Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2013, Article ID 651091, 6 pages
http://dx.doi.org/10.1155/2013/651091

Research Article

Statistical Evaluation of a Fully Automated Mammographic Breast Density Algorithm

Mohamed Abdolell,[1,2,3] Kaitlyn Tsuruda,[2] Gerry Schaller,[1,2] and Judy Caines[1,2,4]

[1] *Department of Diagnostic Radiology, Dalhousie University, Halifax, NS, Canada B3H 2Y9*
[2] *Department of Diagnostic Imaging, Capital District Health Authority, Halifax, NS, Canada B3H 2Y9*
[3] *Division of Medical Education/Informatics, Dalhousie University, Halifax, NS, Canada B3H 2Y9*
[4] *Nova Scotia Breast Screening Program, Halifax, NS, Canada B3H 2Y9*

Correspondence should be addressed to Mohamed Abdolell; mo@dal.ca

Received 2 October 2012; Revised 5 April 2013; Accepted 9 April 2013

Academic Editor: Giner Alor-Hernández

Copyright © 2013 Mohamed Abdolell et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Visual assessments of mammographic breast density by radiologists are used in clinical practice; however, these assessments have shown weaker associations with breast cancer risk than area-based, quantitative methods. The purpose of this study is to present a statistical evaluation of a fully automated, area-based mammographic density measurement algorithm. Five radiologists estimated density in 5% increments for 138 "For Presentation" single MLO views; the median of the radiologists' estimates was used as the reference standard. Agreement amongst radiologists was excellent, ICC = 0.884, 95% CI (0.854, 0.910). Similarly, the agreement between the algorithm and the reference standard was excellent, ICC = 0.862, falling within the 95% CI of the radiologists' estimates. The Bland-Altman plot showed that the reference standard was slightly positively biased (+1.86%) compared to the algorithm-generated densities. A scatter plot showed that the algorithm moderately overestimated low densities and underestimated high densities. A box plot showed that 95% of the algorithm-generated assessments fell within one BI-RADS category of the reference standard. This study demonstrates the effective use of several statistical techniques that collectively produce a comprehensive evaluation of the algorithm and its potential to provide mammographic density measures that can be used to inform clinical practice.

1. Introduction

Breast density refers to fibroglandular tissue in the breast and is one of the top major risk factors for breast cancer. Women with extremely dense breasts (75% or greater mammographic density) have a four- to sixfold increase in the risk of developing breast cancer compared to those with fatty breasts (less than 25% density) [1–3].

Traditionally, visual assessment by radiologists has been used to characterize and quantify mammographic density (and a woman's risk for breast cancer) using Wolfe Grades, Tabar Patterns, Boyd Scales, or the American College of Radiologists' (ACR) Breast Imaging Reporting and Data System (BI-RADS) density lexicon [4–7]. Despite good reproducibility, methods used to characterize mammographic density have shown weaker associations with breast cancer risk compared to methods quantifying mammographic density [2, 3, 8] and suffer from inter- and intraobserver variability.

The ACR has stated that radiologists' visual assessments of percent breast density using the BI-RADS lexicon are "not reliably reproducible" [9]. This fundamental lack of reproducibility has led to the development of various semi- and fully automated algorithms to quantify percent breast density as a means to overcome inter- and intraobserver variability. It is therefore important to apply rigorous statistical methods to evaluate the performance of these algorithms.

1.1. State of the Art. Area-based methods used to quantify mammographic density have produced reliable and standardized mammographic density measurements on a continuous scale. The de facto standard of such methods is the Cumulus software [10, 11]. Using Cumulus, a digitized film-screen mammogram is displayed and a trained user selects a threshold value to separate the breast area from the background (i.e., the region of interest). A second threshold is

then selected to identify regions of dense breast tissue, and the percent breast density is calculated as the area of dense tissue divided by the area of the region of interest. Despite being a proven predictor of breast cancer risk, the semiautomated nature of Cumulus' breast density assessments is susceptible to inter- and intraobserver variability and could be improved by a fully automated method. Additionally, this software is intended for use with digitized film-screen mammograms. As 90% of certified mammography units in the USA are now full-field digital [12], a software for use with full-field digital mammograms (FFDMs) is needed.

