Breast cancer is globally a major threat for women’s health. Screening and adequate follow-up can significantly reduce the mortality from breast cancer. Human second reading of screening mammograms can increase breast cancer detection rates, whereas this has not been proven for current computer-aided detection systems as “second reader”. Critical factors include the detection accuracy of the systems and the screening experience and training of the radiologist with the system. When assessing the performance of systems and system components, the choice of evaluation methods is particularly critical. Core assets herein are reference image databases and statistical methods. We have analyzed characteristics and usage of the currently largest publicly available mammography database, the Digital Database for Screening Mammography (DDSM) from the University of South Florida, in literature indexed in Medline, IEEE Xplore, SpringerLink, and SPIE, with respect to type of computer-aided diagnosis (CAD) (detection, CADe, or diagnostics, CADx), selection of database subsets, choice of evaluation method, and quality of descriptions. 59 publications presenting 106 evaluation studies met our selection criteria. In 54 studies (50.9%), the selection of test items (cases, images, regions of interest) extracted from the DDSM was not reproducible. Only 2 CADx studies, not any CADe studies, used the entire
DDSM. The number of test items varies from 100 to 6000. Different statistical evaluation methods are chosen. Most common are train/test (34.9% of the studies), leave-one-out (23.6%), and N-fold cross-validation (18.9%). Database-related terminology tends to be imprecise or ambiguous, especially regarding the term "case". Overall, both the use of the DDSM as data source for evaluation of mammography CAD systems, and the application of statistical evaluation methods were found highly diverse. Results reported from different studies are therefore hardly comparable. Drawbacks of the DDSM (e.g. varying quality of lesion annotations) may contribute to the reasons. But larger bias seems to be caused by authors' own decisions upon study design. RECOMMENDATIONS/CONCLUSION: For future evaluation studies, we derive a set of 13 recommendations concerning the construction and usage of a test database, as well as the application of statistical evaluation methods.