# Confidence Analysis for Breast Mass Image Classification

Andrik Rampun*, Hui Wang*, Bryan Scotney*, Philip Morrow* and Reyer Zwiggelaar‡
*School of Computing, Ulster University and ‡Department of Computer Science, Aberystwyth University, UK

## 1. Introduction

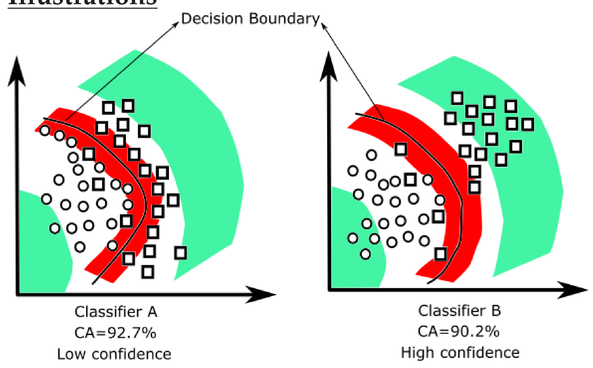
Many recent studies discussed the failure of Computer-Aided Diagnosis (CAD) systems. Recent studies of Kholi and Jha [1] and Jorritsma *et al.* [2] revealed that one of the main issues in the deployment of CAD systems is lack of 'trust' of clinicians in the CAD system, increasing the possibility of the system not being used.

## 2. Why CAD Failed

Computer scientists tend to report accuracy ($ACC$) and Area Under the Curve ($AUC$) to measure the performance of the method developed.

Unfortunately, these metrics do not measure the degree of confidence in individual recommendations. For example, a CAD method may produce an $AUC$ value of 0.97 but the majority of the cases are classified with a low confidence, which could reduce the acceptability of the classifications by radiologists.

**Illustrations**

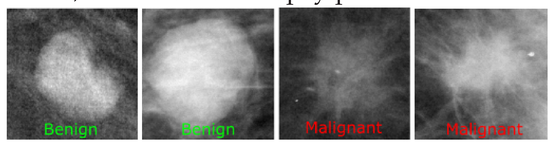



## 3. What Radiologist do

If the radiologist is not confident about his/her assessment, he/she could turn to the CAD system for support as a 'second reader'. If the CAD provides recommendation with a confidence of 0.60, the radiologist may deem it useless, however if the confidence measure is 0.90, the radiologist may deem it useful [2]. **Our goal is to investigate whether $ACC$ and $AUC$ have a direct correlation to confidence measure or not.**

## 4. Dataset

The CBIS-DDSM database [3] contains 1593 masses (829 benign and 764 malignant) from 838 patients, each case is a biopsy proven.




## 5. Methodology

**Feature Representation**

Each breast mass is represented based on the following characteristics: (1) Breast density ($F_d \in \{1,2,3,4\}$); (2) Mass shape ($F_s$): 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 ($F_s^6$), asymmetric breast tissue ($F_s^7$), architectural distortion ($F_s^8$), and irregular ($F_s^9$); (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 ($F_m^7$); (4) Subtlety ($F_t$) with values 1 to 5; and (5) BI-RADS assessment ($F_a$) with values 0 to 5. Each feature is concatenated as $F = \{F_d, F_s, F_m, F_t, F_a\}$.

**Classification Approach**

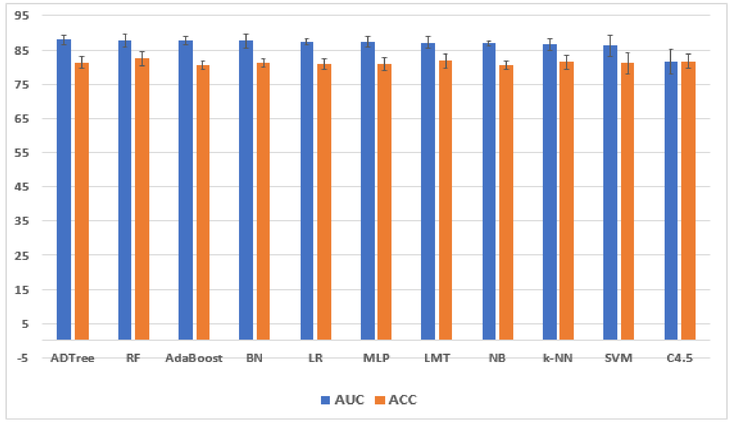
A patient based stratified ten-runs 10-fold cross-validation (10-FCV) scheme was employed. Eleven machine learning algorithms were employed with each optimised using the CVParameterSelection or GridSearch technique in the WEKA data mining suite.

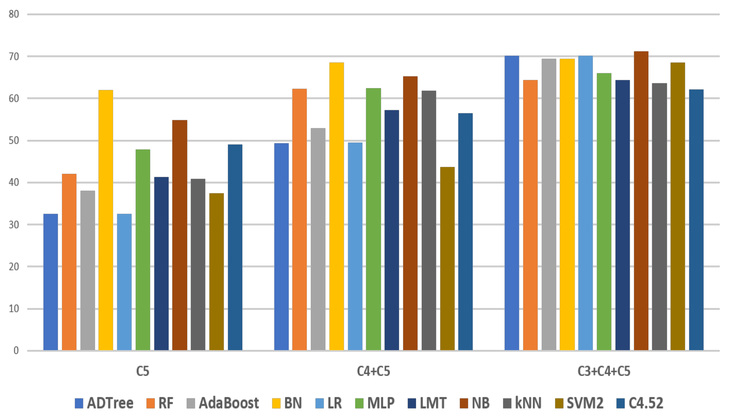
**Confidence metrics**

We used the probability outputs ($P \in (0,1]$) of the classifier as a confidence indication for each case being 'benign'/'malignant'. We categorise the values into the following classes:

- Confidence 1 ($C_1$): $0.50 \leq P \leq 0.59$
- Confidence 2 ($C_2$): $0.60 \leq P \leq 0.69$
- Confidence 3 ($C_3$): $0.70 \leq P \leq 0.79$
- Confidence 4 ($C_4$): $0.80 \leq P \leq 0.89$
- Confidence 5 ($C_5$): $0.90 \leq P \leq 1.0$

## 6. Experimental Results




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 $ACC = 82.51\%$ followed by the LMT classifier with $ACC = 81.79\%$, 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 $ACC$ and $AUC$.**

However, in terms of confidence measure 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 $P \geq 0.90$. Furthermore, with $P \geq 0.80$, the BNet once again performed the best with on average $68.48\%$ classified above this threshold value.

## 7. Conclusion

Although most of the classifiers produced similar results in terms of $ACC$ and $AUC$, their performances are different in terms of confidence measure. For example although the ADTree and RF classifiers produced the best $AUC$ and $ACC$ values, respectively, most cases were classified with $P \leq 0.70$.

## Acknowledgment & References

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.

[1] A. Kholi and S. Jha. Why CAD failed in mammography. *American College of Radiology*, 15(B):535–537, 2018.

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

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