Volume-based methods theoretically yield accurate estimates of mammographic density and so it is simply assumed that volume-based density estimates are associated with breast cancer risk, as has been demonstrated to be the case for area-based estimates (both visually and algorithmically assessed) [13, 14]. Volumetric methods use "For Processing" FFDMs and DICOM header information to calculate density. Yet, volume-based estimates have not been shown to demonstrate a similarly strong association with breast cancer risk [11, 15, 16]. Additionally, the underlying distribution of mammographic density estimates from volumetric methods is significantly more left-skewed than that of area-based methods (typical range 0–40% versus 0–100%) [17], making them difficult to interpret by radiologists, who are not simply able to visualize mammographic density as a volumetric construct [11, 15].

The assessment of the agreement between percent breast density algorithms and an expert radiologist should necessarily quantify the consistency or reproducibility of measurements made by these two "raters" on the same set of digital mammograms. The intraclass correlation coefficient (ICC) provides such a measure of agreement [18]. The Bland-Altman plot is another way to assess agreement between raters. Scatter and box plots can also yield insights into the level of agreement between raters. Yet, much of the literature validating emerging density measurement algorithms relies on the use of the Pearson correlation coefficient, ρ, which is a measure of the linear dependence between two raters and can be quite high despite the agreement being poor [18, 19]. Overall percent agreement is another statistic that is used to assess agreement but is also flawed as it does not factor in any inherent inter- and intrarater variability [19]. Reporting of a single numerical measure of agreement alone is one-dimensional and does not present a comprehensive perspective on algorithm performance.

This paper presents several statistical methods that collectively provide a more comprehensive evaluation of the performance of a fully automated area-based image analysis algorithm that generates percent breast density measures from FFDMs.

2. Materials and Methods

138 "For Presentation" FFDMs collected from the Capital District Health Authority in Nova Scotia were retrospectively analyzed. Images were acquired on Siemens full-field digital mammography machines and automatically postprocessed by the manufacture's proprietary software at the time of acquisition. This early stage work has focused on the mediolateral oblique views and excluded craniocaudal views as it has been shown in the literature that mammographic density estimates from only one view are sufficient to indicate breast cancer risk [20]. In addition, the ACRs' National Mammography Database breast density element definition stipulates that "if left and right breasts differ, use the higher density" [21].

2.1. Percent Density Analysis. Percent mammographic density was measured by a fully automated research-based algorithm that uses "For Presentation" FFDMs to calculate an area-based measure of density as a percentage on a continuous scale (Figure 1, Panels 1(a) through 1(d)). Using view position and image laterality information from the DICOM header (elements (0018, 5101) and (0020, 0062), resp.) the software creates and applies a mask to identify the breast envelope (region of interest) by removing the pectoral muscle, subcutaneous fat, and overlay text (Panel 1(b)). A variation of the MaxEntropy and Moments thresholding methods is applied to determine a threshold for dense tissue in the breast [22, 23]. The area of the dense tissue (i.e., the number of pixels of dense tissue) is then calculated (Panel 1(c)), as is the area of the region of interest (i.e., the number of pixels in the breast area, Panel 1(d)), and the final density estimate is calculated as the ratio of dense tissue area to the region of interest. In this manner, the software uniquely generates a reproducible, fully automated, area-based estimate of mammographic density using "For Presentation" FFDM images.

To evaluate the agreement between the algorithm and an expert mammographer, percent mammographic density was visually assessed by five radiologists in 5% density increments (0%, 5%, . . . , 95%, 100%) using five megapixel Barco Screens supported by the Syngo MammoReport Software (VB24D, Siemens AS). Visual assessments were performed by two senior mammographers, one junior mammographer, one senior resident, and one fellow. This 21-point scale was used as a proxy for a continuous measure.

2.2. Statistical Analysis. To quantify the reliability of estimates performed by the radiologists' visual assessments, Intraclass Correlation Coefficients (ICCs) were used to measure interobserver agreement. Although the interpretation of ICCs can vary depending on the context, the ICC is equivalent to a quadratically weighted Kappa, and a widely referenced scale for interpretation of Kappa values can be used as a general guide [24, 25]. Specifically, ICC values of 0.00–0.20, 0.21–0.40, 0.41–0.60, 0.61–0.80, and 0.81–1.00 were used to indicate poor, fair, moderate, substantial, and excellent to perfect agreement, respectively.

