Nature Communications

https://doi.org/10.1038/s41467-026-70637-3

Article in Press

Subgroup performance of a commercial digital breast tomosynthesis model for breast cancer detection

Received: 18 March 2025

Accepted: 24 February 2026

Cite this article as: Brown-Mulry, B., Isaac, R.S., Lee, S.H. *et al.* Subgroup performance of a commercial digital breast tomosynthesis model for breast cancer detection. *Nat Commun* (2026). https://doi.org/10.1038/s41467-026-70637-3

Beatrice Brown-Mulry, Rohan Satya Isaac, Sang Hyup Lee, Ambika Seth, KyungJee Min, Theo Dapamede, Frank Li, Aawez Mansuri, MinJae Woo, Christian Allison Fauria-Robinson, Bhavna Paryani, Judy Wawira Gichoya & Hari Trivedi

Subgroup Performance of a Commercial Digital Breast Tomosynthesis Model for Breast Cancer Detection

Beatrice Brown-Mulry[1,*], Rohan Satya Isaac[1,*], Sang Hyup Lee[2], Ambika Seth[2], KyungJee Min[2], Theo Dapamede[1], Frank Li[1], Aawez Mansuri[1], MinJae Woo[3], Christian Allison Fauria-Robinson[4], Bhavna Paryani[4], Judy Wawira Gichoya[1], and Hari Trivedi[1,†]

[1]HITI Lab, Emory University, Atlanta, GA, USA
[2]Lunit, Seoul, South Korea
[3]Clemson University, Clemson, SC, USA
[4]Emory University, Atlanta, GA, USA

[*]Equal Contributions
[†]Corresponding Author: hari.trivedi@emory.edu

Abstract

AI models show potential to improve breast cancer screening, however detailed subgroup evaluations to uncover the strengths and weaknesses of models are lacking. This study presents a granular evaluation of a commercial AI model for cancer detection on digital breast tomosynthesis (DBT) on a retrospective cohort of 167,860 screening exams in female patients. Performance in distinguishing screen detected cancers (1,368 exams) from negative exams (166,387 exams) is stratified across demographic, imaging, and pathologic subgroups to identify disparities. The overall AUROC is 0.91 and sensitivity is 0.73 with robust performance across demographics. In-situ cancers (AUROC: 0.85, sensitivity: 0.55), calcifications (AUROC: 0.80, sensitivity: 0.66), and dense breast tissue (AUROC: 0.88, sensitivity: 0.63) are associated with lower performance, while masses (AUROC: 0.93, sensitivity: 0.85) and architectural distortions (AUROC: 0.90, sensitivity: 0.83) are associated with higher performance. These results highlight the need for detailed evaluations and vigilance in adopting new clinical tools.

1 Introduction

Breast cancer is the most common type of cancer in women and causes 42,000 deaths each year in the United States [1]. Screening mammography has historically been associated with a 38–48% reduction in breast cancer mortality through early detection of abnormalities and enabling earlier treatment [2–4], although recent analyses of contemporary observational data suggest that the observed protective effects may not be solely attributable to screening itself [5]. The Breast Imaging Reporting and Data System (BI-RADS), developed by the American College of Radiology, is a standardized framework used to categorize breast imaging findings and guide management. BI-RADS includes both an overall assessment score (ranging from 0 to 6) and specific lesion descriptors such as mass shape, margin, calcification morphology, and asymmetry type. Each BI-RADS category corresponds to a recommended clinical action: for example, Category 1 (negative) and 2 (benign) require routine screening, Category 3 (probably benign) suggests short-interval follow-up, and Categories 4 and 5 (suspicious and highly suggestive of malignancy) prompt biopsy [6]. Despite this, the sensitivity and specificity of 2D screening mammography remain around 86.9% and 88.9% resulting in unnecessary recalls and biopsies, as well as missing cancers [7]. Earlier data from the Breast Cancer Detection Demonstration Project (BCDDP) also showed age-related sensitivities of 90% for women under 50 and 95% for women 50 and older [8].

