Classification of mammographic microcalcification clusters with machine learning confidence levels

Conference Paper · July 2018

DOI: 10.1117/12.2318058

5 authors, including:

Andrik Rampun
The University of Sheffield
28 PUBLICATIONS **144** CITATIONS
SEE PROFILE

Hui Wang
Zhengzhou University of Light Industry
56 PUBLICATIONS **452** CITATIONS
SEE PROFILE

Bryan Scotney
Ulster University
269 PUBLICATIONS **1,886** CITATIONS
SEE PROFILE

Reyer Zwiggelaar
Aberystwyth University
246 PUBLICATIONS **2,456** CITATIONS
SEE PROFILE

Classification of Mammographic Microcalcification Clusters With Machine Learning Confidence Levels

Andrik Rampun[a], Hui Wang[b], Bryan Scotney[a], Philip Morrow[a] and Reyer Zwiggelaar[c]

[a] School of Computing, Ulster University, Coleraine, Northern Ireland, UK
[b] School of Computing, Ulster University, Jordanstown, Newtownabbey, Northern Ireland, UK
[c] Department of Computer Science, Aberystwyth University, UK

ABSTRACT

This paper presents a novel investigation of machine learning performance by examining probability outputs in conjunction with classification accuracy (CA) and area under the curve (AUC). One of the main issues in the deployment of computer-aided detection/diagnosis (CAD) systems is lack of 'trust' of clinicians in the CAD system, increasing the possibility of the system not being used. Whilst most authors evaluate the performance of their breast CAD systems based on CA and AUC, we study the distribution of the classifiers' probability outputs and use it as an additional confidence level metric to indicate the reliability of a computer system. Experimental results suggest that although most classifiers produce similar results in terms of CA and AUC (less than 2% variation), their performances are significantly different when considering confidence level (10 to 25% difference). This study may provide opportunities for refining radiologists' interaction with CAD systems and improving the reliability of CAD systems as well as diagnostic decision making in medicine with high CA or AUC with high degree of certainty.

Keywords: Microcalcification, Confidence Levels, Breast Mammography, Computer-Aided Diagnosis

1. INTRODUCTION

Screening mammography is one of the most common imaging techniques for early detection of breast cancer. According to the American College of Radiology, breast composition, masses and calcifications are among the most important features in mammography assessment because they could indicate an early sign of abnormality. Nowadays, computer-aided diagnosis/detection (CAD) systems are being widely used in hospitals to speed up diagnosis. In breast mammography, CAD can be used to estimate breast density,[1–3] measuring the characteristics of breast mass,[4,5] modelling the distribution of microcalcification cluster,[6] for risk assessment[7] and as a pre-processing method.[8] Breast CAD systems in clinical environments are seen as an invaluable tool and are often used as a 'second reader' opinion. This means, that breast CAD system systems are used as a tool to provide additional information beyond human perception, hence supporting clinicians/radiologists when they are unsure about their decisions.

Although several studies have shown that radiologists' diagnostic performances have improved when assisted by a CAD system, accuracy is often lower than might be expected based on a radiologist's individual performance and the CAD system in isolation. This indicates 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 a real clinical environment from reaching their full potential. According to a recent study conducted by Jorritsma *et al.*,[9] one of the reasons is lack of trust of radiologists in the system. For example, a method may produce a CA value of 93% but the majority of the cases are classified with a confidence value between 0.50 to 0.60, so the radiologist would be likely to ignore the computer output, resulting in under-trust from a user perspective when using the system. Figure 1 shows a synthetic illustration of this example. It can be observed that for classifier A, many cases are located within the red area (close to the decision boundary; low confidence) despite having a higher CA value than classifier B. On the other hand, many cases are located within the green area (far from the decision boundary (high confidence)) for classifier B although they have a lower CA value.

