identifiable as being associated with a specific class (such as the dominance of cystic-benign points in the bottom right corner of the 2D plot); however, PCA generates a relatively homogeneous, nearly spherical distribution of points. Reflective of its mathematical basis, PCA representations provide primarily global information content, lacking the capability to represent rich local data structure. t -SNE generates a dramatically different type of low dimensional representation. As shown in Figs. 4(c) and 4(d), t -SNE produces a highly nonlinear, jagged, and highly sparse data mapping. Many isolated “islandlike” subgroupings are identifiable in the t -SNE visual representations. As predicted by the high classification performance even for 2D and 3D, t -SNE manages to clearly capture inherent class structure associations. Lastly, the Laplacian eigenmap [Figs. 4(e) and 4(f)] creates globally sparse yet locally smooth representations. As captured by the figures, the distinctly triangular form in 2D is revealed as a projected aspect of a more complex, yet smoothly connected 3D geometric structure. As evident by upper “ridge” of malignant (red) lesion points and broad cystic (yellow) “fin” on the left, the Laplacian eigenmap also manages to capture inherent class associations.

The FFDM and DCE-MRI visual representations are noisier than the U.S. due to the smaller sample size. A few examples are provided in Figs. 5(a) and 5(b). The MRI data set clearly exhibits a sparse arclike geometric structure using the Laplacian eigenmap. This structure seemingly separates the bulk of benign (green) lesions from the IDC (red) while dispersing the DCIS (blue) cases in between.

V. DISCUSSION

V.A. Dimension reduction in CADx

Three major conclusions can be made regarding the use of DR techniques in breast CADx from this study. First, and most importantly, information critical for the classification of breast mass lesions contained within the original high dimensional CADx feature vectors is not destroyed by applying the unsupervised, nonlinear DR and representation techniques of t -SNE and Laplacian eigenmaps. This observation is strongly supported by the robustness of the classification performance across the three different imaging modalities, U.S., DCE-MRI, and FFDM.

Second, according to the statistical resampling validation methods, the DR-based classification performance characteristics appear to potentially rival or in some cases exceed that of traditional feature selection based techniques. Additionally, both the linear PCA and nonlinear t -SNE and Laplacian eigenmap methods often generated “tighter” 95% empirical bootstrap intervals, implying reduced variance in classifier output, as compared to the feature selection based approaches, especially ARD (see Fig. 1). For instance, in the large U.S. data set, the performance for 13 ARD selected features was $AUC_{0.632+}=0.88$ with 95% empirical bootstrap interval $[0.787;0.895]$ and for four LSW selected features was $AUC_{0.632+}=0.87$ with interval $[0.817;0.906]$ compared to 4D t -SNE mapping (from the original 81D feature space) giving $AUC_{0.632+}=0.90$ with interval $[0.847;0.919]$. These findings imply that the generally nonlinear manifold, on which U.S. feature data exist, embedded in four dimensional Euclidean space can adequately represent the critical information for classification. These results build evidence for some potential benefits of employing the information pre-

serving DR techniques in place of explicit feature selection, including the avoidance of the curse of dimensionality.

Third, the nonlinear DR techniques generated visually rich embedded mappings with a geometric structure that often presented sparse separation between class categories, as demonstrated in Fig. 4(b): Malignant, benign, cyst, and Fig. 5(a): Benign, DCIS, IDC. The natural class associations visible in the mappings are not totally unexpected since, as explored above, the classification performance results clearly demonstrate the reduced mapping's capacity to retain sufficient information for class discrimination. The large sample number of the U.S. data set provided the most vivid visualizations, highlighting both the geometric forms and sparse quality of the nonlinear embeddings. Although PCA retained high supervised classification performance, unlike the nonlinear Laplacian eigenmaps and *t*-SNE embeddings [Figs. 4(d) and 4(f)], PCA is not capable of adequately representing the data's inherent local structural properties [Fig. 4(b)], leading to less informative visualizations. Yet, the two nonlinear methods offer distinct perspectives on the data structures. The Laplacian eigenmap appears to perhaps frame the lesions in a more globally smooth context as evidenced by the gradual transitions between distant regions of the geometric form, whereas *t*-SNE creates many distinct jagged "islands" of clustered lesion points. These emergent characteristics reflect the theoretically motivated principles driving the respective nonlinear DR algorithms.