Over the past two decades, digital breast tomosynthesis (DBT) has emerged as an advanced imaging modality designed to address the limitations of 2D mammography. DBT provides pseudo-three-dimensional imaging of the breast by reconstructing multiple 2D projections into image slices and can improve specificity by reducing the masking effect of overlapping tissues [9, 10]. These advantages have led to widespread adoption of the technology since its FDA approval in 2011 [11], however research has demonstrated weaknesses in some areas compared to conventional mammography. These include the increased interpretation times of DBT images – often up to twice as long as 2D mammography–and their potential for reduced visibility of calcifications, a key indicator for certain types of breast cancer [12–14].

Over the past several years, multiple deep learning models have been developed both in the research and commercial settings for diagnosis of breast cancer on 2D mammography. A recent meta-analysis demonstrated AI model performance for 2D mammography outperformed radiologists (AUC: 0.87 vs 0.81, P = 0.002) in reader studies [15]. However, reader studies are difficult to translate to real-world experience due to a manufactured reading environment and upsampling of abnormal cases. In purely retrospective studies, AI model performance for 2D mammography was not shown to be better than radiologists (AUC: 0.89 vs 0.96, P = 0.152) [15]. For DBT, three reader studies and one retrospective study demonstrated generally higher area under the receiver operating characteristics curve (AUC) for AI compared to radiologists, however with wide confidence intervals [15–19].

Little work currently exists that studies the performance of breast AI models across various subgroups of patient demographics, imaging characteristics, and pathology subtypes. For AI to gain a foothold in clinical care, clinicians must be able to understand conditions under which models can fail [20–24]. Broad interpretation of model performance across all cases can be misleading, since a model that detects the same cancers as a radiologist is not clinically meaningful. Rather, the radiologist and AI model should augment one another with complementary strengths to yield a net positive effect in detection and reduction of false positives [25–27].

In this study, we evaluate a commercial AI model for breast cancer detection on digital breast tomosynthesis (DBT) on a large retrospective cohort of 167,860 screening mammography exams from 61,332 women [28], and compare its performance to previously published metrics as well as its FDA validation study. Similar to prior work, we examine model performance across demographics, including race, ethnicity, and age groups. Additionally, performance is considered across clinically meaningful subtypes of cancer (invasive, DCIS) and subtypes of negative exams (i.e. negative at screening, negative on diagnostic, negative on biopsy). Further, we evaluate performance across finding types (mass, calcification, architectural distortion, and asymmetry) and across specific BI-RADS imaging descriptors.

2 Results

A total of 166,387 negative exams and 1,368 positive exams were included in the final evaluation of the model (Figure 1). A summary of the distribution is shown in Table 1. The majority of exams were for Black (77,306, 46.1%) and White patients (72,498, 43.2%), with 653 (47.7%) and 619 (45.2%) screen-detected cancers, respectively. Non-Hispanic or Latino patients (139,088, 82.9%) represented the major ethnicity. Breast density was distributed as A (18,282, 11.0%), B (68,861, 41.3%), C (70,438, 42.2%), and D (9229, 5.5%).

2.1 Overall Model Performance

The model achieved an overall AUC of 0.91 (95% CI: 0.90–0.92) in distinguishing between screen-detected cancers and negative exams, with a recall of 0.73, FPR of 0.07, TNR of 0.93, and FNR of 0.27 (Table 2). Performance was not statistically different across Race, Ethnicity, or Age. In addition, model performance was evaluated across multiple subgroups including screening outcome, breast density, imaging features, and pathology subtype.

2.2 Model Trends by Outcome Subgroups

We observed that model prediction scores trended slightly higher between screen negative, diagnostic negative, and confirmed benign exams, with decreasing percentages of exams being classified as normal according

to the operating point of 0.1 (Figure 2). The model classified 93% (137,240/146,903) of screening studies as negative, 90% (14,058/15,557) of diagnostic negative exams as negative, and 81% (3,191/3,927) biopsy proven benign exams as negative. The model classified 73% (1,002/ 1,368) of screen-detected cancers as positive. Finally, the model classified 31% (33/105) of interval cancers as positive.

