Artificial Intelligence for Digital Breast Tomosynthesis - Reader Study Results

WHITE PAPER

Abstract

ProFound AI™ is the latest artificial intelligence algorithm developed by iCAD, Inc., Nashua, NH using deep learning technology that is intended to be used concurrently by radiologists while reading digital breast tomosynthesis (DBT) exams.

The algorithm detects soft tissue densities (masses, architectural distortions and asymmetries) and calcifications in 3D DBT slices (Figures 1 and 2). The suspicious areas that are detected and highlighted and the unique certainty of finding and case scores assist radiologists in identifying and assessing soft tissue densities and calcifications that may be confirmed or dismissed by the radiologist.

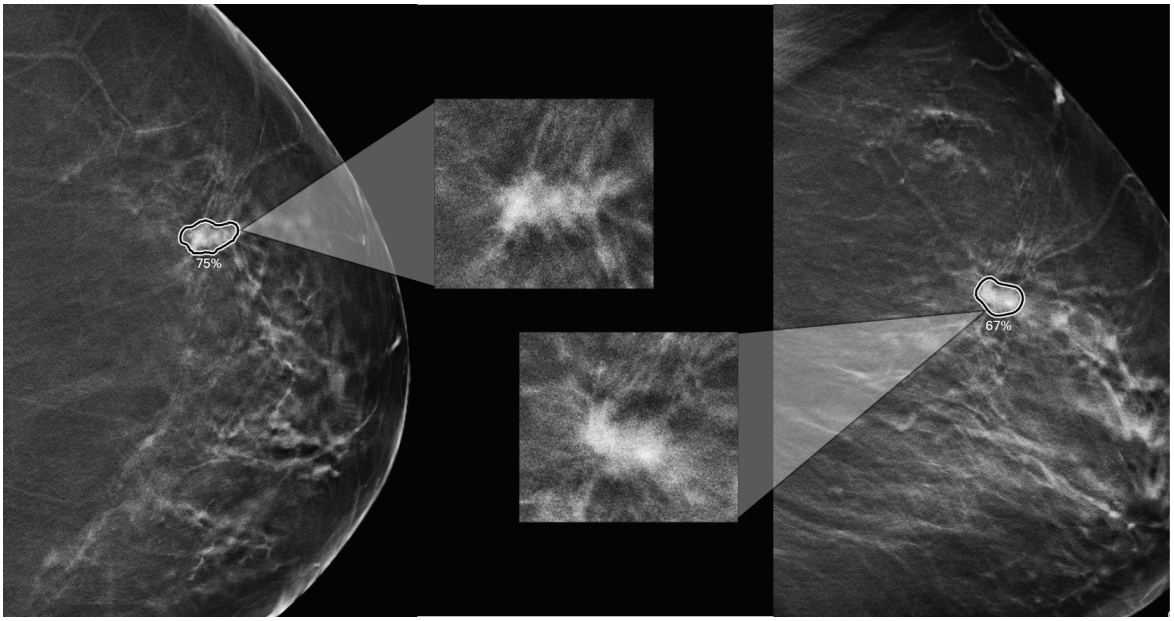

Figure 1: ProFound AI Detections and Certainty of Findings Scores for a Soft Tissue Density