In this paper, we investigate the performance of machine learning algorithms in classification of microcalcification clusters from a confidence level perspective. We are interested in the question: if a case is classified as

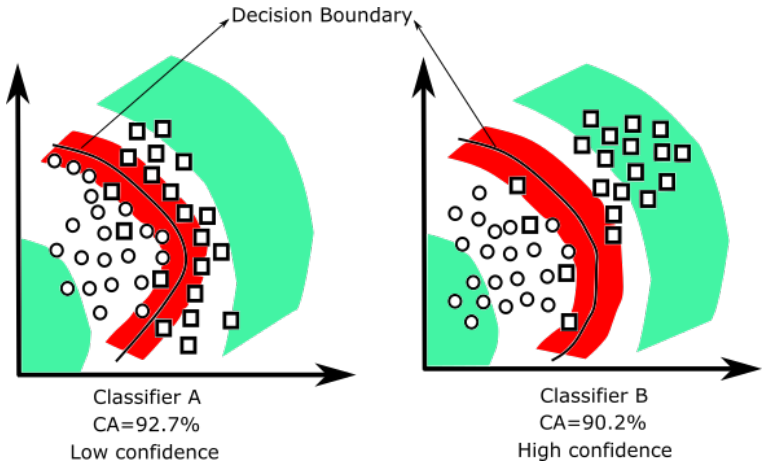

Figure 1. A synthetic illustration of the confidence levels investigation in this study. The green and red regions are areas with high and low confidence levels, respectively.

'malignant', what is the degree of certainty of the case being actually malignant? We seek to answer this question by looking at the probability outputs produced by the classifier. The majority of the breast CAD studies in the literature use AUC, accuracy, sensitivity and specificity as evaluation metrics, which are already widely accepted. However, these metrics can provide incorrect representation of the actual reliability of the system (as illustrated in Figure 1). In contrast, we propose the use of confidence levels as a new alternative evaluation metric in conjunction with the existing metrics such as AUC, accuracy, sensitivity and specificity. For this purpose, we investigate 11 machine learning algorithms and show how their performances can be significantly different in terms of confidence levels despite having small variations in AUC or CA.

2. LITERATURE REVIEW

Microcalcifications are small specks of calcium deposits scattered in the breast tissue and can be seen as small white dots or dashes in mammograms. Although they are extremely common in women, the presence of microcalcification clusters could be an indicative of acute or potential breast cancer. In the last decades, many authors have developed CAD methods for classification of microcalcification clusters and reported very promising results of up to 100% and 0.98 AUC value. However, the majority of the studies are evaluated based on small datasets. In terms of feature descriptors, many have been studied in the literature such as shape, morphological, cluster, intensity-based, and texture features. Figure 2 shows examples of malignant and benign cases.

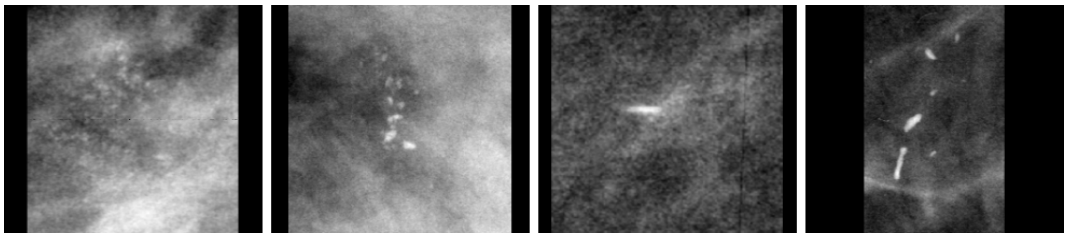
Figure 2. From left to right the first two and last two images are malignant and benign, respectively.

In an early study, Dhawan *et al.*[10] used texture and cluster information to determine whether a microcalcification cluster was benign or malignant using an artificial neural network (ANN) ($AUC = 0.82$ based on 191 cases). The study of Chen *et al.*[6] proposed a method by modelling the topological structure of the microcalcification in a graph-based representation and extracting a set of features using different scales before using the k-nearest neighbors algorithm for classification ($AUC = 0.96$ and $CA = 96\%$). Strange *et al.*[11] developed a multi-scale shape descriptor using persistent discrete mereotopology inspired by discrete mereotopology and computational topology ($AUC = 0.82$ based on 300 samples). Ren *et al.*,[12] using a combination of different features such as

texture, morphology, cluster, clinical information and first order statistical features reported an $AUC = 0.98$ based on 159 patches using the ANN classifier.