2.3 Model Trends by Pathological Subgroups

For binary classification, we considered invasive cancers and Ductal Carcinoma In Situ (DCIS) as positive, and all other cases as negative (screen negative, diagnostic negative, benign, borderline, and high-risk lesions). We observed that model prediction scores trended higher as pathology subtypes became more abnormal or at higher risk of becoming cancer. At the operating point of 0.1, the model correctly classified 84% (2,634/3,132) of screening studies with a benign lesion, 74% (26/35) of studies with a borderline lesion and 70% (531/760) of studies with a high-risk lesion as negative. For screening studies with DCIS and invasive cancer outcomes, the model correctly identified 55% (248/454) of DCIS and 82% (754/914) of invasive cancers (Figure 3).

Performance was generally good across invasive cancer subtypes but trended lower for certain rare subtypes. Invasive Ductal Carcinoma, Not Otherwise Specified (IDC NOS) represented the majority of cases (544/734), and had 86% (468/544) of cases classified as positive. Invasive Lobular Carcinoma (ILC) was classified correctly in 82% (82/100) of cases and Invasive Mammary Carcinoma (IMC) in 84% (56/67) of cases (Figure 4). Note that the aforementioned counts consider the number of invasive pathologies present, which differs from the total number of screen detected invasive cancers exams as some contain multiple invasive pathologies.

Model performance was significantly worse for DCIS (AUC 0.85, 95% CI: 0.83-0.87; Recall 0.55, 95% CI: 0.50-0.59; n=454; $p<0.001$) compared to invasive cancers (AUC 0.94, 95% CI: 0.93–0.95; Recall 0.83, 95% CI: 0.80-0.85; n=914; $p<0.001$) across all DCIS sub-types and grades (Figure 5). Ductal Carcinoma In Situ, Not Otherwise Specified (DCIS NOS) represented the majority of cases (201/361) and had 55% (111/201) of cases classified as positive. Note that the aforementioned counts consider the number of DCIS pathologies present which will differ than the total number of screen detected DCIS, as some exams contained multiple DCIS pathologies.

2.4 Model Trends by Imaging Features

Model performance varied based on the presence of specific findings in screening studies (Figure 6). Because exams could contain multiple findings and model results are returned at the exam level, our results are aggregated for the presence of any imaging finding at the study level. For example, if an exam has both a mass and calcification, the model output is attributed to both the mass and calcification subgroups. 84% (17,533) of abnormal screening exams had only one BI-RADS 0 lesion, while 16% (3319) of abnormal exams contained more than one. For exams containing a mass, the model correctly classified 85% (2368 of 2780) of cases in the screen negative group, 90% (2484 of 2766) in the diagnostic negative group, 84% (582 of 695) in the confirmed benign group, and 85% (196 of 231) in the screen-detected cancer group. For studies with asymmetry, the model correctly classified 92% (2799 of 3057) in the screen negative group, 91% (7992 of 8737) in the diagnostic negative group, 83% (1206 of 1449) in the confirmed benign group, and 78% (414 of 532) in the screen-detected cancer group. For studies with architectural distortion, the model correctly classified 72% (83 of 116) of cases in the screen negative group, 87% (1245 of 1431) in the diagnostic negative group, 58% (204 of 352) in the confirmed benign group, and 83% (177 of 213) in the screen-detected cancer group. For studies with calcification, the model correctly classified 80% (1735 of 2173) of cases in the screen negative group, 85% (1873 of 2196) in the diagnostic negative group, 79% (1381 of 1741) in the confirmed benign group, and 65% (408 of 623) in the screen-detected cancer group. Outputs for imaging features were further stratified by BI-RADS descriptors such as mass shapes and margins (Supplementary Figures S1 and S2), calcification distribution and morphologies (Supplementary Figures S3, S4, and S5), and architectural distortion and asymmetry subtypes (Supplementary Figures S6 and S7).

