CONFIDENCE ANALYSIS FOR BREAST MASS IMAGE CLASSIFICATION

Andrik Rampun[⋆]*, Hui Wang*[†]*, Bryan Scotney*[⋆]*, Philip Morrow*[⋆] *and Reyer Zwiggelaar*[‡]

[⋆]School of Computing, Ulster University, Coleraine, Northern Ireland, UK
[†]School of Computing, Ulster University, Jordanstown, Newtownabbey, Northern Ireland, UK
[‡]Department of Computer Science, Aberystwyth University, UK

ABSTRACT

Computer-aided diagnosis (CAD) has great potential in providing real benefits to doctors and patients. Recent studies have, however, found lack of trust in CAD by radiologists in clinical diagnostic decision making. One of the main reasons is the lack of an appropriate confidence measure. This paper presents the first-ever study of classification confidence in the context of breast mass classification. We evaluated 11 state-of-the-art classification algorithms on breast mass image data using their confidence of classification metric, in addition to other standard evaluation metrics including accuracy and area under the curve (ROC). Experimental results show that although most classifiers produced very similar results with less than 2% difference in terms of accuracy and ROC, their performances are significantly different in terms of confidence levels. We suggest that the confidence measure should be used in conjunction with the existing performance metrics such as accuracy and ROC.

Index Terms— Confidence Level, Breast Mass Classification, Computer Aided Diagnosis, Machine Learning

1. INTRODUCTION

CAD systems have been shown to have great potential in providing support to doctors such as performing routine work, quantifying tissue changes that correspond with tumour grade in an accurate and reproducible manner and extracting biomarkers which are otherwise difficult for human doctors [1, 2, 3, 4, 5, 6, 7]. This will ultimately benefit patients and health systems. However, many studies have reported that the interactions between radiologists and CAD systems are not optimal and radiologists often lack trust in the system [8, 9, 10, 11], even though some CAD systems have high accuracies and high area under the curve (AUC) values. For example, in some clinical cases if the radiologist is not confident about his/her assessment, he/she would turn to the CAD system for support as a second reader. If the CAD provides recommendation with a confidence of 0.60, the radiologist would deem it useless. If the CAD recommends with a confidence of 0.90, the radiologist would deem it useful [8]. The question is, is the confidence value a true measure of the recommendation being true?

In the study of medical CAD systems, a large majority of the studies tend to report the performance of their methods based on the accuracy, sensitivity, specificity, precision, recall or AUC metrics [6]. Unfortunately, these metrics do not measure the degree of confidence in individual recommendations (i.e. classifications, predictions). For example, a CAD method may produce an AUC value of 0.97 but the majority of the cases are correctly classified with a confidence value between 0.50 to 0.60, which could reduce the acceptability of the classifications by radiologists. The confidence outputs of a CAD system (or a machine learning system) are able to assist radiologists in improving their diagnostic decisions [8, 11, 12, 13]. Philpotts [12] studied the limitations of a CAD system in mammographic interpretation and found that accuracy alone is insufficient and could erode radiologists confidence in making a diagnostic decision. In a study of learning health care systems, Cahan and Cimino [13] suggested that CAD support systems (DSSs) should consider a reliable confidence measure as one of the key elements. Furthermore, the study of Jorritsma *et al.* [8] also suggested that using a confidence rating is an important factor in building trust between radiologists and CAD systems.

In this paper, we investigate the reliability of the confidence measures of 11 machine learning algorithms, testing them on the challenging breast mass classification problem. To the best of our knowledge, this is the first ever study on the reliability performance of machine learning algorithms in medical CAD systems. This study is motivated by the issues recently raised by Jorritsma *et al.* [8]. Based on their study, the authors concluded that the interaction between clinicians and the computer system is not optimal during the diagnostic decision-making process, thus preventing the use of CAD systems in real clinical environments from reaching their full potential.

2. MATERIALS AND EXPERIMENTS

The dataset used in this study is taken from the Curated Breast Imaging Subset of Digital Database for Screening Mammog-

raphy (CBIS-DDSM) [14]. In total it contains 1593 masses (829 benign and 764 malignant) from 838 patients. Each case is a biopsy proven 'benign' versus 'malignant' annotation by expert radiologists.