In recent years, studies in microcalcification diagnosis used several datasets. Kai *et al.*[13] employed an extreme learning machine based on a hidden Markov tree model which is based on the complex wavelet transform technique. The proposed method was evaluated on three datasets (Nijmegen, Mammographic Image Analysis Society (MIAS) and Digital Database for Screening Mammography (DDSM)) with a range of AUC values from 0.91 to 0.99. Yang *et al.*[14] employed the Simplified Pulse Coupled Neural Network (SPCNN) to classify each case based on wavelet high-frequency coefficients and reported an $AUC = 0.97$ with an accuracy of over 93% tested on both the MIAS and Japanese Society of Medical Imaging Technology datasets. Suhail *et al.*[15] used a binary tree-based approach to model the clinical perception features such as the number of microcalcifications and the distribution pattern; which can be mathematically computed based on the topology of the trees and the connectivity of the microcalcifications. The authors reported 91% accuracy based on 288 patches extracted from the DDSM dataset. Based on the same dataset, Suhail *et al.*[16] proposed another method using a scalable linear Fisher discriminant analysis approach for the linear transformation of segmented micro-calcification data followed by the Support Vector Machine (SVM) variant as a classification approach. The authors reported an accuracy of 96% with $AUC = 0.95$, which is comparable to the state-of-the art alternative methods in the literature.

According to the survey conducted by Cheng *et al.*,[17] none of the developed methods reported the performance of their methods in terms of confidence levels or degree of certainty. Moreover, the majority of the studies used AUC, CA, sensitivity and specificity as evaluation metrics. In terms of the size of datasets, most of the previous proposed methods were evaluated based on a small dataset (e.g. under 300 samples/patches). In recent studies, most of the proposed methods are evaluated based on several datasets reporting more extensive results analysis. Nevertheless, the overall number of samples/patches is still considered small in comparison to the number of data available publicly. In our study, we use a larger number of microcalcifications clusters taken from the Curated Breast Imaging Subset of DDSM (CBIS-DDSM):[18] in total 1872 samples covering 753 patients.

3. METHODOLOGY

3.1 Feature Descriptions

In this study, we used clinical features which are already included in the CBIS-DDSM[18] database. The motivations for using clinical features are two-fold; firstly this study aims to investigate the reliability of machine learning algorithms which is the backbone of CAD systems instead of developing new feature descriptors; secondly the associated clinical features included in the dataset are known to be highly correlated with the clinical outcomes.

The following clinical features are used in this study:(1) Breast density ($F_d \in \{1, 2, 3, 4\}$), which indicates the amount of dense tissues within the breast and indicates the risk of developing breast cancer based on the following scales: (a) BI-RADS I (0-25% dense tissues, predominantly fat), (b) BI-RADS II (26-50% dense tissues, fat with some fibro-glandular tissue), (c) BI-RADS III (51-75% dense tissues, heterogeneously dense) and (d) BI-RADS IV (above 75% dense tissues, extremely dense); (2) Calcification type ($F_t \in \{F_t^1 ... F_t^{14}\}$). This describes the shape and appearance of the microcalcification cluster such as amorphous (F_t^1), pleomorphic (F_t^2), round and regular (F_t^3), coarse (F_t^4), lucent center (F_t^5), dystrophic (F_t^6), eggshell (F_t^7), fine linear branching (F_t^8), large rod like (F_t^9), punctate (F_t^{10}), milk of calcium (F_t^{11}), n/a (F_t^{12}), skin (F_t^{13}), and vascular (F_t^{14}). Each cluster can be a combination of two or more types of calcification such as amorphous and pleomorphic; (3) Pattern distribution ($F_n \in \{F_n^1 ... F_n^6\}$). This describes the types of distribution of the microcalcification based on the following patterns: clustered (F_n^1), linear (F_n^2), regional (F_n^3), diffusely scattered (F_n^4), segmental (F_n^5), n/a (F_n^6). Each cluster can be a combination of two or more types of calcification such as linear and regional; (4) Subtlety ($F_s \in \{1, 2, 3, 4, 5\}$) describes the subtlety of the breast tissue around the microcalcification cluster and (5) Clinical assessment ($F_a \in \{0, 1, 2, 3, 4, 5\}$) represents the radiologist's BI-RADS abnormality assessment (different from the BI-RADS density assessment) based on the features above along with other patient clinical information.