V.B. Reduction method parameters

We briefly explored the impact of the parameter selection toward performance and visual appearance. To our knowledge, there is no principled way to optimally select a parameter configuration, thus we simply choose parameters that gave reasonable mappings as discernible in the 2D/3D representations. This is a problem in general for many unsupervised techniques. In fact, as *t*-SNE creators noted,² the method was primarily considered for visualization purposes and not explicitly for DR beyond three dimensions. Performance of *t*-SNE is not well understood for the general purpose of DR and subsequent classification. Future work may be of interest to discover procedures for identifying optimal or "near-optimal" subsets of parameters for CADx or similar machine learning purposes.

V.C. Classifiers and feature selection

In considering classifier design, one desires to be "as simple as possible, but no simpler," meaning the most robust scheme in terms of both performance and stability (low variability in performance between different samples from the same underlying distribution), all while attempting to constrain the number of parameters, namely, the input space dimension. Additionally, simpler models facilitate future repeatability with new contexts and data sets. The degree to which such pursuits are successful is dependent on the interplay of the three main aspects affecting the performances of the classifiers including sample size, data complexity, and model complexity/regularization. Naturally included within

the scope of the model complexity/regularization is the choice of inputs to the classifier, whether in the form of DR mappings or a set of selected features, as this also critically influences ultimate classification capability. Ideally, any classifier's aim is to synthesize the information available from the input space in a complete and unbiased fashion toward accomplishing the decision task. In general, classification of new input based on finite training data set is an "ill-posed" problem, and regardless of the sophistication of regularization employed, instability may persist.¹⁵ For these reasons, both the LDA and MCMC-BANN were investigated. By spanning over three different imaging modalities of varying data set size, using two different classifiers, and employing three different feature space approaches, all three of these key concepts (sample size, sample complexity, and model complexity) were touched in the course of this investigation.

For the relatively large U.S. data set, with 1126 unique lesions making up 2956 lesion images, some of the relative strengths associated with the more general, nonlinear MCMC-BANN were particularly apparent. Specifically, the MCMC-BANN, when paired with either the DR techniques or LSW-FS was able to achieve high $AUC_{0.632+}$ performance, even at low input space dimensions, as seen in Fig. 1(a). This is in part due to the MCMC-BANN ability to generalize to any target distribution, yet remain relatively well-regularized, thereby avoiding "overfitting" and severe underperformance on testing data. Yet, critically, when relying on explicit feature selection, across all input space dimension sizes for the FFDM and MRI data, and when fewer than nine features were selected for the U.S. data, the MCMC-BANN's success was contingent upon the use of LSW-FS over ARD-FS. The MCMC-BANN severely underperformed when coupled with the ARD-FS, especially when limited to picking only a few features. The smaller $AUC_{0.632+}$ and higher bootstrap variability (most dramatically evident for the lower input space dimensions), reveals limitations in ARD-FS ability to consistently identify smaller sub-sets of features capable of robustly contributing to the classification task. This limitation may be in part due to ARD's capacity for discovering nonlinear associations, which may vary highly between different bootstrapped subsamples, as well as its less direct approach (compared to LSW) in feature determination.

Turning to LDA, while not best suited to model the nonlinear DR mappings, the robustness and stability of LDA shines when joined with LSW-FS for classification purposes. LDA is, in a sense, naturally regularized by its linear nature and thus automatically avoids severe overfitting situations. Often, the relative advantage of a more complex classifier, such as MCMC-BANN, over LDA, may begin to erode as sample size decreases, even if the underlying distribution is not completely linear in nature. These phenomena are apparent for the much smaller FFDM (245 unique cases on 735 images) and DCE-MRI (356 unique lesions/images) data sets, as the less sophisticated LDA often produced the highest $AUC_{0.632+}$ values. The LDA classifier showed the greatest strength with the MRI data, nearly matching the LSW-MCMC-BANN and similarly for the DR approaches.

Furthermore, in examining Fig. 2 again, among points falling within desirable performance specifications (upper left hand corner: High classification performance/lower expected variability), it is reasonable to favor configurations which require the lowest input space dimensionality, as discussed previously (either the number selected features or target embedded mapping dimensions). A potential advantage of DR is that it may reduce the amount of necessary parameters (not including the unsupervised transformation characterized by the data itself) required to form a satisfactory data representation suitable for robust classification. In fact, most motivation for performing DR is lost if the target dimension is not considerably lower than the original high dimensional space. This is because such mapped representations become less efficient compared to simply making use of the original feature space or selected subspace as dimensions are added. Thus, within the framework of these criteria, in reviewing the results from the three modalities on whole, one may postulate that as an overall strategy, 4D *t*-SNE appears likely to produce competitive classification performance when used as input into a nonlinear classifier such as the MCMC-BANN. Such classification performance coupled with the intriguing 2D and 3D visualizations of the overall data structure may evoke attractive research potential.