Qualitatively, exams with masses showed the best separation between model prediction scores of positive and negative exams whereas exams with architectural distortion showed the least separation. Examples of true positive, false positive, and false negative model predictions by lesion type are shown in Figure 7.

3 Discussion

This study presents a performance evaluation of a commercial AI model for breast cancer detection across demographic, pathologic, and imaging subtypes. Our results demonstrate that overall model performance was good with an AUC of 0.91 (95% CI: 0.90-0.92), similar to the FDA clearance metrics of 0.93 (95% CI: 0.92-0.94) [29]. The overall recall in our study was 0.73, which is lower than the 0.88 reported in a prior study [30]. Model performance was robust across demographic subtypes of age, race, and ethnicity. We observed a trend of decreased performance in dense breasts (density C) (AUC: 0.90, 95% CI: 0.88-0.91) and extremely dense breasts (density D) (AUC: 0.87, 95% CI: 0.82-0.92) but not a significant difference. This is somewhat expected as dense breasts can mask abnormalities, even with the separation of tissues via DBT.

When evaluating performance across imaging findings, we found that model performance was robust for exams with masses and architectural distortions, however was statistically significantly worse in exams with calcifications (AUC: 0.80, 95% CI: 0.78-0.82). Many studies have suggested that DBT suffers from decreased sensitivity for microcalcifications [31], possibly related to decreased resolution of DBT and/or the visibility of microcalcifications on only a small number of tomosynthesis slices. Therefore, decreased model performance for calcifications could be considered as a limitation of DBT modality rather than the model. We also observed that architectural distortion yielded a higher number of false positives than other imaging findings. Architectural distortions are routinely biopsied unless associated with a known post-surgical scar. These pathologies often indicate benign radial scars after biopsy. We hypothesize that this, combined with their relative rarity and disproportionate association with malignancy compared to other finding types, may contribute to their higher false positive rate. A unique feature of the BI-RADS lexicon is the various subtypes of lesions that are described; for example, a mass is described by its shape, margins and density. When evaluating performance by these features, we observe further interesting trends such as the propensity for malignant oval or circumscribed masses to have lower scores and benign irregular masses to have higher scores. We also observe that malignant exams with grouped, linear, and clustered calcifications tend to have lower scores than regional and segmental calcifications.

When evaluating across pathologic outcomes, we observed a statistically significant decrease in model performance for DCIS compared to invasive cancers (AUC: 0.85, n=454 vs 0.94, n=914; $p<0.001$). Further, the recall for DCIS was only 0.55 indicating that the model misses nearly half of DCIS cases. We hypothesize that this is due to the tendency of DCIS to present as micro-calcifications. Nevertheless, this highlights the need for radiologists to understand the differences between classification metrics such as AUC, recall, and precision, as a misunderstanding of a high AUC score can lead to unintentional automation bias. Our failure analysis (Supplementary Table S1) reveals that 44% of missed cancers were invasive, 56% were DCIS, and 11%, 61%, 36%, 10% contained masses, calcifications, asymmetries, and architectural distortions, respectively. We also noted that subtypes of IDC, such as tubular, papillary, and mucinous carcinoma had higher rates of false negatives.

By examining model prediction scores at the exam level, we can detect trends that provide additional insight. For example, the number of false positives increased across screen negative, diagnostic negative, and biopsy-proven benign exams. This is not unexpected, as the degree of visualized abnormality increases between these exam types and suggests that the model is susceptible to similar false positive errors as humans. Conversely, the high number of true negatives for diagnostic negative and biopsy-proven benign cases (both of which are deemed abnormal on screening by a radiologist), suggests that the model could be beneficial in decreasing false positive recalls and biopsies. Further work is needed to assess whether a higher operating point could result in a meaningful decrease in false positive recalls and biopsies. Lastly, the model identified 31% (33/105) interval cancers suggesting that the model was able to detect features that were either missed by the radiologist or undetectable at the time of screening. The real-world implications of this are difficult to interpret given that retrospective evaluation of interval cancers remains highly variable among radiologists [32,33]. Further evaluation of the value of AI for interval cancers would likely require a either a reader study with re-interpretation of exams with and without AI, or long-term prospective deployment to observe for any decrease in interval cancer rates.