2.1. Feature Descriptions

We used the following clinical features which are already included in the CBIS-DDSM: (1) Breast density (F_d) with values 1 to 4. Breast density describes the amount of dense tissues within the breast. The higher the amount of dense tissues the higher the probability of the patient having cancer, (2) Mass shape (F_s) with the following criterion: N/A (F_s^1), round (F_s^2), oval (F_s^3), lobulated (F_s^4), lymph node (F_s^5), focal asymmetric density (F_s^6), and asymmetric breast tissue (F_s^7), architectural distortion (F_s^8), irregular (F_s^9). Mass shape describes the whole architecture or structure of the mass. The more irregular the shape the more likely it is of being malignant, (3) Mass margin (F_m) with the following criterion: N/A (F_m^1), circumscribed (F_m^2), microlobulated (F_m^3), obscured (F_m^4), ill defined (F_m^5), and spiculated (F_m^7). Mass margin characterises the structures of the mass boundary. A well defined mass margin is more likely to be benign, whereas a mass with spiculated boundary is more likely to be malignant, (4) Subtlety (F_t) with values 1 to 5. Subtlety indicates the changes of the mass from when it was first diagnosed to the subsequent follow up. The more subtle the mass the higher the probability of being malignant, and (5) Assessment (F_a) with values 0 to 5. Each mass is classified using the six BI-RADS scales: (a) BI-RADS 0 (incomplete), (b) BI-RADS 1 (negative), (c) BI-RADS 2 (benign findings), (d) BI-RADS 3 (probably benign), (e) BI-RADS 4 (suspicious abnormality), and (f) BI-RADS 5 (highly suspicious of malignancy).

2.2. Feature Representation

To represent the clinical information as a feature vector (F), each feature is concatenated as follows

$$F = \{F_d, F_s, F_m, F_t, F_a\} \tag{1}$$

where $F_d \in \{1,2,3,4\}$, $F_t \in \{1,2,3,4,5\}$ and $F_a \in \{0,1,2,3,4,5\}$. Since the values of F_s and F_m can be many due to possible combination of different shapes and margins, we present both features as a set of 'binary representations'. This can be done by first sorting shape characteristics from most likely to be benign to most likely to be malignant. Therefore, if the above sets of shape and margin are, respectively,

$$F_s = \{F_s^1, F_s^2, F_s^3, F_s^4, F_s^5, F_s^6, F_s^7, F_s^8, F_s^9\} \tag{2}$$

and

$$F_m = \{F_m^1, F_m^2, F_m^3, F_m^4, F_m^5, F_m^6\} \tag{3}$$

then, for example, if the shape of the mass is a combination of oval (F_s^3), lymph node (F_s^5) and architectural distortion (F_s^8), then resulting F_s is represented as [0,0,1,0,1,0,0,1,0]. This is similar for the mass margin. We tested several feature selection techniques in Waikato Environment for Knowledge Analysis (WEKA) [15] and found that the CfsSubsetEval [16] attribute evaluator and the GreedyStepwise search method produced the best results (hence, employed these techniques throughout this study).

2.3. Machine Learning Algorithms

We employed the following classifiers: Random Forest (RF), Multilayer Perceptron (MLP), Logistic Regression (LR), Naïve Bayes (NB), Bayesian Network (BNet), k-Nearest Neighbours (k-NN), C4.5 (C4.5), Alternate Decision Tree (ADTree), Logistic Model Trees (LMT), AdaBoostM1 (AdaBoost) and Support Vector Machine (SVM). The CVParameterSelection and GridSearch techniques were employed in WEKA [15] for parameter selection.

2.4. Experimental Settings

A patient based stratified ten-runs 10-fold cross-validation (10-FCV) scheme was employed to ensure no masses from the same patient were used in the training and testing phases.

2.5. Evaluation Metrics

We evaluate the performance of each classifier using accuracy (A) and area under the curve (AUC). The accuracy measures the number of masses classified correctly over the total number of masses, whereas the AUC indicates the trade-off between the true positive classification rate against the false positive rate. Subsequently, we evaluate the confidence (C) performance of each classifier. For this purpose, we used the probability outputs ($P \in (0,1]$) as a confidence indication for each instance/case being 'benign' or 'malignant'. We categorise the values into the following classes: (a) Confidence 1 (C_1): $0.50 \leq P \leq 0.59$, (b) Confidence 2 (C_2): $0.60 \leq P \leq 0.69$, (c) Confidence 3 (C_3): $0.70 \leq P \leq 0.79$, (d) Confidence 4 (C_4): $0.80 \leq P \leq 0.89$, and (e) Confidence 5 (C_5): $0.90 \leq P \leq 1.0$. Hence, we investigate P for each correctly classified case (TP and TN) as to the level of confidence with which it is classified as being correct. Note that the P values are the probability outputs for the classifier and describing details is beyond the scope of this paper. We refer the reader to WEKA documentation [15] for computing P values for each classifier.