3.2 Feature Representation

To represent these clinical features as a feature vector (F), each feature is concatenated as $F = \{F_d, F_t, F_n, F_s, F_a\}$. Since the values of F_t and F_n can be many due to possible combination of different types and distributions, we present both features as a set of 'binary representations'. Therefore, if the pattern distribution is $F_n = \{F_n^1, ..., F_n^5\}$ then, for example, if it is a combination of regional (F_n^3) and diffusely scattered (F_n^4), the resulting F_n is represented as [0,0,1,1,0,0] (the same applied to F_t).

The main reason for transforming the clinical features into a 'binary representations' is to limit the variations, hence simplifying the complexity of the data, which can help the classifier to build a more robust predictive model. These new features may have a different interpretation from the original features, but they also could provide more discriminatory power (e.g. feature transformation could smooth class boundaries) in a different space than the original space.

3.3 Machine Learning Algorithms

This study employed 11 machine learning classifiers, some of which are commonly used in CAD systems for microcalcifications. The machine learning algorithms employed in this study are the Random Forest (RF), Multilayer Perceptron (MLP), Logistic Regression (LR), Naïve Bayes (NB), Bayesian Network (BNet), k-Nearest Neighbours (k-NN), C4.5 (J48), Alternate Decision Tree (ADTree), Logistic Model Trees (LMT), AdaBoostM1 (AdaBoost) and Support Vector Machine (SVM). Following our previous studies,[19,20] we used the CVParameterSelection and GridSearch techniques to select associated hyper-parameters for each classifier (available in the WEKA data mining suite).

For the classifiers with only one parameter (e.g. k-Nearest Neighbours), the CVParameterSelection technique was employed in the WEKA data mining suite[21] for parameter selection. In contrast, the GridSearch technique was used to explore two parameters for classifiers with two parameters (e.g. Random Forest and the Support Vector Machine). To select parameter values for each classifier, 10% of the patients in the training set were selected randomly and 3-fold cross validation was used to evaluate the performance for each (or each pair of) tested parameter(s) during the validation process. Subsequently, we employed parameter values that optimised the accuracy during the validation process. Note that investigating optimal parameters is not within the scope of our study. Therefore, other parameters such as kernel type and searching algorithm method are not tested. For experimental settings a patient based stratified ten-runs 10-fold cross validation (10-FCV) scheme was employed to ensure no samples from the same patient were used in both training and testing datasets. Each classifier was trained, and in the testing phase each sample was classified as malignant or benign.

4. EXPERIMENTAL RESULTS

The dataset used in this study is taken from the CBIS-DDSM[18] database, which is an updated and standardised version of the DDSM database. It is publicly available from the Cancer Imaging Archive. Each case is a biopsy proven 'benign'/'malignant' annotated by an expert radiologist. In total it contains 1872 microcalcification clusters (1199 benign and 673 malignant) from 753 patients. 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. We evaluated the performance of each classifier using the most common evaluation metrics in the literature, namely classification accuracy (CA) and area under the curve (AUC).

Subsequently, we evaluated 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'. 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[21] for computing P values for each classifier. We categorised the values into the following classes: (a) Confidence 1 ($C1$): $0.50 \leq P \leq 0.59$, (b) Confidence 2 ($C2$): $0.60 \leq P \leq 0.69$, (c) Confidence 3 ($C3$): $0.70 \leq P \leq 0.79$, (d) Confidence 4 ($C4$): $0.80 \leq P \leq 0.89$, and (e) Confidence 5 ($C5$): $0.90 \leq P \leq 1.0$. In our study, we calculated the confidence level as the percentage of correctly classified cases (true positive (TP) and true negative (TN), e.g. $\sum(TP + TN)$) falling into each category, which means there are five confidence levels for each classifier. The higher the cumulative percentage for $C4$ and $C5$, the better the result in terms of confidence level.

