Repeatability in computer-aided diagnosis: Application to breast cancer diagnosis on sonography

Karen Drukker,^{a)} Lorenzo Pesce, and Maryellen Giger

Department of Radiology, The University of Chicago, 5841 S. Maryland Ave., MC 2026 Chicago, Illinois 60637

(Received 2 October 2009; revised 14 April 2010; accepted for publication 17 April 2010; published 19 May 2010)

Purpose: The aim of this study was to investigate the concept of repeatability in a case-based performance evaluation of two classifiers commonly used in computer-aided diagnosis in the task of distinguishing benign from malignant lesions.

Methods: The authors performed .632+ bootstrap analyses using a data set of 1251 sonographic lesions of which 212 were malignant. Several analyses were performed investigating the impact of sample size and number of bootstrap iterations. The classifiers investigated were a Bayesian neural net (BNN) with five hidden units and linear discriminant analysis (LDA). Both used the same four input lesion features. While the authors did evaluate classifier performance using receiver operating characteristic (ROC) analysis, the main focus was to investigate case-based performance based on the classifier output for individual cases, i.e., the classifier outputs for each test case measured over the bootstrap iterations. In this case-based analysis, the authors examined the classifier output variability and linked it to the concept of repeatability. Repeatability was assessed on the level of individual cases, overall for all cases in the data set, and regarding its dependence on the case-based classifier output. The impact of repeatability was studied when aiming to operate at a constant sensitivity or specificity and when aiming to operate at a constant threshold value for the classifier output.

Results: The BNN slightly outperformed the LDA with an area under the ROC curve of 0.88 versus 0.85 ($p < 0.05$). In the repeatability analysis on an individual case basis, it was evident that different cases posed different degrees of difficulty to each classifier as measured by the by-case output variability. When considering the entire data set, however, the overall repeatability of the BNN classifier was lower than for the LDA classifier, i.e., the by-case variability for the BNN was higher. The dependence of the by-case variability on the average by-case classifier output was markedly different for the classifiers. The BNN achieved the lowest variability (best repeatability) when operating at high sensitivity ($>90\%$) and low specificity ($<66\%$), while the LDA achieved this at moderate sensitivity ($\sim 74\%$) and specificity ($\sim 84\%$). When operating at constant 90% sensitivity or constant 90% specificity, the width of the 95% confidence intervals for the corresponding classifier output was considerable for both classifiers and increased for smaller sample sizes. When operating at a constant threshold value for the classifier output, the width of the 95% confidence intervals for the corresponding sensitivity and specificity ranged from 9 percentage points (pp) to 30 pp.

Conclusions: The repeatability of the classifier output can have a substantial effect on the obtained sensitivity and specificity. Knowledge of classifier repeatability, in addition to overall performance level, is important for successful translation and implementation of computer-aided diagnosis in clinical decision making. © 2010 American Association of Physicists in Medicine.

[DOI: [10.1118/1.3427409](https://doi.org/10.1118/1.3427409)]

Key words: computer-aided diagnosis, ultrasound, repeatability

I. INTRODUCTION

Computer-aided diagnosis (CADx) has been an area of active research, not only within breast imaging¹⁻⁴ but also within other medical imaging subdisciplines including chest imaging.⁵⁻⁷ The aim of CADx is to provide the radiologist with a second opinion on the characterization of an abnormality found in a medical imaging exam. The radiologist indicates, i.e., detects, abnormalities of interest and computerized analysis is performed based on those identifications and the image data. It is important to note that computer-

aided detection—(CADE), in which the computerized analysis involves a localization task—will not be discussed in this paper. In the case of CADx (a characterization task), the computerized analysis provides a metric for the likelihood that an abnormality indicated by a radiologist is, in fact, the disease of interest. In this paper, we will focus on CADx for the diagnosis of breast cancer on sonographic breast images. Depending on the type of classifier used in the computerized analysis, the CADx output may be a probability of malignancy, a level of suspicion score, or a related metric. Note

that here we defined CADx in a strict sense in that we only consider a system to be a true CADx system if it provides the user with numbers related to the likelihood of disease. At this time, there are quite a few FDA-approved CADe systems in clinical use but no FDA-approved CADx systems, although there are a few commercial systems available that may help radiologists in their diagnoses such as the B-CAD[®] system (Medipattern Corporation, Toronto, Canada).

A widely used and accepted method for meaningful evaluation of a binary diagnostic technique (normal versus abnormal, cancer versus benign, etc.) is receiver operating characteristic (ROC) analysis.^{8,9} ROC analysis provides information about the performance based on the distribution of the relative ranks of correct and incorrect decisions, or just the average relative rank in the case of the area under the curve (AUC).¹⁰ While the standard error of the AUC value provides an estimate of the uncertainty in a cumulative manner, ROC analysis does not provide detailed information on a case-by-case basis or metrics related to the specific scale used in the analysis. It is a reasonable assumption that CADx has the potential to help radiologists as long as it provides information that in some way differs from the information that radiologists can extract on their own. In order to achieve such synergy between radiologist and CADx, knowledge of the actual classifier output behavior and predictability is essential. Indeed, two classifiers with very similar stand-alone performances in terms of areas under the ROC curve may have dramatically different impact when used as reading aids. For example, a CADx scheme which is accurate on trivial cases is much less likely to improve the radiologists' performance than a scheme suited for more difficult cases.¹¹ In spite of its importance, detailed case-by-case performance analysis within CADx has, to date, received very limited attention and performance analyses have been generally based on ROC analysis methods. In the past, we have investigated the robustness of our CADx scheme for breast ultrasound with respect to different radiologists,¹² different imaging sites,¹³ different manufacturers' scanners,¹⁴ different image selection protocols,¹⁵ and different populations¹⁶ using widely accepted ROC analysis methods. While some other authors have indicated that it may be important to investigate classifier behavior in a more detailed manner beyond the summary performance measure of "area under the ROC curve,"¹¹ to our knowledge there currently is no "gold standard" on how to do this within CADx.