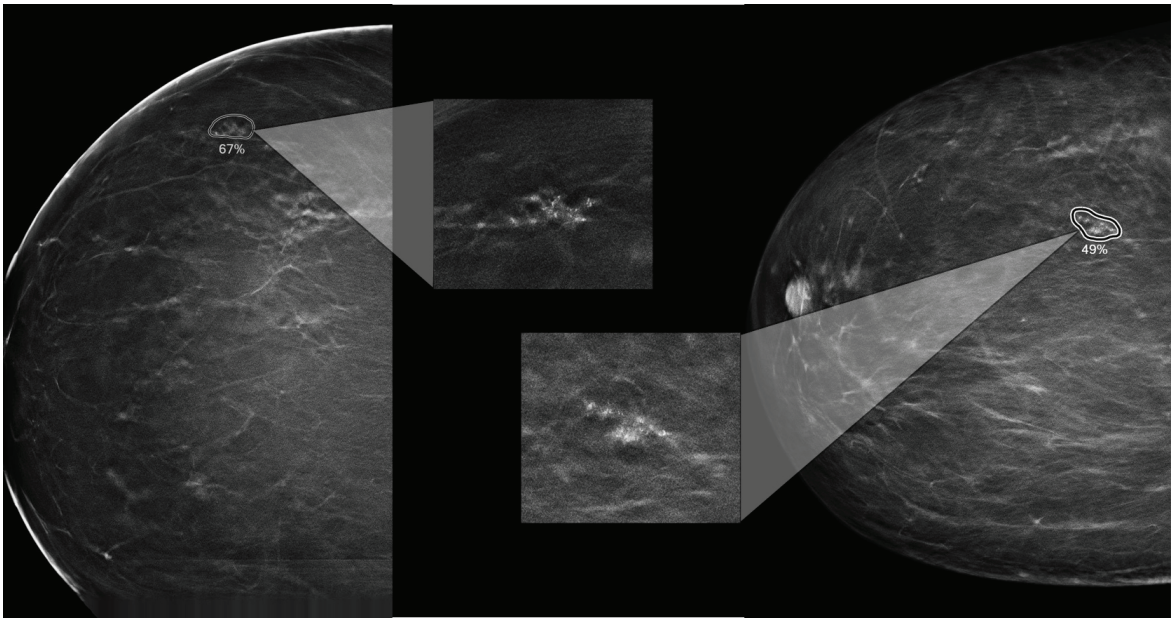

Figure 2: ProFound AI Detections and Certainty of Finding Scores for Calcifications

Introduction

The addition of DBT to full-field digital mammography (FFDM) improves radiologist performance by increasing cancer detection rates [1-4] and lowering recall rates [2-7], but also increases reading time almost two-fold [1, 8, 9], compared to 2D alone. Thus, ProFound AI was designed to maintain or improve radiologist clinical performance, while significantly reducing reading time.

Materials and Methods

Study Design

A retrospective, fully-crossed, multi-reader, multi-case (MRMC) clinical reader study with iCAD's ProFound AI algorithm was conducted with 24 radiologists reading an enriched set of 260 DBT cases, including 65 cancer cases with a total of 66 malignant lesions (Figure 3). The purpose of the study was to compare clinical performance of radiologists using ProFound AI and its certainty of finding and case scores to that of radiologists reading DBT without ProFound AI.