3. EXPERIMENTAL RESULTS

3.1. Quantitative Results

Table 1 summarises the quantitative results for 11 classifiers employed in this study in terms of A and AUC.

Table 1. Quantitative results based on A and AUC (ordered by AUC).

Classifiers	Accuracy (%)	AUC (%)
ADTree	81.35±1.64	87.96±1.30
RF	82.51 ±1.99	87.89±1.91
AdaBoost	80.57±1.25	87.86±1.27
BN	81.25±1.24	87.60±2.20
LR	80.94±1.53	87.50±0.85
MLP	80.87±1.83	87.38±1.48
LMT	81.79±1.95	87.21±1.78
NB	80.62±1.14	87.00±0.71
k-NN	81.47±1.99	86.63±1.77
SVM	81.12±3.05	86.30±3.14
C4.5	81.76±1.98	81.46±3.60

It can be observed that ensemble-based classifiers (e.g ADTree, RF and AdaBoost) outperformed the other classifiers in terms of AUC. In terms of accuracy, the RF classifier produced the highest $A = 82.51\%$ followed by the LMT classifier with $A = 81.79\%$, which is only 0.3% higher than the third best classifier (C4.5). Overall, most classifiers produced very similar results when evaluated using common performance metrics such as A and AUC.

3.2. Results of P Distribution

The distribution is clustered into five categories C_1, C_2, C_3, C_4 and C_5. For each category we calculate the percentage probability true positive (TP) and true negative (TN) cases falling into these categories. Table 2 shows the results of P distribution for all employed classifiers.

Generally, it can be observed that the percentages of cases falling into these categories vary across classifiers. For example, only 0.41% of the TN cases are in $C_1^{TN}(0.50 \leq P \leq 0.59)$ for the BNet classifer whereas the corresponding value is over 10% for the ADTree classifer. Similarly, over 36% of the TN cases fall into $C_5^{TN}(0.90 \leq P \leq 1.0)$ for the BNet classifier and only 3.49% for the ADTree classifier. This indicates that although these classifiers have small differences in terms of A and AUC, in terms of confidence measure these results show they are significantly different. In terms of P distribution for TN cases, the BNet and NB classifiers are the top two with $C_5^{TN} = 36.02\%$ and 29.60%, respectively. Although the RF classifier achieved the highest accuracy, only 19.12% of the TN cases have probability outputs ≥ 0.90. The RF and LMT classifier are the top two in terms of accuracy performance (82.51% and 81.79%, respectively) but on average more than 17% of the TP and TN cases have $P \leq 0.69$.

3.3. Confidence Measure

Figure 1 shows the total percentage of cases for $\sum C_5^{TP+TN}$, $\sum(C_5^{TP+TN} + C_4^{TP+TN})$ and $\sum(C_5^{TP+TN} + C_4^{TP+TN} + C_3^{TP+TN})$ across different classifiers. It can be observed that on average more than 60% of the correctly classified cases have probability outputs $P \geq 0.90$ when using the BNet classifier, followed by the NB classifier with approximately 55%. This indicates that these two classifiers are very reliable in terms of degree of certainty if $P \geq 0.90$.

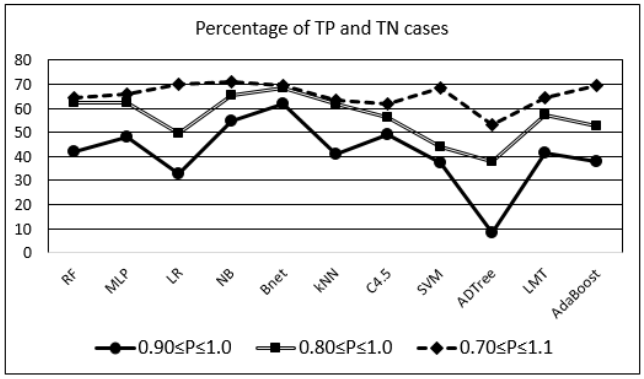

Fig. 1. Percentage of TP and TN cases for $\sum P \geq 0.9$ (top graph), $\sum P \geq 0.8$ (middle graph) and $\sum P \geq 0.7$ (bottom graph).

Furthermore, with $P \geq 0.80$, the BNet once again performed the best with on average 68.48% of the TP and TN cases classified above this threshold value, followed by the NB classifier (65.23%). In a case of $P \geq 0.70$ in Figure 1 shows the average percentage of TP and TN cases classified with $P \geq 0.70$. It can be observed that the NB classifier ranked first, producing 71.18%, closely followed by the LR (71.04%) and BNet classifiers (69.41%). The remaining classifiers achieved over 60% except the ADtree with 53.33%, which suggests that a large number of the TP and TN cases were classified with $0.50 \leq P \leq 0.69$.