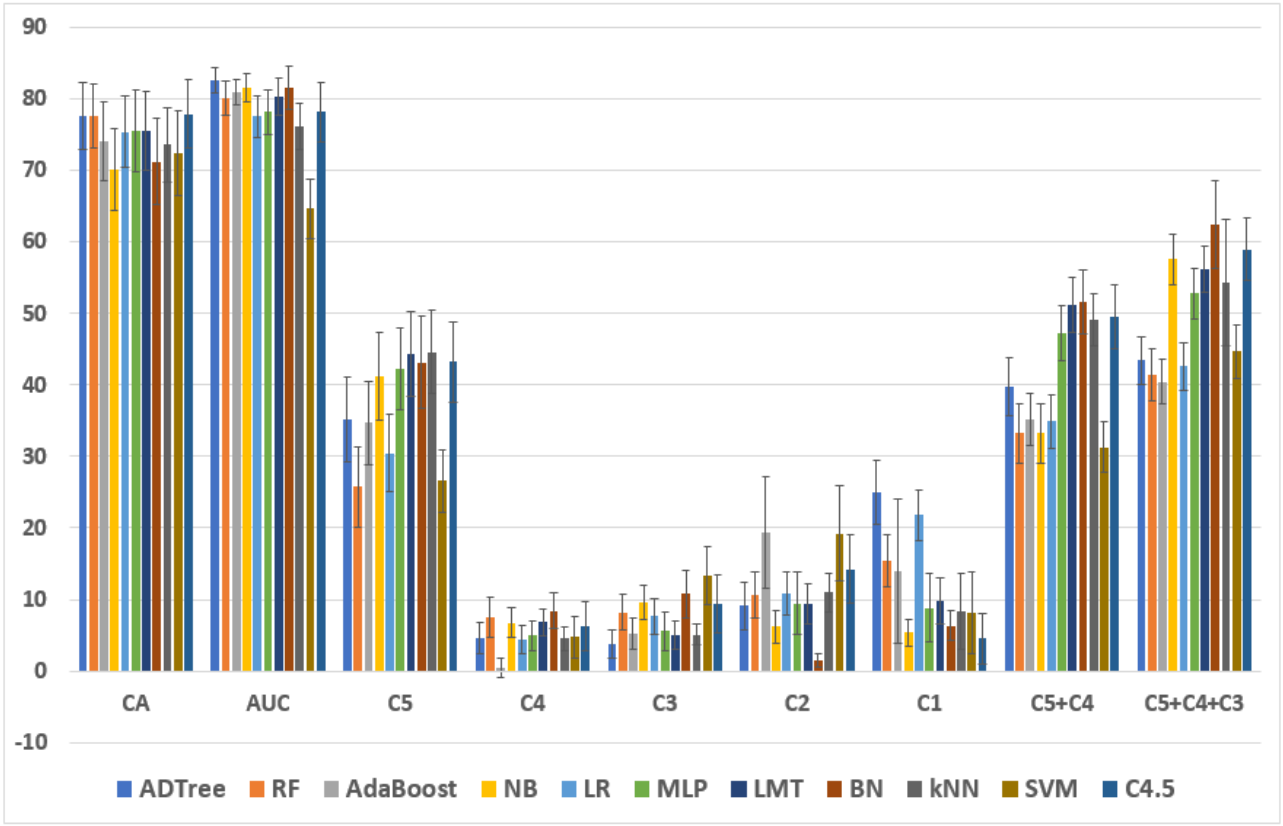

Figure 3. Quantitative results based on CA, AUC, $C5$, $C4$, $C3$, $C2$, $C1$, $C5 + C4$ and $C5 + C4 + C3$. The x-axis and y-axis represents the percentage (%) and evaluation metric, respectively.

Figure 3 shows quantitative results for all 11 classifiers, which shows that in terms of CA the C4.5 (77.8%), RF (77.6%) and ADTree (77.5%) are in the top three, but in terms of AUC performance the ADTree (82.5%), NB and BN (81.5%) and AdaBoost (80.9%) produced the top three results. The worst performing CA and AUC are the NB (70%) and the SVM (64.5%), respectively. In general, most classifiers produced very similar results with around 5% difference in both metrics except the AUC value produced by the SVM classifier (65%). However, in terms of confidence levels the k-NN and LMT classifiers have the highest number of TP and TN (around 44.6%) classified with $0.90 \leq P \leq 1.0$ followed by the C4.5 (43.2%). The worst performing classifier is RF with only 25.75% of the correctly classified cases falling in the $C5$ category although it has one of the highest classification accuracies. The ADTree, which has the highest AUC, classified 35.15% of the cases with high confidence ($P \geq 0.9$). Furthermore, only a small number of cases was classified with $0.80 \leq P \leq 0.89$ across all classifiers. The BN classifier produced 8.45% followed by RF (7.47%). For cases correctly classified with $0.50 \leq P \leq 0.59$ (e.g. low confidence levels), the ADTree produced 24.92% followed by the LR (21.79) and RF (15.46) classifiers, which indicates that there are a large number of cases classified with a low degree of certainty (hence resulting in a system having low confidence). When taking $\sum(C5 + C4)$ into account as a measurement of degree of certainty, the BN, LMT and C4.5 classifiers produced three best results of 51.57%, 51.16% and 49.55%, respectively. Other classifiers such as k-NN produced 48.34% and MLP 47.14%. The ADTree and RF were once again among the worst performing classifiers with less than 40%. On the other hand, taking $C3$ as a threshold for the degree, of certainty the BN, C4.5 and NB are the three best performing classifiers with $\sum(C5 + C4 + C3)$ values of 62.34%, 58.96% and 57.51%, respectively. The AdaBoost classifier has the lowest confidence level of 40.41% and the RF classifier achieved 41.47%.