We believe that a more detailed case-by-case performance analysis is of importance since, in the future, radiologists will likely interact in clinical practice with a CADx scheme on a case-by-case basis and interpret cases using specific values of the CADx output. For example, a radiologist may interpret a CADx output score of z larger than a given value z_θ to indicate that the case is deemed malignant by the CADx and a score of z' smaller than the value z_θ to indicate it as benign. Hence, it is desirable for a CADx scheme to not only achieve high and stable, i.e., reliable, performance within the ROC paradigm, but to also give very similar output for the same case when comparable classifier training scenarios are used. (There is, of course, no reason to expect similar output

when the training sets are sampled from different populations or obtained using an incompatible sampling scheme.)

In this work, we investigated the performance of two commonly used classifiers for the task of distinguishing benign from malignant breast lesions on breast ultrasound. While, for completeness, classifier performance was also assessed within the ROC framework, our main focus was to investigate by-case performance. Simply put, when a classifier indicated that the level of suspicion for an unknown case equaled z , was it really z or might it as well have been $z + \Delta z$ (and how large was Δz)? In the assessment of the by-case classifier performance we used the concept of *repeatability*, as explained in Sec. II. The secondary focus of this paper was the impact of classifier by-case performance on the potential clinical usefulness of CADx schemes in terms of the obtained sensitivity and specificity. In plain words (for the CADx schemes considered here), how "believable" was the CADx output and what, if any, effect would this have on the resulting CADx benign versus malignant assessment?

It is important to note that this study was designed to yield insights in by-case classifier behavior, not to determine which classifier is best for CADx, whether applied to breast sonography or in general. Also, only stand-alone classifier performance was considered here, and the impact on the use of CADx as a reader-aid remains a topic of future investigations.

II. MATERIAL AND METHODS

II.A. Patient data

All patient data used in this work were collected under Internal Review Board approved HIPAA compliant protocols. We employed a database of diagnostic breast ultrasound images acquired in our Radiology Department with a Philips[®] HDI 5000 scanner containing images of 1251 cases (breast lesions) with a total of 2855 images. Of these cases, 212 were cancers and 1039 were benign lesions, bringing the cancer prevalence in this data set to 17% (212/1251). The majority of this database (1126 cases including 158 cancers) consisted of sonographic images from patients who presented for diagnostic breast ultrasound after a potential abnormality was noted on mammography. The remainder of the database (125 lesions including 54 cancers) was acquired as part of the high-risk screening protocol at our site for patients for whom a potential abnormality was first noted on breast MRI. There were no other selection criteria for inclusion in the database and it forms a realistic representation of patients seen in our diagnostic ultrasound clinic. Several U.S. views are typically acquired for each physical lesion (usually at least two perpendicular views) and the number of images per lesion ranged from one to 20 with an average of 2.3 images per lesion (2855/1251) for the data set used here. It is important to note that all computer analysis was performed on a "by lesion" basis, i.e., on a "by-case" basis, *not* on a "by-image" basis, and the computer outputs obtained from different views of a given lesion were averaged before any further analysis. The pathology of all malignant lesions was

confirmed by either biopsy or surgical excision. The pathology of the benign lesions was confirmed either by biopsy or by imaging follow-up. The average imaging follow-up time before inclusion in the database was 3 yr (range 2–5 yr). Note that throughout this paper, “case” refers to a physical breast lesion.

II.B. Computer-aided diagnosis scheme

The computer-aided diagnosis (lesion characterization) scheme employed here has been described extensively in previous publications.^{12,17,18} Since we were interested in lesion characterization here (not detection), the only input the CADx scheme required was the image data and the approximate location of the lesion center for each image. In this work, the locations of lesion centers corresponded to the geometric center of lesion outlines as drawn by an expert Mammography Quality Standards Act certified radiologist with over 11 years of experience. Reader variability was not considered here.

The subsequent analyses of the lesions were all automatic. After a preprocessing stage, lesions were automatically segmented from the parenchymal background using a cost function-based contour optimization technique.¹⁸ To facilitate lesion characterization, four lesion descriptors were extracted for each lesion describing the lesion shape, margin, posterior acoustic behavior, and texture.¹⁷ These lesion descriptors, i.e., lesion features, were chosen based on their successful use in past CADx applications. Our use of these features here is unlikely to *a priori* favor one classifier over the other since these features were selected in the past based on lesion characteristics used by radiologists in clinical practice¹⁷ and their performance was at that time evaluated using images obtained with older equipment than used in this study. No feature selection was performed for the purpose of this paper.

The lesion characterization, i.e., classification, task of interest was the distinction between benign and malignant lesions. We investigated two classifiers for this purpose: (1) Linear discriminant analysis (LDA)¹⁹ and (2) a Bayesian Neural Net (BNN)²⁰ with five hidden units. The four lesion descriptors formed the input to the classifiers. The classifier output provided an estimate for the level of suspicion for each lesion. Here, the output of the BNN is a measure for the probability of malignancy of a lesion (given the cancer prevalence in the training set), while such an interpretation is lacking for the LDA score. In this paper we used the “raw” LDA output since conversion of this output to probabilities of malignancy or other output metrics was not of interest here.

