Consistent performance measurement of a system to detect masses in mammograms based ... (2013)

TÃ-tulo:Consistent performance measurement of a system to detect masses in ammograms based on blind feature extraction
Autores: Antonio GarcÃ-aâ€“Manso, Carlos J. GarcÃ-aâ€“Orellana, Horacio M. GonzÃ¡lezâ€“Velasco, RamÃ³n Gallardoâ€“Caba and Miguel MacÃ-asâ€“MacÃ-as Revista:Â Biomedical Engineering OnLine
Vol./Pag.:Â 12, 1-18Ed./AÃ±o:Â BioMed Central LTD (Reino Unido), 2013DOI:10.1186/1475-925X-12-2Abstract:Background
Breast cancer continues to be a leading cause of cancer deaths among women, especially in Western countries. In the last two decades, many methods have been proposed to achieve a robust mammography-based computer aided detection (CAD) system. A CAD system should provide high performance over time and in different clinical situations. I.e., the system should be adaptable to different clinical situations and should provide consistent performance.
Methods
We tested our system seeking a measure of the guarantee of its consistent performance. The method is based on blind feature extraction by independent component analysis (ICA) and classification by neural networks (NN) or SVM classifiers. The test mammograms were from the Digital Database for Screening mammography (DDSM). This database was constructed collaboratively by four institutions over more than 10 years. We took advantage of this to train our system using the mammograms from each institution separately, and then testing it on the remaining mammograms. We performed another experiment to compare the results and thus obtain the measure sought. This experiment consists in to form the learning sets with all available prototypes regardless of the institution in which them were generated, obtaining in that way the overall results.
Results
The smallest variation from comparing the results of the testing set in each experiment (performed by training the system using the mammograms from one institution and testing with the remaining) with those of the overall result, considering the success rate for an intermediate decision maker threshold, was roughly 5%, and the largest variation was roughly 17%. But, if we considere the area under ROC curve, the smallest variation was close to 4%, and the largest variation was about a 6%.

Conclusions
Considering the heterogeneity in the datasets used to train and test our system in each case, we think that the variation of performance obtained when the results are compared with the overall results is acceptable in both cases, for NN and SVM classifiers. The present method is therefore very general in that it is able to adapt to different clinical situations and provide consistent performance.