It has repeatedly been shown that radiologists' visual assessments of mammographic density are associated with breast cancer risk [1, 3, 4, 10, 26]. As such, the median of the visual assessments performed by the five participating radiologists was considered to be the reference standard for this analysis. The algorithm was considered promising in informing clinical practice if the agreement between the

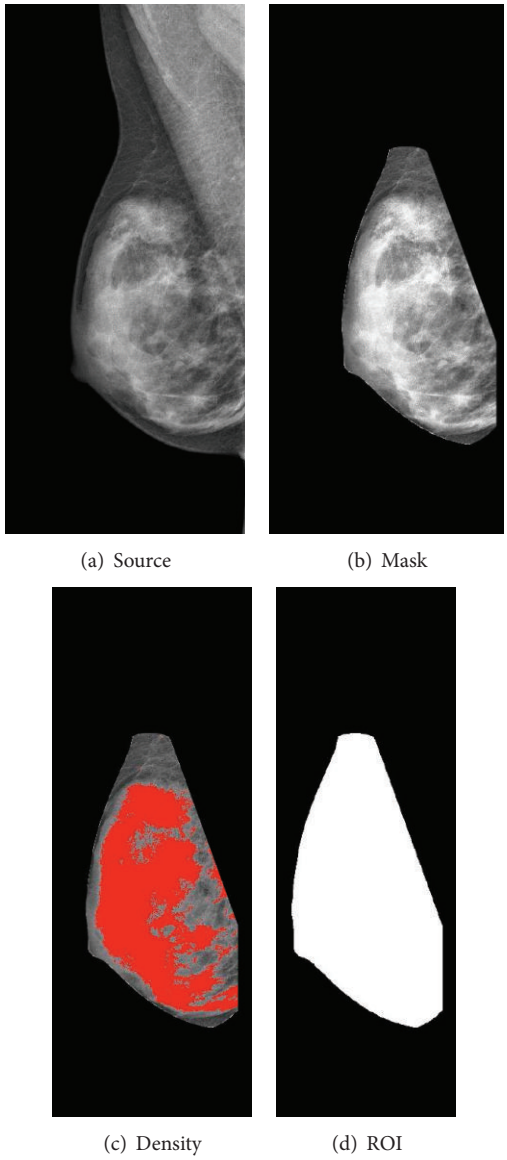

Figure 1: A sequence of processed images generated at various steps of the algorithm for estimating area-based mammographic density: (a) a "For Presentation" mammogram from our sample; (b) the image after a mask has been applied to identify the breast envelope; (c) the area of dense tissue (red pixels); and (d) the region of interest as a binary map of the breast envelope. The algorithm calculates percent breast density as the number of red pixels in Panel (c) divided by the number of white pixels in Panel (d).

algorithm and the reference standard fell within the 95% CI of the ICC of the radiologists.

The ICC was used to quantify the level of agreement between the algorithm and the reference standard, and a scatterplot was used to demonstrate the relationship between the two. A Bland-Altman difference plot was used to analyze the agreement between the algorithm and the reference standard and to quantify the amount and direction of bias as well as the upper and lower limits of agreement (bias $\pm 1.96\sigma$ of the difference) [27]. Lastly, a box-and-whisker plot was used to visualize the results in terms of the BI-RADS density lexicon (0–24%, 25–49%, 50–74%, and 75–100%) [7].

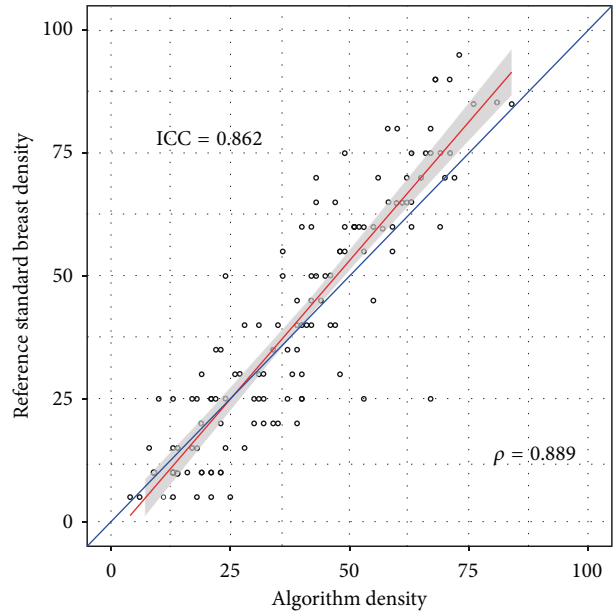

Figure 2: Scatter plot showing the relationship between the mammographic density estimates produced by the algorithm (x-axis) and the reference standard (y-axis). The blue line indicates perfect agreement between the algorithm and the reference standard, in which case all points would fall exactly on the line of agreement. The red line is the line of the best fit determined by linear least squares regression analysis and shows that the algorithm tends to slightly overestimate density compared to the reference standard for lower densities and slightly underestimate density compared to the reference standard for higher densities.