II.C. Training and testing protocols

We performed classifier training and testing within several 0.632 bootstrap analyses (Table I)^{21,22} (the .632+ bootstrap differs from the 0.632 bootstrap only in the bias correction of the AUC value, which will be discussed in Sec. II D). Each iteration of a 0.632 (or .632+) bootstrap training/testing protocol, cases are randomly selected to be part of the training

TABLE I. Training/testing protocols for .632+ bootstrap experiments with fixed 17% cancer prevalence throughout.

Sample size (<i>n</i>)	Number of bootstrap iterations (<i>I</i>)	Average number of times each case (of <i>N</i> =1251) was part of the test set
1251	1000	368
1251	500	184
625	1000 ^a	184
250	2500 ^a	184

^aFor each bootstrap iteration, a subset (size *n*) was randomly selected from the overall data set (size *N*=1251).

set, and those cases not selected form an independent test set. The 0.632 bootstrap is constructed in such a way that, on average, for each iteration 63.2% of the cases serve as the training set and the remaining 36.8% serve as the test set. We performed sampling in such a way that in all bootstrap iterations, the cancer prevalence in the training set (and testing set) was equal to that in the full data set (17%). We first verified that the number of bootstrap iterations *I* was sufficient by investigating *I*=500 and *I*=1000 bootstrap iterations when sampling from the database of *N*=1251 cases in a “straightforward” 0.632 manner. Furthermore, we investigated the effect of sample size *n* by randomly selecting *each iteration* samples of size *n* from the database and performing 0.632 bootstrap training and testing within these randomly selected subsets. The number of bootstrap iterations was adjusted so that each case in the database (*N*=1251) was, on average, part of the test set as frequently as for the bootstrap analysis with the *N*=*n*=1251, and *I*=500 (Table I). Note that all analysis was performed on a by-case basis such that all lesion views (images) of a given case were either all part of the training set or all part of the test set in a given bootstrap iteration, and that the classifier outputs obtained from different views of a given case were averaged to form the case-based classifier output.

II.D. Statistical analysis

We assessed classifier performance both on a “global” scale using ROC analysis⁸ and on a more “local” case-based scale investigating the repeatability²³ of classifier output. We also assessed the impact of repeatability on the obtained sensitivity and specificity.

II.D.1. ROC analysis

For all training/testing protocols, the area under the ROC curve (AUC) and 95% confidence interval thereof, was estimated from the nonparametric Wilcoxon areas using the .632+ bootstrap correction.^{21,22} Note that in this work, we used a nonparametric approach and all estimates including confidence intervals were obtained empirically from the bootstrap samples. For a two-sided superiority test, a 95% confidence interval that contains zero signifies a failure to

TABLE II. Overview of case-based performance analysis performed in this work.

Repeatability	
Level of analysis	Metric
Individual case	Standard deviation, σ , of classifier output for individual test cases
Overall (for entire data set)	Mean, standard deviation and histogram of case-based σ 's
Dependence on classifier output	Case-based σ as a function of case-based mean value of classifier output (<i>repeatability profile</i>)
Impact of repeatability	
Operating condition	Impact
Operating at constant sensitivity or specificity	Resulting median and 95% confidence intervals of the corresponding classifier output value
Operating at constant threshold value for classifier output	Resulting (sensitivity, specificity) operating points

achieve statistical significance for ($\alpha=0.05$ level) and an interval excluding zero indicates statistical significance ($\alpha=0.05$ level).

II.D.2. By-case performance analysis

For the local by-case performance, it is informative to relate the analysis to the range in classifier output values. Since the output range of the LDA classifier was not known *a priori*, while that of the BNN classifier was by definition bound between 0 and 1, we *a posteriori* determined the 95% confidence interval for the output values for each classifier and its width, r for the LDA and r' for the BNN classifier, respectively. We chose to use 95% confidence intervals to define outliers (and eliminate them from plots) and these intervals were determined *a posteriori* for each classifier based on *all* experiments. We will refer to the width of the 95% confidence intervals for the range in classifier output (r and r' , respectively) as the “width of the output range” in the remainder of the manuscript.

In this paper we will use the concept of repeatability²³ to assess the classifier by-case performance. Here, a classifier will be deemed *repeatable* if it consistently gives the same output value for a given case under different classifier training scenarios. In other words, a “high repeatability” means a “low variability” in classifier output on a case-by-case basis. Repeatability is a concept closely related to variability, reproducibility, and precision, and in this paper we chose to mainly use “repeatability” as the terminology of choice since our analysis is loosely based on the work presented by Bland.²³ The two main foci of the local by-case performance analysis were (1) repeatability and (2) its impact on the potential clinical usefulness of CADx schemes in terms of sensitivity and specificity as summarized in Table II and detailed below.

II.D.2.a. Repeatability. In the work of Bland,²³ it was shown that repeatability of two series of measurements for J subjects can be accurately assessed by plotting the observed difference between the measurements for each subject

j ($j=1, \dots, J$) versus the observed mean of those measurements (rather than by calculating the correlation coefficient between the two series of measurements as is often erroneously done). Since in our work each case served in the testing capacity multiple times (Table I), rather than the two times in the work of Bland,²³ we assessed repeatability by calculating the standard deviation of measurements, i.e., classifier outputs, for each case.