Figure 4 shows the confidence level performances in terms of precision (PR). In this study, precision measures the accuracy of correctly classified cases for each category or confidence level. The PR metric can be calculated as $\frac{TP+TN}{TP+TN+FP+FN}$, where FP and FN are false positive and false negative, respectively. It can be observed that

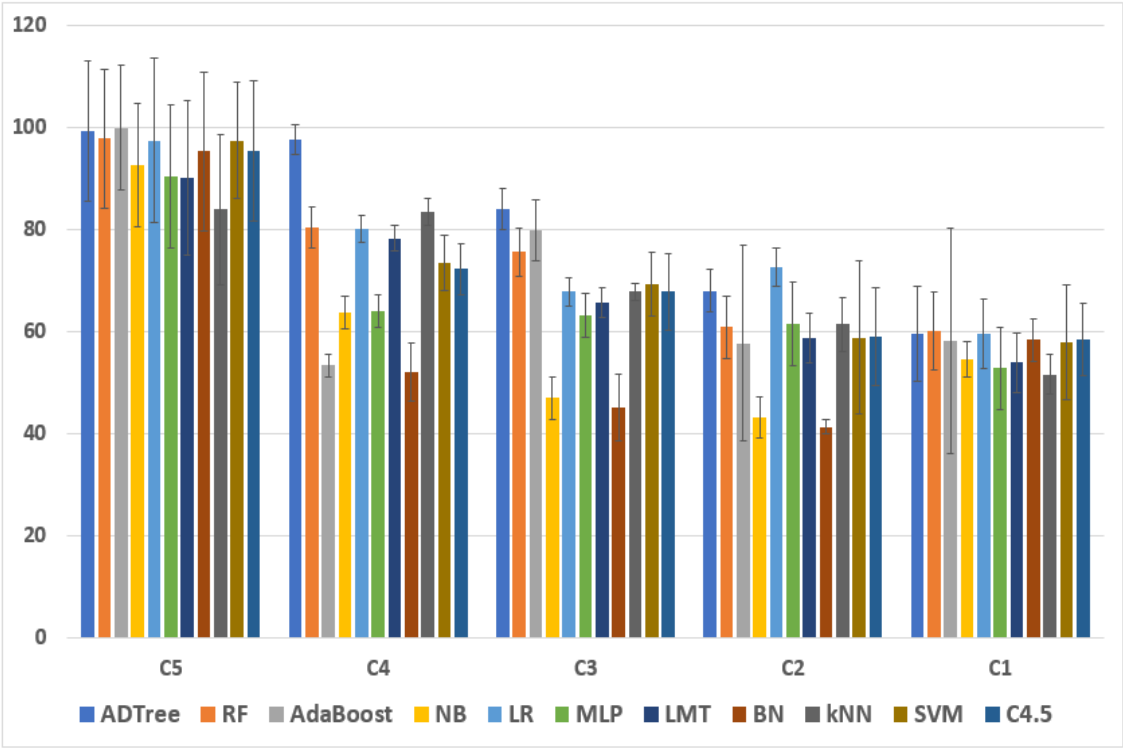

Figure 4. Quantitative results based on precision. The y-axis and x-axis represents the evaluation metric and percentage (%), respectively.