3. Results

Five radiologists visually assessed 138 images to estimate mammographic density, and the algorithm was applied to those same 138 images to generate a fully automated density assessment for each of the images.

The radiologists' visual assessments were in excellent agreement with an ICC = 0.884, 95% CI (0.854, 0.910). The algorithm demonstrated excellent agreement with the reference standard with an ICC = 0.862, which fell within the 95% CI of the agreement between the radiologists' visual assessments. The algorithm is validated well on an external set of 261 mammograms, ICC = 0.841.

The Pearson correlation coefficient between the algorithm and the reference standard assessments was $\rho = 0.889$.

The algorithm slightly overestimated low densities and underestimated high densities compared to the reference standard (Figure 2). Overall, there was a small, positive bias in the reference standard assessments compared to the algorithm assessments, as measured by the mean difference between the reference standard and the algorithm assessments (bias = 1.86%) (Figure 3). Additionally, the upper and lower agreement levels were both less than 25%, and thus

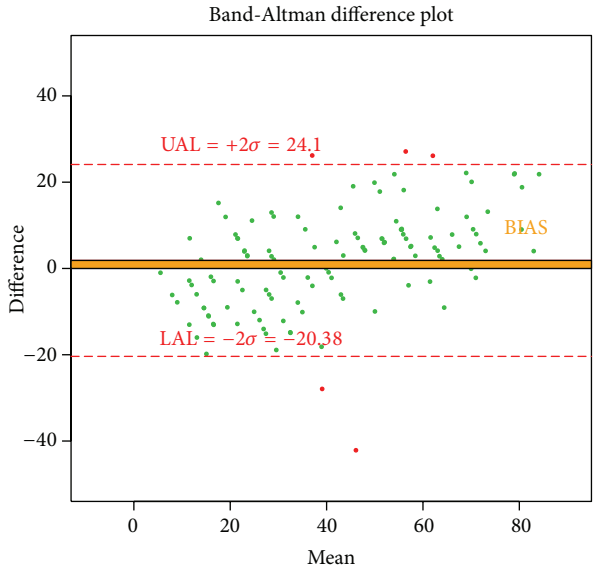

Figure 3: Bland-Altman difference plot showing agreement between the algorithm and the reference standard measures of mammographic density. The difference refers to the reference standard minus the algorithm assessment. The absolute values of the upper and lower agreement limits are <25%, which is the span of a single category in the 4-level BI-RADS density classification scheme. A bias of +1.86%, as indicated by the orange band above the horizontal zero difference line, shows that the reference standard density is on average only slightly higher than the density generated by the algorithm.

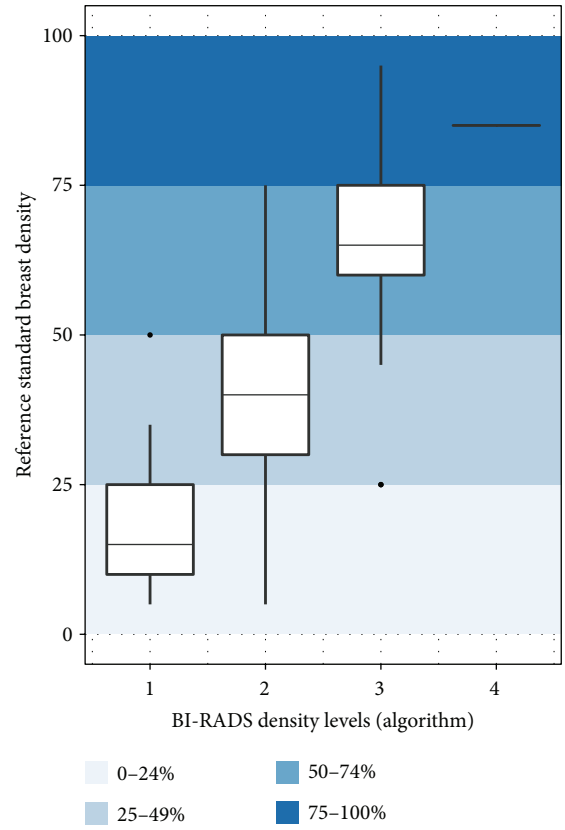

Figure 4: Box-and-whisker plot displaying the distribution of reference standard mammographic density assessments falling into the algorithm-derived classifications designated by the standard 4-level BI-RADS density lexicon. Ideally, each of the boxes and their whiskers should be entirely contained in their respective BI-RADS levels. The reference standard mammographic density assessments in the lowest and the highest BI-RADS levels are well classified, while the middle two levels overlap in both directions into adjacent BI-RADS levels.

approximately 95% of the data classified by the algorithm was within one BI-RADS category of the reference standard classification (Figure 3).