We performed the repeatability analysis on several levels (Table II). (a) We assessed the repeatability of the classifier output for each *individual case* by calculating the case-based standard deviation σ of the classifier output for each case when used in the testing capacity in the bootstrap iterations. (b) Then we assessed the *overall*, or average, repeatability of each classifier for the entire data set by calculating the mean, standard deviation, and a histogram of the individual case-based standard deviations σ . (c) We evaluated the *dependence* of the classifier repeatability on the classifier output value by plotting the case-based standard deviation σ versus the case-based mean value of the classifier output for all test cases (analogous to the difference plots presented by Bland²³). In order to avoid “overcrowding” in these *repeatability profiles*, we performed block averaging of both the case-based mean and the case-based standard deviation σ of the classifier output over blocks of $N_{\text{block}}=60$ cases (plus a single block of the remaining 49 cases) *after* ordering all cases by their mean case-based output (and eliminating the outliers based on the 95% confidence intervals for the mean case-based output r and r' as discussed in the previous paragraph). In these block-averaged repeatability profiles, the average case-based standard deviation σ_c (and standard deviation thereof) were calculated for each block of cases and plotted as a function of the case-based mean output in the form of histogram “bins” marked, on the x -axis, by the minimum case-based mean and the maximum case-based mean for each block. Note that the use of a constant block size ($N_{\text{block}}=60$ cases) for these plots resulted in histograms with a variable bin-width and that the “skinnier” the bins, the

TABLE III. Estimates for the area under the ROC curve (AUC value) and 95% confidence interval (CI) for different sample sizes (see Table I for training/testing bootstrap protocols).

Sample size (<i>n</i>)	Iterations (<i>I</i>)	AUC for LDA mean (95% CI)	AUC for BNN mean (95% CI)	95% CI for difference in AUC (BNN versus LDA)
1251	1000	0.85 (0.82;0.87)	0.88 (0.86;0.90)	(0.017;0.047)
1251	500	0.85 (0.82;0.87)	0.88 (0.86;0.90)	(0.018;0.045)
625	1000	0.84 (0.81;0.88)	0.87 (0.83;0.91)	(−0.001;0.055)
250	2500	0.84 (0.76;0.90)	0.85 (0.74;0.92)	(−0.062;0.069)

higher the number of cases corresponding to that region of case-based mean classifier output. Furthermore, the relative malignant versus benign composition was determined within each block of $N_{\text{block}}=60$ cases.

To recap, we will use the standard deviation of the classifier output for individual cases to assess repeatability on an *individual* case basis, where a small value for the standard deviation means high repeatability and vice versa. The case-based standard deviations for all cases will be used to assess the overall classifier repeatability for the entire data set. For each classifier, in the assessment of the dependence of repeatability on the classifier output values (for this particular data set), the output values corresponding to the “best” or “highest” repeatability will be determined from the observed minimum in their respective repeatability profile. Please note again that the ranges of the classifier output, r and r' , were used *only* to generate the repeatability plots in a manner facilitating comparison. (Effectively, this means that of the $N=1251$ cases, 1189 were used to generate the repeatability plots.) The standard deviations of the classifier output will be reported as a function of r and r' , respectively, but for completeness “raw” unscaled values for the standard deviations of the classifier output will be provided in these plots as well on a second x or y -axis as appropriate.

II.D.2.b. From repeatability to sensitivity and specificity. We assessed the *impact* of repeatability on the potential clinical usefulness of CADx in two ways, both involving sensitivity and specificity measurements (Table II). These assessments were performed to provide insight in the answers to two important questions. The first important question we considered is “I want to operate at a certain sensitivity (or specificity), what value of the classifier output does this translate into?” The second important question is “If I use a constant threshold for the classifier output in my assessment of the disease status, what sensitivity and specificity am I operating at?” Hence, we performed two types of analyses: (a) We assessed the impact of operating at a constant target level of sensitivity (or specificity) on the corresponding cutoffs for the classifier output, and (b) we evaluated the impact of operating at a constant threshold value for the classifier output on the resulting (sensitivity, specificity) operating points analogous to the use of operating points in ROC analysis. In more detail, in (a) we empirically determined each bootstrap iteration, which threshold value for the classifier output resulted in a predetermined “high sensitivity” and which threshold value resulted in a “high specificity” operating condition. From all bootstrap iterations, the median

value and 95% confidence intervals for these threshold values were calculated and indicated in the repeatability plots. We then used, in (b), the median values obtained in assessment (a) for the thresholds corresponding to “high sensitivity” and “high specificity” as constant cutoffs for the classifier output in the task of determining whether a case was benign or malignant and calculated the resulting distribution in both sensitivity and specificity in terms of (sensitivity, specificity) operating points. In these assessments, we selected 90% sensitivity and 90% specificity to represent “high sensitivity” and “high specificity” operating conditions, respectively. Of course, it is the goal to operate at both high sensitivity and high specificity at the same time, but one cannot define cutoff values for the classifier output that will achieve both simultaneously.¹⁰ One should note that the values for the “target” sensitivity and target specificity of most clinical interest will be different for different CADx applications and that the values of interest here (90% sensitivity and 90% specificity, respectively) were selected as examples to gain insight in classifier behavior. It is also important to note that it would be desirable for a classifier to have the best repeatability (lowest variability) in the neighborhood of the output value that translates into the desired sensitivity or specificity.

III. RESULTS

III.A. ROC analysis

The AUC values and 95% confidence intervals thereof obtained for the analysis of $N=n=1251$ cases with $I=500$ and $I=1000$ bootstrap iterations strongly suggest that convergence of the AUC value has been achieved after $I=500$ bootstrap iterations (Table III). For these bootstrap analyses, the difference in performance obtained by the LDA and that obtained by the BNN is small but of statistical significance, with the BNN classifier outperforming the LDA classifier. This difference in performance fails to reach statistical significance for the smaller sample sizes $n=625$ and $n=250$, i.e., those 95% confidence intervals for the difference in AUC included zero (Table III). Note again, however, that we did not intend to prove one classifier better than the other and that these ROC results were mainly intended to provide a context for the repeatability analysis.