In practice, it should be noted that with the sole intention of maximizing classification performance based on finite sample training data, there may be no clear advantage for use of DR techniques over traditional feature selection. Although, again, due to the curse of dimensionality, as the input space dimension for classification becomes higher in dimension, eventually cross-validation based performance will stagnant or even begin to regress lower. This occurs as the data set sample size is not sufficient to adequately isolate a unique classifier solution (as many, potentially infinite, become possible) and marginal, if not none at all, new information is gained by the additional dimensions. Thus, for these reasons and in order to compare each data set on common ground, the tests were limited to 1D–13D.

VI. CONCLUSION

The ability to capture high dimensional data structure in a human interpretable low dimensional representation is a powerful research tool. The above findings strongly suggest the relevance of nonlinear DR and representation techniques to future CADx research. DR cannot be expected to replace the benefits of feature selection based approaches in many cases. Yet, these techniques, in addition to competitive classification performance, do offer complementary information and a fresh perspective on interpreting the overall structure of the feature data. Of interest to future studies is to further investigate the origin, meaning, and physical interpretation of the discovered structures present in the CADx lesion data as revealed by these nonlinear, local geometry preserving representations. Such rich data structure representations may offer novel insights and useful understandings of clinical CADx image data.

ACKNOWLEDGMENTS

This work is partially supported by U.S. DoD Grant No. W81XWH-08-1-0731, from the U.S. Army Medical Research and Material Command, NIH Grant No. P50-CA125138, and DOE Grant No. DE-FG02-08ER6478. The authors would like to gratefully acknowledge Lorenzo Pesce, Richard Zur, Jun Zhang, and Partha Niyogi for their thoughtful discussion and insightful suggestions. Additionally, the authors thank Weijie Chen contributing breast MRI feature data. The authors are grateful to Geoffrey Hinton and Laurens van der Maaten for freely distributing their algorithm code as well as the very handy dimension reduction MATLAB toolbox. We would also like to gratefully acknowledge the SIRAF shared computing resource, supported in parts by NIH Grant Nos. S10 RR021039 and P30 CA14599, and its excellent administrator, Chun-Wai Chan. And lastly the authors thank the reviewers for their useful suggestions.

M.L.G. is a stockholder in R2 Technology/Hologic and received royalties from Hologic, GE Medical Systems, MEDIAN Technologies, Riverain Medical, Mitsubishi and Toshiba. It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interest that would reasonably appear to be directly and significantly affected by the research activities.

APPENDIX: DIMENSION REDUCTION ALGORITHMS

I. Laplacian eigenmaps algorithm outline

Beginning with k input points $x_1, \dots, x_k$ in R^l :

Step 1: Construct the adjacency graph. Generate a graph with edges connecting nodes i and j if x_i and x_j are “close.” Closeness is defined by the nodes included in the N nearest neighbors. This relation is naturally symmetric between points i and j . The parameter N must be selected.

Step 2: Choosing weight. The “heat kernel” is used to assign weights to edge connected nodes i and j : $W_{ij} = \exp(-\|x_i - x_j\|^2/t)$. Otherwise, use $W_{ij}=0$ for unconnected vertices. See Belkin and Niyogi¹ for kernel justification. The parameter t is user defined. If t is set very high or approximately $t=\infty$, the edge connected node weights are essentially $W_{ij}=1$. This option can be used to avoid parameter selection.

Step 3: Computing eigenmaps. Assuming a connected graph G generated in step 1, solve for the following eigenvector and eigenvalues: $Lf = \lambda Df$, where D is the diagonal weight matrix, defined by summing over the rows of W . $D_{ii} = \sum_j W_{ij}$, and L is the Laplacian matrix defined as $L = D - W$. Symmetric and positive semidefinite, conceptually, the Laplacian matrix acts as an operator on functions defined by graph G 's vertices. Solving the equation, let $f_0, \dots, f_{k-1}$ be the eigenvectors, arranged in accordance to their eigenvalues $0 = \lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_k$. $Lf_0 = \lambda_0 Df_0, \dots, Lf_{k-1} = \lambda_{k-1} Df_{k-1}$.