the AdaBoost classifier has the highest precision for $C5$ with 99.85% of the cases that have $P \geq 0.9$ correctly classified, which indicates that only a very small number of cases are FP or FN. The ADTree came second with precision 99.19% followed by the RF (97.8%) classifier. The worst performing classifier is the k-NN classifier with only 84% of the cases with $P \geq 0.9$ correctly classified. This means that 16% of the cases with high probability output values are either FP or FN. However, the precision in the $C4$ category for the AdaBoost classifier has the second lowest with 53.29%. The worst performing classifier in terms of precision in this category is the BN classifier with 52.04%. The ADTree classifier remains the top performing classifier with over 97% precision and has among the best precision value across the five different categories. The BN classifier produced good PR for $C5$ and $C1$ but is among the worst classifiers for $C4$, $C3$ and $C2$. For $C1$, all classifiers produced low precision results with on average below 60% with small variations.

5. DISCUSSION

From this study, experimental results indicate that:

1. In breast CAD systems, the CA or AUC metric alone does not provide a complete representation of the reliability of the system particularly from a confidence perspective. In this study, we have shown that the ADTree and the RF classifiers are among the top performing classifiers based on CA and AUC but many of the cases are classified with a low degree of certainty. From a confidence level perspective, in category $C5$ the k-NN and LMT classifiers have the highest number of cases classified correctly with 44.6%. The RF classifier achieved only 25.75%, which is among the worst performing classifiers.

2. Although most of the classifiers employed in this study produced similar results in terms of CA and AUC, their performances are significantly different when investigating the probability outputs for each confidence category. Based on our experimental results, the standard deviations across categories $C5$, $C4$, $C3$, $C2$ and $C1$ are 7.2%, 2.2%, 3.1%, 5.3% and 6.8%, respectively.

3. In terms of precision, the majority of the classifiers achieved 97% for $C5$, which indicates that most of the cases with high probability outputs are correctly classified. Our experimental results suggest that the ADTree classifier is the most consistent learning algorithm across different confidence levels.

4. This study shows the importance of investigating confidence levels in the development of CAD systems in improving interaction between clinicians and CAD systems which was also mentioned in the study by Jorritsma *et al.*[9] The current metrics such as CA, AUC, sensitivity and specificity need to be used in conjunction with a confidence measure as well as precision to provide a more transparent representation of the actual reliability of CAD systems.

5. This study may lead to the development of a new evaluation metric which may be useful to measure the degree of confidence of the system and model selection in ensemble based classifiers. For example, selecting a set of weak classifiers based on the confidence measure rather than based on AUC or CA alone.

6. CONCLUSION

In conclusion, we have studied the reliability of 11 classifiers used in the classification of microcalcification clusters. In contrast to the majority of the current breast CAD systems, we used confidence levels to measure the reliability of a system in conjunction with CA and AUC, and we further investigated the precision for each confidence category. Experimental results suggest that using CA and AUC alone are insufficient to provide the actual performance of CAD systems and they need to be evaluated using confidence levels and precision for each category. Therefore, based on the results of our study we suggest the use of CA and AUC metrics in conjunction with precision and confidence measure as an additional evaluation metric in medical-based CAD systems.

In future work we plan to investigate the probability outputs of deep learning based methods in classification particularly in the field of medical image analysis. This could lead to two research directions: firstly, by analysing the probability outputs of TP and TN cases we may reveal the reliability from a degree of certainty perspective of deep learning based methods rather than relying on just the CA or AUC performances. Secondly, since most deep learning based methods use image features for classification, it would be informative to study the level of correlation between the extracted features and the probability outputs. We also plan to extend this investigation to different problem domains such as breast lesion calcification and density classification as these could indicate early signs of abnormality in breast cancer.

REFERENCES

[1] Rampun, A., Scotney, B. W., Morrow, P. J., Wang, H., and Winder, J., "Breast density classification using local quinary patterns with various neighbourhood topologies," *Journal of Imaging* **4**(1), 14 (2018).

[2] Rampun, A., Morrow, P. J., Scotney, B. W., and Winder, J., "Breast density classification using local ternary patterns in mammograms," in [*Proc. 14th International Conference Image Analysis and Recognition (Montreal, Canada, July 2017)*], *Lecture Notes in Computer Science* **10317**, Springer (2017).