FIG. 1. The repeatability of the classifier output for a randomly selected individual case (when used in the testing capacity) as measured by the standard deviation of its output over the bootstrap iterations (recall Table II) for $N=n=1251$ cases and $I=500$ bootstrap iterations (a) for the LDA classifier and (b) for the BNN classifier. The same case was used for both histograms. The 95% CI of the classifier output for all cases determined the scale of the x-axes (r for the LDA and r' for the BNN, respectively, as indicated).

III.B. By-case performance analysis

The width of the 95% confidence interval for the range in the classifier output for *all* experiments, r (LDA) and r' (BNN), was 5.22 and 0.74, respectively (see also Fig. 1).

III.B.1. Repeatability

Focusing on the analysis using $N=n=1251$ cases and $I=500$ bootstrap iterations, the assessment of repeatability on an *individual case basis* is illustrated for a randomly selected, and in this instance malignant, case (Fig. 1). For this particular case, the LDA obtained a lower case-based σ (in terms of r) than the BNN (in terms of r'), and hence was more consistent in its benign versus malignant assessment or, in other words, the LDA output displayed better repeatability for this particular case. The LDA also performed better when assessing the overall (or average) repeatability for the entire

FIG. 2. The overall repeatability of the classifier output for the entire data set as measured by the standard deviations of the classifier output for all cases (when used in the testing capacity) (recall Table II) for $N=n=1251$ cases and $I=500$ bootstrap iterations (a) for the LDA classifier and (b) for the BNN classifier. The bottom x-axes provide values in terms of the classifier output range r and r' , respectively, and the top x-axes provide raw values.

data set (Fig. 2), with the LDA obtaining an approximately 35% lower average case-based σ (in terms of r : $0.036 r$) than the BNN (in terms of r' : $0.055 r'$) with a narrower distribution. The dependence of repeatability on the classifier output, i.e., repeatability profiles, showed marked differences between the two classifiers (Fig. 3). While the LDA obtained the best repeatability in the middle of its output range (close to zero) [Fig. 3(a)], the BNN obtained the best repeatability (lowest variability) around the minimum of its output range (also close to zero) [Fig. 3(b)]. Outside of the range $[0-r']$ used for plotting, the BNN classifier also obtained high repeatability (for output values close to one).

For the analyses with $n < N$ cases (Table I) the *case-based* repeatability and overall repeatability declined for both classifiers. For the $n=625$ analysis, the average case-based σ increased with a factor of 1.7 with respect to those observed in the $N=n=1251$ and $I=500$ analysis (Fig. 2) for both clas-

FIG. 3. Repeatability dependence on the classifier output, i.e., repeatability profiles, for the entire data set as measured by the standard deviation of the classifier output as a function of the mean of the classifier output for all cases (recall Table II), for $N=n=1251$ cases and $I=500$ bootstrap iterations (a) for the LDA classifier and (b) for the BNN classifier. The impact of operating at constant sensitivity (90%) or constant specificity (90%) on the corresponding classifier output value is indicated by the shaded areas. The repeatability profiles have been “locally averaged” by first raking all cases according to their average classifier output and subsequently averaging the case-based σ 's over blocks of $N_{\text{block}}=60$ cases. The error bars indicate ± 1 standard deviation, and the left y-axes provide values in terms of the classifier output range r and r' , respectively, and the right y-axes provide raw values. Relative malignant/benign composition within each block of $N_{\text{block}}=60$ cases ordered by mean classifier output (c) for the LDA classifier and (d) for the BNN classifier.

sifiers. For the $n=250$ analysis, this factor was 3.0 for the LDA and 2.7 for the BNN. Note that for these $n < N$ analyses, each case (of $N=1251$) served as a test case the same number of instances as for the $N=n=1251$ cases and $I=500$ iterations analysis (Table I). For the $n=625$ and $n=250$ analyses, the same trends were observed for the dependence of repeatability on the classifier output and the shape of the repeatability profiles remained the same as for the $N=n=1251$ analysis.

III.B.2. From repeatability to sensitivity and specificity

Focusing again on the $N=n=1251$ cases and $I=500$ iterations analysis, the LDA classifier obtained the best repeatability for output values close to the middle of its output range as discussed previously [as indicated in Fig. 3(a)] which translated to operating at moderate sensitivity ($\sim 74\%$) and moderate specificity ($\sim 84\%$). The BNN

achieved the best repeatability at output values close to the minimum in its output range [as indicated in Fig. 3(b)] corresponding to operating at high sensitivity ($>90\%$) and low specificity ($<66\%$). Note again that it would be desirable for a classifier to have the best repeatability (lowest variability) in the neighborhood of the output value that translates into the desired sensitivity or specificity. For the BNN, the output region of best repeatability overlapped with that of the highest case density [Fig. 3(d)], while for the LDA it did not [Fig. 3(c)].

When the aim was to operate at constant sensitivity or constant specificity (Table II) relatively wide 95% confidence intervals for the threshold value of the classifier output resulted as indicated in Fig. 3. For the analysis of $N=n=1251$ cases, the difference in the obtained median threshold values for the classifier output resulting in “high sensitivity” (90% sensitivity) and “high specificity” (90% specificity) was the largest observed, but relatively small at

TABLE IV. The impact of operating at a *constant sensitivity* or *constant specificity* (recall Table II) on corresponding classifier output value as measured by the mean and 95% confidence intervals for the classifier output values corresponding to 90% sensitivity and 90% specificity, respectively (see Table I for training/testing bootstrap protocols). Note that for $n=250$, the confidence intervals for each operating condition overlap for each classifier.