Finally, the k input data points in R^l are embedded in m

dimensional Euclidean space using the m eigenvectors after the zero eigenvalued f_0 , $x_i \rightarrow (f_1(i), \dots, f_m(i))$.

II. t-SNE algorithm outline

Beginning with k input points $\{x_1, \dots, x_k\}$ in R^l , set perplexity parameter $Perp$, number of iterations T , learning rate η , and momentum $\alpha(t)$.

Step 1: Compute similarities. Compute pairwise p_{ji} probabilities using the σ_i found with perplexity $Perp$, and use symmetrized conditional probability distributions $p_{ij} = (p_{ji} + p_{ij})/2k$.

Step 2: Initialize solution sample. Sample from $N(0, 10^{-4}I^m)$ for initial points $\{y_1, \dots, y_k\}$.

Step 3: Execute T , update iterations on Y . Compute low dimension similarities q_{ij} using Eq. (4) and gradient using Eq. (5). Update Y using $Y^{(t)} = Y^{(t-1)} + \eta \delta C / \delta y_i + \alpha(t) \times (Y^{(t-1)} - Y^{(t-2)})$. Output: Low dimension mapping $\{y_1, \dots, y_k\}$ in R^m .

^{a)}Electronic mail: andrewj@uchicago.edu

¹M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.* **15**, 1373–1396 (2003).

²L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

³M. L. Giger, H. Chan, and J. Boone, "Anniversary Paper: History and status of CAD and quantitative image analysis: The role of medical physics and AAPM," *Med. Phys.* **35**, 5799–5820 (2008).

⁴Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," *Acad. Radiol.* **5**, 155–168 (1998).

⁵M. A. Kupinski and M. Giger, "Automated seeded lesion segmentation on digital mammograms," *IEEE Trans. Med. Imaging* **17**, 510–517 (1998).

⁶Z. Huo, M. L. Giger, C. J. Vyborny, U. Bick, P. Lu, D. E. Wolverton, and R. A. Schmidt, "Analysis of spiculation in the computerized classification of mammographic masses," *Med. Phys.* **22**, 1569–1579 (1995).

⁷K. Drukker, M. L. Giger, K. Horsch, M. A. Kupinski, C. J. Vyborny, and E. B. Mendelson, "Computerized lesion detection on breast ultrasound," *Med. Phys.* **29**, 1438–1446 (2002).

⁸W. Chen, M. L. Giger, U. Bick, and G. M. Newstead, "Automatic identification and classification of characteristic kinetic curves of breast lesions on DCE-MRI," *Med. Phys.* **33**, 2878–2887 (2006).

⁹W. Chen, "Computerized interpretation of breast MRI: Investigation of enhancement-variance dynamics," *Med. Phys.* **31**, 1076 (2004).

¹⁰K. Drukker, M. L. Giger, C. J. Vyborny, and E. B. Mendelson, "Computerized detection and classification of cancer on breast ultrasound," *Acad. Radiol.* **11**, 526–535 (2004).

¹¹M. Giger, "Computer-aided diagnosis of breast lesions in medical images," *Comput. Sci. Eng.* **2**, 39–45 (2000).

¹²G. D. Tourassi, B. Harrawood, S. Singh, J. Y. Lo, and C. E. Floyd, "Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms," *Med. Phys.* **34**, 140–150 (2007).

¹³Y. Yuan, M. L. Giger, H. Li, K. Suzuki, and C. Sennett, "A dual-stage method for lesion segmentation on digital mammograms," *Med. Phys.* **34**, 4180–4193 (2007).

¹⁴K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. (Academic, Boston, 1990).

¹⁵C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2006).

¹⁶S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Comput.* **4**, 1–58 (1992).

¹⁷B. Sahiner, H. Chan, N. Petrick, R. F. Wagner, and L. Hadjiiski, "Feature selection and classifier performance in computer-aided diagnosis: The effect of finite sample size," *Med. Phys.* **27**, 1509–1522 (2000).

¹⁸M. A. Kupinski and M. L. Giger, "Feature selection with limited datasets," *Med. Phys.* **26**, 2176–2182 (1999).

¹⁹W. Chen, R. M. Zur, and M. L. Giger, "Joint feature selection and clas-

sification using a Bayesian neural network with automatic relevance determination priors: Potential use in CAD of medical imaging" in *Medical Imaging 2007: Computer-Aided Diagnosis*, 2007, edited by M. Giger and N. Karssemeijer [Proc. SPIE **6514**, 65141G–65141G–10 (2007)].