4. DISCUSSION AND CONCLUSIONS

The A and AUC metrics do not provide a complete representation of the 'usefulness' of medical CAD systems. In this study, we have shown that although the ADTree and RF classifiers produced the best AUC and A values, respectively, the majority of TP and TN cases were classified with $P \leq 0.70$. Although most of the classifiers employed in this study produced similar results in terms of A and AUC, their performances are significantly different when using the probability outputs to measure their reliability. This suggests that the selection of classifiers to be employed in the CAD systems is

Table 2. The average percentage of P distribution for all the classifiers employed in this study ($\% \pm \sigma$).

	$0.9 \leq P \leq 1.0$	$0.80 \leq P \leq 0.89$	$0.70 \leq P \leq 0.79$	$0.60 \leq P \leq 0.69$	$0.50 \leq P \leq 0.59$
Random Forest (RF)					
TP	22.91±4.90	4.71±3.02	0.71±1.32	6.71±3.20	0.71±0.91
TN	19.12±6.32	15.50±5.40	1.42±3.50	0.62±1.10	10.11±3.42
Multilayer Perceptron (MLP)					
TP	25.52±5.20	2.41±3.13	1.82±2.50	4.63±3.43	2.60±3.73
TN	22.31±6.60	12.21±6.61	1.70±3.54	2.52±4.40	5.15±4.72
Logistic Regression (LR)					
TP	11.91±3.30	11.11±3.50	10.51±3.32	0.92±1.53	0.85±1.61
TN	20.60±5.40	5.81±4.90	10.20±5.52	0.11±0.42	8.90±3.73
Naïve Bayes (NB)					
TP	25.30±4.53	5.52±2.01	0.03±0.21	3.01±2.11	1.33±2.32
TN	29.60±5.50	4.81±2.81	1.01±2.43	0.42±0.64	9.59±3.12
Bayesian Network (BNet)					
TP	26 ± 4.51	6.44 ± 2.83	0.69 ± 2.13	0.52 ± 0.77	0.32 ± 0.95
TN	36.02 ± 6.41	0 ± 0	0.26 ± 1.75	10.59 ± 3.59	0.41 ± 1.14
k-Nearest Neighbours (*k*-NN)					
TP	22.27 ± 4.86	4.94 ± 2.72	0.47 ± 1.03	6.84 ± 3.16	0.71 ± 0.66
TN	18.64 ± 5.92	16.04 ± 4.99	1.21 ± 3.17	0.64 ± 0.92	9.71 ± 3.09
C4.5					
TP	23.52 ± 4.39	3.96 ± 2.91	2.47 ± 5.63	4.48 ± 4.18	0.29 ± 0.53
TN	25.59 ± 8.95	3.34 ± 7.50	3.20 ± 7.98	16.33 ± 5.71	0.58 ± 1.57
Support Vector Machine (SVM)					
TP	14.12 ± 4.01	6.86 ± 8.42	10.25 ± 9.30	0.11 ± 0.97	11.21 ± 3.84
TN	22.37 ± 6.34	0.42 ± 2.99	14.5 ± 4.1	0 ± 0	0.28 ± 1.83
Alternate Decision Tree (ADTree)					
TP	4.88 ± 3.67	18.06 ± 5.06	3.63 ± 2.82	3.08 ± 3.13	5.51 ± 3.54
TN	3.49 ± 3.3	11.43 ± 4.58	12.06 ± 5.14	8.79 ± 4.03	10.6 ± 3.18
Logistic Model Tree (LMT)					
TP	22.68 ± 4.21	5.27 ± 2.15	1.05 ± 1.93	5.42 ± 3.28	1.31 ± 1.95
TN	18.61 ± 4.51	10.62 ± 3.69	6.07 ± 3.1	3.61 ± 4.4	7.15 ± 5.63
AdaBoostM1 (AdaBoost)					
TP	20.82 ± 3.84	2.59 ± 3.39	10.16 ± 4.30	0.15 ± 0.57	1.05 ± 1.88
TN	17.21 ± 5.71	11.96 ± 5.59	6.62 ± 4.52	1.16 ± 1.31	8.85 ± 3.38

very important. The investigation of confidence levels in the development of CAD systems is essential as mentioned in a recent study[8]. This study could lead to a development of a new evaluation metric which could be used to estimate the degree of certainty of the system, hence provides positive potential impacts of CAD systems [8, 17]. This study indicates that using A and AUC in conjunction with confidence measure to provide a more transparent representation of the actual reliability of CAD systems which is similar to our previous study [18]. In conclusion, we have studied the reliability of the confidence measures of 11 different classifiers for the breast mass classification which indicates that although a system could produce high accuracy or AUC, it does not provide a full indication about the confidence reliability of the system. In future work we plan to compare recent popular algorithms and investigate the probability outputs of deep learning based methods in classification particularly in the field of medical image analysis.