Sample size (n)	Iterations (I)	Threshold values for the classifier output [mean (95% CI)]			
		LDA		BNN	
		90% sens.	90% spec.	90% sens.	90% spec.
1251	500	0.73 ^a (0.26; 1.21)	-0.55 ^b (-0.94; -0.27)	0.13 ^a (0.07; 0.20)	0.36 ^b (0.30; 0.43)
625	1000	0.77 (0.14; 1.48)	-0.55 (-1.24; -0.16)	0.11 (0.04; 0.23)	0.37 (0.27; 0.51)
250	2500	0.78 (-0.18; 2.14)	-0.61 (1.93; 0.16)	0.08 (0.004; 0.28)	0.40 (0.21; 0.63)

^aValues used to generate (sensitivity, specificity) operating points in Fig. 4 for operating at the selected example target high sensitivity (90% sensitivity).

^bValues used to generate (sensitivity, specificity) operating points in Fig. 4 for the example high specificity operating point (90% specificity).

24% of r (LDA) and 32% of r' (BNN) (Fig. 3 and Table IV). These confidence intervals widened considerably for the smaller values of n investigated and those intervals even overlapped for the $n=250$ analysis (Table IV). It is important to note that the values of the classifier output (Table IV) lack an easily interpretable meaning in the case of the LDA classifier and are an estimate of the likelihood of malignancy relative to the cancer prevalence in the training set (which was constant in this paper) in the case of the BNN classifier.

When the aim was to operate at a constant preselected threshold for the classifier output (Table II) the resulting spread in (sensitivity, specificity) operating points was considerable. The $N=n=1251$ and $I=500$ analysis showed the least variability in resulting operating points, but also for this analysis we found that the impact of operating at the two threshold values corresponding to “high sensitivity” and to “high specificity,” respectively (see footnote of Table IV), caused large variability in corresponding sensitivity and specificity (Table V and Fig. 4). While the standard errors for the AUC values were respectable for the $N=n=1251$ and $I=500$ analysis at 0.013 and 0.010 for the LDA and BNN, respectively [as determined from the 95% confidence intervals for the AUC values (Table III) assuming normality], the

moderate overall repeatability of the classifier output had a considerable impact on the obtained (sensitivity, specificity) operating points when operating at a constant threshold value for the classifier output (corresponding to, on average, 90% sensitivity or to 90% specificity) in the decision whether or not a case was malignant. The width of the 95% confidence intervals for the obtained sensitivity and specificity ranged from 9 percentage points (pp) to 30 pp (Table V). The former was obtained by the LDA for the specificity when operating at “high specificity” and the latter was obtained by the LDA for the specificity when operating at “high sensitivity”.

IV. DISCUSSION AND CONCLUSION

In this paper, we investigated the performance of two classifiers in the task of distinguishing between benign and malignant lesions imaged with diagnostic breast ultrasound. It is important to note that it was not the intention to prove that one classifier was better than the other. Rather, the focus was on investigating the concept of repeatability of classifier output within bootstrap training and testing protocols. Both classifiers have been employed in the past for the classification task of interest and we used the unaltered CADx

TABLE V. The impact of operating at a *constant threshold* value for the classifier output on the resulting (sensitivity, specificity) operating points (recall Table II) as measured by the spread in operating points (sensitivity, specificity) resulting from the use of a fixed threshold value for the classifier output ($N=n=1251$ cases, $I=500$ bootstrap iterations) in the distinction between benign and malignant cases (see also Fig. 4).

		Resulting sensitivity and specificity [median (95% CI)]	
		LDA (%)	BNN (%)
90% sensitivity output threshold	Sens.	90 (80;98)	90 (78;97)
	Spec.	63 (44;74)	66 (55;75)
90% specificity output threshold	Sens.	53 (40;68)	62 (46;75)
	Spec.	90 (85;94)	90 (84;94)

FIG. 4. The impact of operating at a constant threshold value for the classifier output on the resulting (sensitivity, specificity) operating points (recall Table II and see also Table V). Depicted are the envelopes of the operating points for the entire data set resulting from selecting a constant value for the threshold of the classifier output (indicated in Table IV) in the task of distinguishing between benign and malignant cases for $N=n=1251$ cases and $I=500$ bootstrap iterations. Operating points are depicted with respect to the traditional ROC axes (1-specificity, sensitivity) and the envelopes were determined from 95% of the operating points closest to the target sensitivity (90%) and target specificity (90%).

schemes without any additional modification or optimization (other than classifier training) for either classifier. Had the intent been to decide which of the two classifiers was best for use within CADx for diagnostic breast ultrasound, optimization in terms of automatic lesion segmentation, feature extraction, and feature selection would have been necessary for each classifier within the training scenarios. The comparisons performed in this work used identical classifier input for both classifiers and other parameters, such as the number of hidden nodes for the BNN, were constant and based on previous experience.

Both classifiers achieved respectable overall classification performance as assessed with ROC analysis with the BNN slightly outperforming the LDA for the $N=n=1251$ case-data set ($p < 0.05$). For both classifiers the repeatability remained modest, i.e., there was relatively large variation in output values for each case, even when the full $N=1251$ data set was used in a straightforward .632+ bootstrap analysis ($N=n=1251$). Classifier repeatability can be interpreted from the perspective of case difficulty for the CADx classifier. If a classifier consistently gives the same output for the same test case irrespective of how it was trained, i.e., if the classifier output was repeatable for this test case, the case can be interpreted as “easy” (based on the given input lesion features) and if a classifier has trouble assigning a consistent output

value the case can be interpreted as “difficult” for the classifier. It was obvious from the examination of the output variability for *individual* cases that a given case posed a different degree of difficulty for each classifier. Note, however, that a high degree of repeatability does not necessarily imply a correct classification. It would be interesting to correlate these findings to the case difficulty as determined by an expert radiologist.