²⁰M. A. Anastasio, H. Yoshida, R. Nagel, R. M. Nishikawa, and K. Doi, "A genetic algorithm-based method for optimizing the performance of a computer-aided diagnosis scheme for detection of clustered microcalcifications in mammograms," *Med. Phys.* **25**, 1613–1620 (1998).

²¹G. D. Tourassi, E. D. Frederick, M. K. Markey, and J. Floyd, "Application of the mutual information criterion for feature selection in computer-aided diagnosis," *Med. Phys.* **28**, 2394–2402 (2001).

²²Y. Wang, D. J. Miller, and R. Clarke, "Approaches to working in high-dimensional data spaces: Gene expression microarrays," *Br. J. Cancer* **98**, 1023–1028 (2008).

²³H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.* **24**, 498–520 (1933).

²⁴M. Kirby, *Geometric Data Analysis: An Empirical Approach to Dimensionality Reduction and the Study of Patterns* (Wiley, New York, 2000).

²⁵K. Drukker, N. P. Grusauskas, and M. L. Giger, "Principal component analysis, classifier complexity, and robustness of sonographic breast lesion classification," in *Medical Imaging 2009: Computer-Aided Diagnosis*, 2009, edited by M. Giger and N. Karssemeijer [Proc. SPIE **7260**, 72602B–72602B6 (2009)].

²⁶C. Varini, A. Degenhard, and T. W. Nattkemper, "Visual exploratory analysis of DCE-MRI data in breast cancer by dimensional data reduction: A comparative study," *Biomed. Signal Process. Control* **1**, 56–63 (2006).

²⁷A. Madabhushi, P. Yang, M. Rosen, and S. Weinstein, "Distinguishing lesions from posterior acoustic shadowing in breast ultrasound via nonlinear dimensionality reduction," *Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE* **1**, 3070–3073 (2006).

²⁸M. K. Markey, J. Y. Lo, G. D. Tourassi, and C. E. Floyd, "Self-organizing map for cluster analysis of a breast cancer database," *Artif. Intell. Med.* **27**, 113–127 (2003).

²⁹K. Drukker, K. Horsch, and M. L. Giger, "Multimodality computerized diagnosis of breast lesions using mammography and sonography," *Acad. Radiol.* **12**, 970–979 (2005).

³⁰H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," *Phys. Med. Biol.* **40**, 857–876 (1995).

³¹B. Sahiner, H. Chan, and L. Hadjiiski, "Classifier performance prediction for computer-aided diagnosis using a limited dataset," *Med. Phys.* **35**, 1559 (2008).

³²R. M. Neal, *Bayesian Learning for Neural Networks* (Springer-Verlag, New York, 1996).

³³M. A. Kupinski, D. Edwards, M. Giger, and C. Metz, "Ideal observer approximation using Bayesian classification neural networks," *IEEE Trans. Med. Imaging* **20**, 886–899 (2001).

³⁴M. E. Tipping, *Advanced Lectures on Machine Learning* (Springer, Berlin/Heidelberg, 2004), pp. 41–62.

³⁵I. Nabney, *Netlab* (Springer-Verlag, London, Berlin, Heidelberg, 2002).

³⁶E. Levina and B. Bickel, *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, 2005).

³⁷M. Belkin and P. Niyogi, "Towards a theoretical foundation for Laplacian-based manifold methods," *J. Comput. Syst. Sci.* **74**, 1289–1308 (2008).

³⁸L. van der Maaten, "MATLAB toolbox for dimensionality reduction" (2008).

³⁹G. Hinton and S. Roweis, *Advances in Neural Information Processing Systems 15* (MIT Press, Cambridge, 2003), pp. 833–840.

⁴⁰L. van der Maaten, "t-SNE Files" (2008).

⁴¹L. L. Pesce and C. E. Metz, "Reliable and computationally efficient maximum-likelihood estimation of 'proper' binormal ROC curves," *Acad. Radiol.* **14**, 814–829 (2007).

⁴²C. E. Metz, "Basic principles of ROC analysis," *Semin. Nucl. Med.* **8**, 283–298 (1978).

⁴³J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology* **143**, 29–36 (1982).

⁴⁴B. Efron and R. Tibshirani, "Improvements on cross-validation: The 632+ bootstrap method," *J. Am. Stat. Assoc.* **92**, 548–560 (1997).