When the algorithm and reference standard estimates were classified using the BI-RADS density lexicon, the box-and-whisker plots showed good agreement within categories (Figure 4). Each box was contained in the accordant colour bar, and, as expected from the Bland-Altman difference plot, the tails on the graphs did not exceed the adjacent BI-RADS categories.

4. Discussion

The algorithm demonstrates excellent agreement with radiologists' visual assessments of mammographic density. Critically, the observed magnitude of this agreement falls within the 95% CI of agreement observed between radiologists. This algorithm is unique in that it generates fully automated mammographic density measurements that can be straightforwardly compared with visually determined radiologists' estimates, which are well accepted as being associated with breast cancer risk.

The sole use of the Pearson correlation coefficient (ρ) provides a one-dimensional and overinflated impression of the level of agreement.

The statistical evaluation presented in this paper used ICCs and Bland-Altman, scatter, and box plots to quantify agreement and bias in breast density assessment between a fully automated algorithm and radiologists' assessments. This multifaceted methodology can be employed to comprehensively evaluate the performance of any breast density measurement algorithm and provides an alternative to the often reported Pearson correlation coefficient and percent agreement statistics which do not consider random chance agreement and cannot quantify bias between different raters.

As breast density legislation gains momentum in the USA and mammography providers are required to disclose breast density in the lay report, there will be an increasing need for automated solutions that provide reliable and accurate measurements of breast density. A woman's breast density will be used to determine her optimal followup, and thus the performance of these algorithms must be evaluated using robust statistical methodologies.

5. Conclusion

Further work is needed to extend the applicability of the breast density algorithm to FFDMs from other manufacturers

as each manufacturer has their own proprietary image processing algorithms that generate "For Presentation" images. Additionally, as radiologists use both mediolateral and craniocaudal views to assess breast density in a clinical setting, the present algorithm must also be extended to accommodate the analysis of craniocaudal views.

The present algorithm is an effective research tool and shows promise in its ability to provide automated mammographic density measurements that can be used to inform clinical practice. The Pearson correlation coefficient (ρ) provides an inadequate, inflated, and overoptimistic measure of the level of agreement. The statistical methods employed provide a comprehensive evaluation of the level of agreement between the algorithm and the reference standard and confirm that the algorithm has an excellent level of agreement with the reference standard. Agreement between raters can only be adequately assessed using multiple statistical methods.

Ethical Approval

This study was approved by the Capital District Health Authority Research Ethics Board: CDHA-RS/2007-365.

Conflict of Interests

Mohamed Abdolell is the founder of Densitas Inc. and is a shareholder in the company. Kaitlyn Tsuruda is an employee at the company.

Acknowledgments

The authors would like to thank Dr. Melanie McQuaid, Dr. Ieva Klavina, and Dr. Chris Lightfoot for providing visually assessed mammographic density measurements used in this study. This research was supported by the Department of Diagnostic Radiology, Dalhousie University and Capital District Health Authority, and the Canadian Breast Cancer Foundation (Atlantic Region). The study sponsors had no involvement in the study design; in the collection, analysis, and interpretation of data; in the writing of the paper; nor in the decision to submit the paper for publication.

References

[1] N. F. Boyd, H. Guo, L. J. Martin et al., "Mammographic density and the risk and detection of breast cancer," *New England Journal of Medicine*, vol. 356, no. 3, pp. 227–236, 2007.

[2] J. Brisson, C. Diorio, and B. Mâsse, "Wolfe's parenchymal pattern and percentage of the breast with mammographic densities: redundant or complementary classifications?" *Cancer Epidemiology Biomarkers and Prevention*, vol. 12, no. 8, pp. 728–732, 2003.

[3] V. A. McCormack and I. dos Santos Silva, "Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis," *Cancer Epidemiology Biomarkers and Prevention*, vol. 15, no. 6, pp. 1159–1169, 2006.

[4] J. N. Wolfe, "Risk for breast cancer development determined by mammographic parenchymal pattern," *Cancer*, vol. 37, no. 5, pp. 2486–2492, 1976.