The LDA achieved slightly better overall repeatability (less variability) with respect to its output range. Note that in the global performance analysis, the width of the 95% confidence intervals for the AUC values was slightly larger for the LDA at 0.05 compared to 0.04 for the BNN. This seeming contradiction between the standard error in ROC analysis and the observed variations in the local repeatability analysis arises because in ROC analysis the performance assessment is based on the *relative ranking* of cases in each bootstrap iteration, while in the case-based repeatability analysis is based on the *actual value* for the classifier output for each case. The LDA was overall more consistent in assigning values to individual cases, while the BNN was slightly better at ranking them. From a CADx user perspective, this means that the output of the LDA was on average more “believable” than that of the BNN.

While the overall repeatability of the classifier output as discussed above is of interest, the dependence of the repeatability on the classifier output may be of greater importance. The repeatability profiles displayed different characteristics for the LDA and the BNN, with the LDA achieving the best repeatability in the middle of its output range and the BNN performing best for values close to the minimum of its output range. From a CADx user perspective, this means that the “believability” of the classifier output depended on the output itself. In practice, it would be desirable for a classifier to have the best repeatability (lowest variability) in the neighborhood of the output value that translates into the desired sensitivity or specificity, which depends on the clinical application. The trend in repeatability was the same for all training/testing protocols investigated, but decreasing sample size resulted in decreasing repeatability (increasing variability) even though the classifier training and testing protocols had been designed to result in the same number of samples for each of the $N=1251$ cases. It is important to note that the smaller sample sizes of $n=250$ and $n=625$ (selected randomly each bootstrap iteration from the full data set for 0.632 training and testing) appear small in this work, but this size of data set is commonly used in the literature on CADx development and application.²⁴

The impact of classifier repeatability on sensitivity and specificity was considerable, both when the aim was to operate at a constant sensitivity or specificity, and when the aim was to operate at a constant threshold for the classifier output. Perhaps most notably, the moderate overall repeatability of the classifier output had a large impact on the obtained (sensitivity, specificity) operating points when operating at a constant threshold value for the classifier output (corresponding to, on average, 90% sensitivity or to 90% specificity) in the decision whether or not a case was malignant. This is

consistent with the fact that tests based on sensitivity or specificity tend to have lower power.¹⁰ As a specific example, when a threshold value for the BNN classifier output was used aiming to operate at 90% sensitivity, the actual sensitivity for that classifier output value lied between 78% and 97% for every 19 out of 20 bootstrap iterations (i.e., 95% of the time, and outside that range 5% of the time) (Table V). From a CADx user perspective, this means that there was a large uncertainty in the obtained sensitivity and specificity when using a constant cutoff z_θ for the classifier output in the task of classifying for malignancy.

The results of the presented repeatability analysis indicate that it may be important to investigate classifier behavior in a more detailed manner beyond the summary performance measure of area under the curve (and standard error or confidence interval thereof) commonly used in ROC analysis.¹¹ We believe that these results should serve as a warning that in instances where ROC analysis suggests promising performance, further case-based classifier performance analysis may be warranted. In instances in which either the area under the ROC curve is not satisfactory or its standard error is large, case-based analysis becomes superfluous since results can be deemed unsatisfactory based on ROC analysis alone.

There were several limitations to this study. First, although the number of lesions available for the generation of classifier training sets was large ($N=1251$), it may not have been large enough to truly represent the patient population for diagnostic breast ultrasound. The data set was a realistic representation of cases seen in our diagnostic ultrasound clinic, however, and therefore expected to be a good representation of a test set. For classifier training purposes, on the other hand, the relatively low cancer prevalence (17%) may have been suboptimal. However, in related works not discussed here because of space limitations, we investigated the effect of the cancer prevalence in the training set and obtained similar results regarding repeatability (variability of the classifier output), both in trend and magnitude.

A second limitation was that we investigated CADx stand-alone performance and did not assess the interaction between radiologist and CADx. In practice, each radiologist operates with a different “internal prevalence”²⁵ and hence is expected to interpret CADx results differently as well. Whether or not the variation in classifier output of a CADx scheme has a significant impact on the performance of radiologists using such a scheme is of great interest for potential future applications of CADx in clinical practice.

We adapted the concept of repeatability²³ to investigate the properties of two classifiers trained for the task of discriminating between benign and malignant breast lesions. Our analysis emphasized the differences between the classifiers and stressed the potential impact of repeatability on the clinical interpretation of CADx output results.

ACKNOWLEDGMENTS

The project described was supported in part by Grant Numbers R01-CA89452, R21-CA113800, and P50-CA125183 from the National Institutes of Health (NIH). The

content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. M. L. G. is a stockholder in R2 Technology/Hologic and receives royalties from Hologic, GE Medical Systems, MEDIAN Technologies, Riverain Medical, Mitsubishi and Toshiba. L. L. P. is a consultant for Carestream Health Inc. and Siemens AG. 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.

^a)Electronic mail: kdrukker@uchicago.edu

¹M. L. Giger, H. P. 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**(12), 5799–5820 (2008).