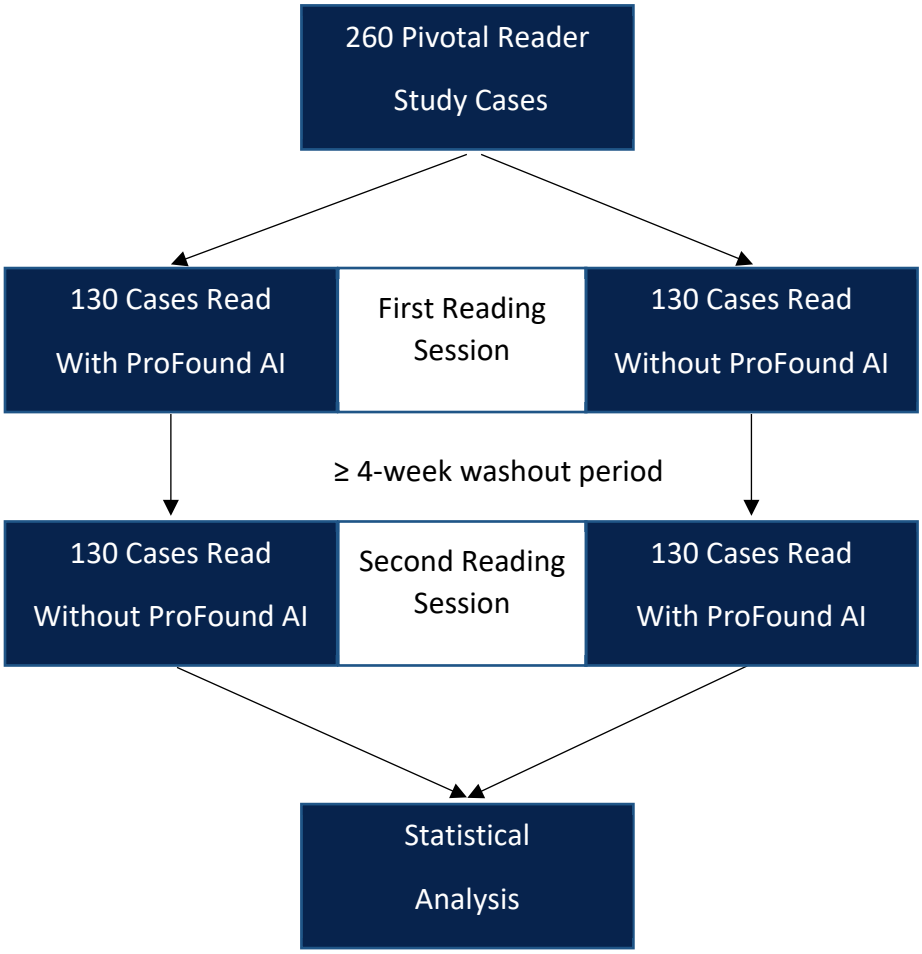

Figure 3: Study Design

Objectives

The objectives of the reader study were the following:

A. Co-primary objectives.

The co-primary objectives were to determine:

1. Whether radiologist performance when using ProFound AI with DBT images is non-inferior to radiologist performance when reading DBT images without ProFound AI, and

2. Whether radiologist reading time when using ProFound AI with DBT images is superior to (shorter than) radiologist reading time when reading DBT images without ProFound AI.

Radiologist performance was assessed by measuring case-level area under the receiver operating characteristic curve (AUC) for the detection of malignant lesions, where malignant lesion localization was required for a reader to correctly detect cancer in a case.

The study was considered to have successfully demonstrated safety and effectiveness of using ProFound AI with DBT compared to reading DBT without ProFound AI if the null hypothesis associated with non-inferiority of AUC is rejected and the null hypotheses associated with superiority of radiologist reading time is rejected. The hypothesis tests for the co-primary objectives were as follows:

1. Test the null hypothesis associated with non-inferiority of case-level AUC at two-sided statistical significance level alpha = 0.05. This null hypothesis will be rejected if the lower limit of the two-sided 95% confidence interval for the difference in average AUC with ProFound AI – without ProFound AI lies above the negative of the non-inferiority margin, -0.05.

2. Test the null hypothesis associated with superiority of radiologist reading time at two-sided statistical significance level alpha = 0.05. This null hypothesis will be rejected if the upper limit of the two-sided 95% confidence interval for the difference in average reading time with ProFound AI – without ProFound AI lies below zero, i.e., if reading time decreases.

B. Secondary objectives.

The secondary objectives of the reader study included the following for radiologists when using ProFound AI with DBT compared to using DBT without ProFound AI:

1. Superiority of case-level AUC

2. Non-inferiority (with non-inferiority margin delta = 0.05) of sensitivity at the case level

3. Superiority of sensitivity at the case level

4. Non-inferiority (with non-inferiority margin delta = 0.05) of sensitivity at the lesion level

5. Superiority of sensitivity at the lesion level

6. Non-inferiority (with non-inferiority margin delta = 0.05) of specificity (case-level)

7. Non-inferiority (with non-inferiority margin delta = 0.05) of recall rate in non-cancers (case-level)

Estimates and corresponding 95% confidence intervals illustrating precision in the estimates were provided for all secondary objectives. The study employed a fully-crossed design in which all readers reviewed images from all cases in two visits separated by a memory washout period of 4 weeks or more between readings of the same case with and without ProFound AI. Each reader was assigned to review half the cases with ProFound AI and the other half without ProFound AI during the first visit and the complementary with and without ProFound AI cases during the second visit, in a counterbalanced fashion, such that all the cases were read by each reader both with and without ProFound AI. The case reading order was randomized separately for each reader. Readers were informed that reading time was being measured and that ProFound AI is intended to reduce reading time, but readers were blinded to the reading time measurements for each case.