Acknowledgment

This research was undertaken as part of the Decision Support and Information Management System for Breast Cancer (DESIREE) project. The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 690238.

5. REFERENCES

[1] A. Rampun, L. Zheng, P. Malcolm, B. Tiddeman, and Reyer Zwiggelaar, "Computer-aided detection of prostate cancer in t2-weighted mri within the peripheral zone," *Physics in Medicine & Biology*, vol. 61, no. 13, pp. 4796, 2016.

[2] A. Rampun, Z. Chen, P. Malcolm, B. Tiddeman, and R. Zwiggelaar, "Computeraided diagnosis: detection and localization of prostate cancer within the peripheral zone," *International journal for numerical methods in biomedical engineering*, vol. 32, no. 5, pp. e02745, 2016.

[3] A. Rampun, P. J. Morrow, B. W Scotney, and J. Winder, "Fully automated breast boundary and pectoral muscle segmentation in mammograms," *Artificial intelligence in medicine*, vol. 79, pp. 28–41, 2017.

[4] A. Rampun, B. Tiddeman, and P. Malcolm R. Zwiggelaar, "Computer aided diagnosis of prostate cancer: A texton based approach," *Medical physics*, vol. 43, no. 10, pp. 5412–5425, 2016.

[5] A Rampun, B. W. Scotney, P. J. Morrow, H. Wang, and J Winder, "Breast density classification using local quinary patterns with various neighbourhood topologies," *Journal of Imaging*, vol. 4, no. 10, 2016, doi: 10.3390/jimaging4010014.

[6] A. Hamidinekoo, E. Denton, A. Rampun, K. Honnor, and R. Zwiggelaar, "Deep learning in mammography and breast histology, an overview and future trends," *Medical image analysis*, vol. 47, pp. 45–67, 2018.

[7] P. Shi, J. Zhong, A. Rampun, and H. Wang, "A hierarchical pipeline for breast boundary segmentation and calcification detection in mammograms," *Computers in biology and medicine*, vol. 96, pp. 178–188, 2018.

[8] W. Jorritsma, F. Cnossen, and P. M. van Ooijen, "Improving the radiologist-cad interaction: designing for appropriate trust.," *Clin Radiol.*, vol. 70, no. 2, pp. 115–122, 2015.

[9] A. Cahan, D. Gilon, O. Manor, and O. Paltiel, "Probabilistic reasoning and clinical decision-making: do doctors overestimate diagnostic probabilities?," *QJM: An International Journal of Medicine*, vol. 96, no. 10, pp. 763–769, 2003.

[10] L. Tabár, B. Vitak, T. H-H. Chen, A. M-F. Yen, A. Cohen, T. Tot, S. Y-H. Chiu, S. L-S. Chen, J. C-Y. Fann, J. Rosell, H. Fohlin, R. A. Smith, and S. W. Duffy, "Swedish two-county trial: Impact of mammographic screening on breast cancer mortality during 3 decades," *Radiology*, vol. 260, no. 3, pp. 658–663, 2011.

[11] K. Doi, "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential," *Computerized medical imaging and graphics*, vol. 31, no. 5, pp. 198–211, 2007.

[12] L. E. Philpotts, "Can computer-aided detection be detrimental to mammographic interpretation?," *Radiology*, vol. 253, no. 1, pp. 17–22, 2009.

[13] A. Cahan and J. J. Cimino, "A learning health care system using computer-aided diagnosis," *Journal of Medical Internet Research.*, vol. 19, no. 3, pp. e54, 2017.

[14] R. S. Lee, F. Gimenez, A. Hoogi, and D. Rubin, "Curated breast imaging subset of DDSM. the cancer imaging archive," 2016.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[16] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *In Proceedings 17th International Conference of Machine Learning*, 2000, pp. 359–366.

[17] S. Astley, R. Zwiggelaar, C. Wolstenholme, K. Davies, T. Parr, and C. Taylor, *Prompting in Mammography: How Good Must Prompt Generators Be?*, pp. 347–354, Springer Netherlands, 1998.

[18] A. Rampun, P. Morrow H. wang, B. Scotney, and R. Zwiggelaar, "Classification of mammographic microcalcification clusters with machine learning confidence levels," in *Proc. 14th International Workshop on Breast Imaging (IWBI-2018)*, in press.