[5] I. T. Gram, E. Funkhouser, and L. Tabár, "The Tabar classification of mammographic parenchymal patterns," *European Journal of Radiology*, vol. 24, no. 2, pp. 131–136, 1997.

[6] N. F. Boyd, H. M. Jensen, G. Cooke, and H. L. Han, "Relationship between mammographic and histological risk factors for breast cancer," *Journal of the National Cancer Institute*, vol. 84, no. 15, pp. 1170–1179, 1992.

[7] *Breast Imaging Reporting and Data Systems (BIRADS)*, American College of Radiology, Reston, Va, USA, 1993.

[8] M. Garrido-Estepa, F. Ruiz-Perales, J. Miranda et al., "Evaluation of mammographic density patterns: reproducibility and concordance among scales," *BMC cancer*, vol. 10, supplement 485, 2010.

[9] American College of Radiology, "ACR Statement on Reporting Breast Density in Mammography Reports and Patient Summaries," 2012, http://www.acr.org/About-Us/Media-Center/Position-Statements.

[10] N. F. Boyd, J. W. Byng, R. A. Jong et al., "Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian National Breast Screening Study," *Journal of the National Cancer Institute*, vol. 87, no. 9, pp. 670–675, 1995.

[11] N. Boyd, L. Martin, A. Gunasekara et al., "Mammographic density and breast cancer risk: evaluation of a novel method of measuring breast tissue volumes," *Cancer Epidemiology Biomarkers and Prevention*, vol. 18, no. 6, pp. 1754–1762, 2009.

[12] FDA, "MQSA National Statistics," 2013, http://www.fda.gov/Radiation-EmittingProducts/MammographyQualityStandards-ActandProgram/FacilityScorecard/ucm113858.htm.

[13] O. Alonzo-Proulx, N. Packard, J. M. Boone et al., "Validation of a method for measuring the volumetric breast density from digital mammograms," *Physics in Medicine and Biology*, vol. 55, no. 11, pp. 3027–3044, 2010.

[14] M. J. Yaffe, "Mammographic density. Measurement of mammographic density," *Breast Cancer Research*, vol. 10, no. 3, article 209, 2008.

[15] D. Kontos, P. R. Bakic, R. J. Acciavatti, E. F. Conant, and A. D. A. Maidment, "A comparative study of volumetric and area-based breast density estimation in digital mammography: results from a screening population," in *Digital Mammography*, vol. 6136 of *Lecture Notes in Computer Science*, pp. 378–385, 2010.

[16] J. Ding, R. Warren, I. Warsi et al., "Evaluating the effectiveness of using standard mammogram form to predict breast cancer risk: case-control study," *Cancer Epidemiology Biomarkers and Prevention*, vol. 17, no. 5, pp. 1074–1081, 2008.

[17] R. Highnam, S. Brady, M. Yaffe, N. Karssemeijer, and J. Harvey, "Robust breast composition measurement-volpara TM," in *Proceedings of the 10th International Conference on Digital Mammography (IWDM '10)*, vol. 6136 of *Lecture Notes in Computer Science*, pp. 342–349, 2010.

[18] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet*, vol. 1, no. 8476, pp. 307–310, 1986.

[19] R. J. Hunt, "Percent agreement, Pearson's correlation, and kappa as measures of inter-examiner reliability," *Journal of Dental Research*, vol. 65, no. 2, pp. 128–130, 1986.

[20] J. Stone, J. Ding, R. M. L. Warren, and S. W. Duffy, "Predicting breast cancer risk using mammographic density measurements from both mammogram sides and views," *Breast Cancer Research and Treatment*, vol. 124, no. 2, pp. 551–554, 2010.

[21] American College of Radiology, "National mammography database data elements," version 2.0, 2009, https://nrdr.acr.org/Portal/HELP/NMD/nmd_data_elements.pdf.

[22] J. N. Kapur, P. K. Sahoo, and A. K. C. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer Vision, Graphics, & Image Processing*, vol. 29, no. 3, pp. 273–285, 1985.

[23] W. H. Tsai, "Moment-preserving thresholding: a new approach," *Computer Vision, Graphics, & Image Processing*, vol. 29, no. 3, pp. 377–393, 1985.

[24] J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educational and Psychological Measurement*, vol. 33, pp. 613–619, 1973.

[25] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[26] J. A. Harvey and V. E. Bovbjerg, "Quantitative assessment of mammographic breast density: relationship with breast cancer risk," *Radiology*, vol. 230, no. 1, pp. 29–41, 2004.

[27] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet*, vol. 1, no. 8476, pp. 307–310, 1986.