[3] He, W., Juette, A., Denton, E. R. E., Oliver, A., Martí, R., and Zwiggelaar, R., "A review on automatic mammographic density and parenchymal segmentation," *International journal of breast cancer* **2015**, 1–31 (2015).

[4] Jiao, Z., Gao, X., Wang, Y., and Li, J., "A deep feature based framework for breast masses classification," *Neurocomputing* **197**, 221–231 (2016).

[5] Xie, W., Li, Y., and Ma, Y., "Breast mass classification in digital mammography based on extreme learning machine," *Neurocomputing* **193**, 930–941 (2016).

[6] Chen, Z., , Strange, H., Oliver, A., Denton, E. R. E., Boggis, C., and Zwiggelaar, R., "Topological modeling and classification of mammographic microcalcification clusters," *IEEE Trans. Biomed. Eng* **62**(4), 1203–1214 (2015).

[7] Rampun, A., Morrow, P. J., Scotney, B. W., and Winder, J., "A quantitative study of local ternary patterns for risk assessment in mammography," in [*Proc. 14th International Conference Image Analysis and Recognition (Montreal, Canada, July 2017)*], *Smart Innovation, Systems and Technologies* **17**, 283–286, Springer (2017).

[8] Rampun, A., Morrow, P. J., Scotney, B. W., and Winder, J., "Fully automated breast boundary and pectoral muscle segmentation in mammograms," *Artificial Intelligence in Medicine* **79**, 28–41 (2017).

[9] Jorritsma, W., Cnossen, F., and van Ooijen, P. M., "Improving the radiologist-cad interaction: designing for appropriate trust," *Clin Radiol.* **70**(2), 115–122 (2015).

[10] Dhawan, A. P., Chitre, Y., and Kaiser-Bonasso, C., "Analysis of mammographic microcalcifications using gray-level image structure features," *IEEE Trans. Med. Imag.* **15**(3), 246–259 (1996).

[11] Strange, H., Chen, Z., Denton, E. R. E., and Zwiggelaar, R., "Modelling mammographic microcalcification clusters using persistent mereotopology," *Pattern Recog. Lett.* **47**, 157–163 (2014).

[12] Ren, J., Wang, D., and Jiang, J., "Effective recognition of mccs in mammograms using an improved neural classifier," *Engineering Applications of Artificial Intelligence* **24**(4), 638–645 (2011).

[13] Kai, H., Wei, Y., and Xieping, G., "Microcalcification diagnosis in digital mammography using extreme learning machine based on hidden markov tree model of dual-tree complex wavelet transform," *Expert Systems with Applications* **86**, 135–144 (2017).

[14] Yang, Z., Dong, M., Guo, Y., Gao, X., Wang, K., Shi, B., and Ma, Y., "A new method of micro-calcifications detection in digitized mammograms based on improved simplified pcnn," *Neurocomputing* **218**, 79–90 (2017).

[15] Suhail, Z., Denton, E. R. E., and Zwiggelaar, R., "Tree-based modelling for the classification of mammographic benign and malignant micro-calcification clusters," *Multimedia Tools and Applications* (2017).

[16] Suhail, Z., Denton, E. R. E., and Zwiggelaar, R., "Classification of micro-calcification in mammograms using scalable linear fisher discriminant analysis," *Medical & Biological Engineering & Computing* , 1–11 (2018).

[17] Cheng, H. D., Cai, X., Chen, X., Hu, L., and Lou, X., "Computer-aided detection and classification of microcalcifications in mammograms: a survey," *Pattern Recognition* **36**(12), 2967–2991 (2003).

[18] Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., and Rubin, D. L., "A curated mammography data set for use in computer-aided detection and diagnosis research," *Sci Data* **4** (2017).

[19] Rampun, A., Zheng, L., Malcolm, P., Tiddeman, B., and Zwiggelaar, R., "Computer-aided detection of prostate cancer in t2-weighted mri within the peripheral zone," *Physics in Medicine & Biology* **61**(13), 4796–4825 (2016).

[20] Rampun, A., Tiddeman, B., Zwiggelaar, R., and Malcolm, P., "Computer aided diagnosis of prostate cancer: A texton based approach," *Medical physics* **43**(10), 5412–5425 (2016).

[21] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and H.Witten, I., "The weka data mining software: an update," in [*ACM SIGKDD exploration newsletter*], **11**, 10–18 (2009).