Results

The study results (Table 1) showed that both co-primary endpoints and all pre-specified secondary endpoints were met.

Objective	Result
Average Radiologist AUC	5.7% increase
Radiologist Reading Time	52.7% decrease
Radiologist Case-Level Sensitivity	8.0% increase
Radiologist Lesion-Level Sensitivity	8.4% increase
Radiologist Specificity	6.9% increase
Radiologist Recall Rate in non-cancers	7.2% decrease

Table 1: Summary of Reader Study Results

Specifically, evaluation of the co-primary endpoints of the study reader demonstrated the following:

1. Radiologist performance using ProFound AI with DBT was non-inferior to, and statistically significantly superior to, radiologist performance using DBT without ProFound AI. Radiologists had superior per-subject average area under the receiver operating characteristic curve (AUC) with ProFound AI, 0.852, versus without ProFound AI, 0.795. The average difference in AUC was 0.057 (95% CI: 0.028, 0.087; non-inferiority $p < 0.01$ for non-inferiority margin delta = 0.05, and $p < 0.01$ for test of difference). The average empirical receiver operating characteristic (ROC) plot across readers for with CAD is above the average empirical ROC plot for without CAD (Figure 4).

2. Radiologist reading time when using ProFound AI with DBT is superior to (shorter than) radiologist reading time when using DBT without ProFound AI. Reading time improved 52.7% with ProFound AI (95% CI: 41.8%, 61.5%; $p < 0.01$). *

   * Radiologist reading times may vary based on the specific functionality of the viewing application used for interpretation.

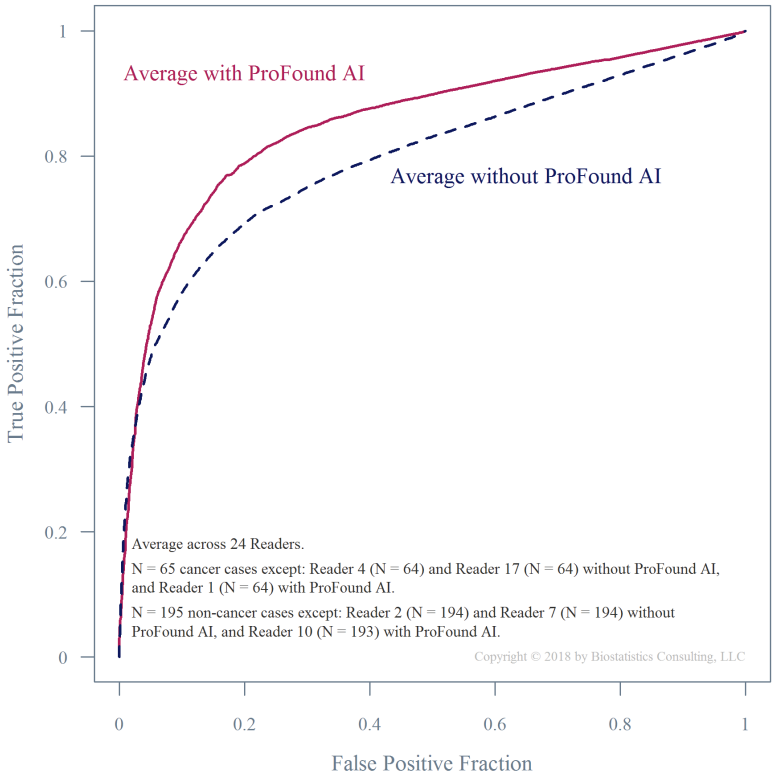

Figure 4. Use of Concurrent DBT ProFound AI System

In addition to superiority of case-level AUC, evaluation of the secondary endpoints of the reader study demonstrated the following:

- Radiologists had superior sensitivity at the case level with ProFound AI. Average sensitivity increased by 0.080 (95% CI: 0.026, 0.134; non-inferiority $p < 0.01$ for non-inferiority margin delta = 0.05, and $p < 0.01$ for test of difference). Average case-level sensitivity was 0.770 without ProFound AI and 0.850 with ProFound AI.
- At the lesion level, radiologists also had superior sensitivity with ProFound AI. Average per-lesion sensitivity across readers increased by 0.084 (95% CI: 0.029, 0.139; non-inferiority $p < 0.01$ for non-inferiority margin delta = 0.05, and $p < 0.01$ for test of difference), from 0.769 without ProFound AI to 0.853 with ProFound AI.
- Radiologists had non-inferior specificity with ProFound AI. Specificity was 0.627 without ProFound AI and 0.696 with ProFound AI, for an average increase of 0.069 (95% CI: 0.030, 0.108; non-inferiority $p < 0.01$ for non-inferiority margin delta = 0.05).
- Finally, radiologists had non-inferior recall rate in non-cancer cases with ProFound AI. In non-cancer cases, lower recall rates are better than higher recall rates. Average recall rate in non-cancer cases was 0.380 without ProFound AI and 0.309 with ProFound AI, for an average reduction of 0.072 (95% CI: 0.031, 0.112; non-inferiority $p < 0.01$ for non-inferiority margin delta = 0.05).

In this study the following were observed:

- Average sensitivity increased by 0.120 (SE = 0.040) in the subgroup of 15 cancer cases with only calcifications.
- Average sensitivity increased by 0.068 (SE = 0.031) in the subgroup of 50 cancer cases with at least one soft tissue density or mixed lesion.

- Average specificity decreased by 0.027 (SE=0.038) in the subgroup of 24 benign and recalled (non-cancer) cases with only calcifications.
- Average specificity increased by 0.079 (SE=0.028) in the subgroup of 62 benign and recalled (non-cancer) cases with at least one soft tissue density or mixed lesion.
- Average specificity increased by 0.084 (SE=0.021) in the subgroup of 109 non-cancer cases with no lesions.

References

[1] Skaane P, Bandos AI, Gullien R, et al. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. Radiology 2013; 267:47-56.

[2] Friedewald SM, Rafferty EA, Rose SL, et al. Breast cancer screening using tomosynthesis in combination with digital mammography. JAMA 2014; 311:2499-2507.

[3] Sharpe RE, Venkataraman S, Phillips J, et al. Increased cancer detection rate and variations in the recall rate resulting from implementation of 3D digital breast tomosynthesis into a population-based screening program. Radiology 2016; 278:698-706.

[4] Hooley RJ, Durand MA, Philpotts LE. Advances in digital breast tomosynthesis. AJR American Journal of Roentgenology 2017; 208:256-266.

[5] McDonald ES, Oustimov A, Weinstein SP, Synnestvedt MB, Schnall M, Conant EF. Effectiveness of digital breast tomosynthesis compared with digital mammography: outcomes analysis from 3 years of breast cancer screening. JAMA Oncology 2016; 2:737-743.

[6] Tucker L, Gilbert FJ, Astley SM, et al. Does reader performance with digital breast tomosynthesis vary according to experience with two-dimensional mammography? Radiology 2017; 283:371-380.

[7] Gilbert FJ, Tucker L, Gillan MGC, et al. Accuracy of digital breast tomosynthesis for depicting breast cancer subgroups in a UK retrospective reading study (TOMMY Trial). Radiology 2015; 277:697-706.

[8] Bernardi D, Ciatto S, Pellegrini M, et al. Application of breast tomosynthesis in screening: incremental effect on mammography acquisition and reading time. The British Journal of Radiology 2012; 85:e1174-1178.

[9] Dang PA, Freer PE, Humphrey KL, Halpern EF, Rafferty EA. Addition of tomosynthesis to conventional digital mammography: effect on image interpretation time of screening examinations. Radiology 2014; 270:49-56.

DMM253 Rev. A