²H. Li, M. L. Giger, Y. D. Yuan, W. J. Chen, K. Horsch, L. Lan, A. R. Jamieson, C. A. Sennett, and S. A. Jansen, “Evaluation of computer-aided diagnosis on a large clinical full-field digital mammographic dataset,” *Acad. Radiol.* **15**(11), 1437–1445 (2008).

³B. Sahiner, H. P. Chan, L. M. Hadjiiski, M. A. Roubidoux, C. Paramagul, J. E. Bailey, A. V. Nees, C. E. Blane, D. D. Adler, S. K. Patterson, K. A. Klein, R. W. Pinsky, and M. A. Helvie, “Multi-modality CADx: ROC study of the effect on radiologists’ accuracy in characterizing breast masses on mammograms and 3D ultrasound images,” *Acad. Radiol.* **16**(7), 810–818 (2009).

⁴M. Samulski, N. Karssemeijer, P. Lucas, and P. Groot, “Classification of mammographic masses using support vector machines and Bayesian networks—Article No. 65141J,” in *Proceedings of the Medical Imaging 2007 Conference: Computer-Aided Diagnosis*, 2007, Parts 1 and 2, Vol. 6514, p. J5141.

⁵H. P. Chan, L. Hadjiiski, C. Zhou, and B. Sahiner, “Computer-aided diagnosis of lung cancer and pulmonary embolism in computed tomography—A review,” *Acad. Radiol.* **15**(5), 535–555 (2008).

⁶Y. Arzhaeva, D. M. J. Tax, and B. van Ginneken, “Dissimilarity-based classification in the absence of local ground truth: Application to the diagnostic interpretation of chest radiographs,” *Pattern Recogn.* **42**(9), 1768–1776 (2009).

⁷L. E. Dodd, R. F. Wagner, S. G. Armato, M. F. McNitt-Gray, S. Beiden, H. P. Chan, D. Gur, G. McLennan, C. E. Metz, N. Petrick, B. Sahiner, J. Sayre, and the Lung Image Database Consortium, “Assessment methodologies and statistical issues for computer-aided diagnosis of lung nodules in computed tomography: Contemporary research topics relevant to the lung image database consortium,” *Acad. Radiol.* **11**(4), 462–475 (2004).

⁸C. E. Metz, “Basic principles of ROC analysis,” *Semin Nucl. Med.* **8**(4), 283–298 (1978).

⁹R. Wagner, C. E. Metz, and G. Campbell, “Assessment of medical imaging systems and computer aids: A tutorial review,” *Acad. Radiol.* **14**, 723–748 (2007).

¹⁰M. S. Pepe, in *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford Statistical Science Series, edited by A. C. Atkinson *et al.* (Oxford University Press, New York, 2004).

¹¹N. Cook, “Use and misuse of receiver operating characteristic curve in risk prediction,” *Circulation* **115**, 928–935 (2007).

¹²K. Drukker, N. P. Grusauskas, C. A. Sennett, and M. L. Giger, “Breast US computer-aided diagnosis workstation: Performance with a large clinical diagnostic population,” *Radiology* **248**(2), 392–397 (2008).

¹³K. 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).

¹⁴K. Drukker, M. L. Giger, and C. E. Metz, “Robustness of computerized lesion detection and classification scheme across different breast ultrasound platforms,” *Radiology* **237**, 834–840 (2005).

¹⁵N. P. Grusauskas, K. Drukker, M. L. Giger, C. A. Sennett, and L. L. Pesce, “Performance of breast ultrasound computer-aided diagnosis: Dependence on image selection,” *Acad. Radiol.* **15**(10), 1234–1245 (2008).

¹⁶N. P. Grusauskas, K. Drukker, M. L. Giger, R. F. Chang, C. A. Sennett, W. K. Moon, and L. L. Pesce, “Breast US computer-aided diagnosis system: Robustness across urban populations in South Korea and the United States,” *Radiology* **253**(3), 661–671 (2009).

- ¹⁷K. Horsch, M. L. Giger, L. A. Venta, and C. J. Vyborny, "Computerized diagnosis of breast lesions on ultrasound," *Med. Phys.* **29**(2), 157–164 (2002).
- ¹⁸K. Horsch, M. L. Giger, L. A. Venta, and C. J. Vyborny, "Automatic segmentation of breast lesions on ultrasound," *Med. Phys.* **28**(8), 1652–1659 (2001).
- ¹⁹R. A. Fisher, "The precision of discriminant functions," *Ann. Eugen.* **10**, 422–429 (1940).
- ²⁰I. Poli and R. D. Jones, "A neural-net model for prediction," *J. Am. Stat. Assoc.* **89**(425), 117–121 (1994).
- ²¹B. Efron and R. Tibshirani, "Improvements on cross-validation: The .632+ bootstrap method," *J. Am. Stat. Assoc.* **92**(438), 548–560 (1997).
- ²²B. Sahiner, H. P. Chan, and L. Hadjiisk, "Classifier performance prediction for computer-aided diagnosis using a limited data set," *Med. Phys.* **35**, 1559–1570 (2008).
- ²³J. M. Bland and D. G. Altman, "Statistical-methods for assessing agreement between 2 methods of clinical measurement," *Lancet* **1**(8476), 307–310 (1986).
- ²⁴J. Shiraishi, L. L. Pesce, C. E. Metz, and K. Doi, "Experimental design and data analysis in receiver operating characteristic studies: Lessons learned from reports in radiology from 1997 to 2006," *Radiology* **253**(3), 822–830 (2009).
- ²⁵K. Horsch, M. L. Giger, and C. E. Metz, "Potential effect of different radiologist reporting methods on studies showing benefit of CAD," *Acad. Radiol.* **15**(2), 139–152 (2008).