In conclusion, breast cancer, like most diseases, is diverse in its presentation, imaging features, demographic distribution, and severity. Artificially binarizing disease into positive and negative classes without considering clinically meaningful subtypes masks true model performance, as no two cancers are alike. When considering permutations of race, ethnicity, breast density, cancer subtype, and four main imaging findings,

our test set contained 1368 cancers with 420 unique permutations of findings. If adding various subtypes of masses and calcifications, nearly every cancer would be unique. Because of the large variety of presentations, lack of understanding of a models' strengths and weaknesses can ultimately lead under- or overreliance on models, both of which are detrimental to patient care. Our results demonstrate that model performance was robust across demographics, however, was decreased for DCIS, calcifications, and dense breasts. This highlights the need for continued vigilance in evaluation and trust of AI model performance to drive adoption and realize the promised performance gains of AI.

3.1 Limitations

This study has several limitations. This is a single institution study with exams obtained across four sites. As a tertiary referral center, Emory has a slightly higher prevalence of cancer cases compared to the general population, which can affect performance metrics such as the false positive rate (FPR), true negative rate (TNR), and false negative rate (FNR), as these metrics are influenced by disease prevalence. Our exams are exclusively from Hologic devices, so results may not be applicable to other scanners. All DBT exams are part of ComboHD exams, meaning that a full-field digital mammogram (FFDM), DBT, and synthetic 2D exam are obtained for each patient. Therefore, it cannot be certain that each lesion was detectable on DBT. Model results were provided at the exam-level, with only the final model score and the most-severe pathologic label considered for each exam. As a result, we did not analyze model performance at the image- and lesion-level, or whether the INSIGHT DBT model was able to accurately localize lesions. Image finding types used in our analysis were determined based on the radiologist-reported finding at screening. Due to this, the results of the finding type analyses are subject to inter-observer variability and imaging findings on screening may differ from the follow-up diagnostic exams. Additionally, we did not evaluate model performance on patients with breast implants as the EMBED dataset currently does not reliably contain this information. Work is ongoing to identify images with visible implants. The FDA clearance for the model did not specifically address exams with breast implants, however the developer advises caution in interpreting results in exams with implants. Lastly, this model, like most other commercial AI models interprets images from only the current exam and therefore does not benefit from comparison with priors. It is possible that some lesions were apparent to the radiologist only when comparing to prior exams. Future work will include failure analysis to assess whether lesions were visible on DBT alone and whether availability of priors affects radiologist performance.

4 Methods

4.1 Ethics Statement

This study was approved by the Institutional Review Board of Emory University (IRB STUDY00006509). Informed consent was waived for the use of anonymized, retrospective data.

4.2 Model

The INSIGHT DBT (engine version v.1.3.0; product version v.1.1.2.0) model from Lunit, Inc (Seoul, South Korea) was used for inference in the study. The model received FDA 510k clearance in 2023 for use on screening exams from Hologic and GE scanners and is deployed across 16 countries and 44 sites. The model architecture has 23.4 million trainable parameters and consists of a ResNet34 feature extractor and an aggregation module that outputs an image-level malignancy score and a heatmap for localization. The analysis was conducted using exam-level scores which were derived by selecting the highest predicted image malignancy score for each exam. It was trained on approximately 160,000 exams from 2 countries, and 29% were cancers. 87% of the exams were from the USA and 13% from South Korea. 21% of the exams were labeled normal, 50% were benign and 29% were cancer. Of the exams labeled cancer, 78% were invasive. 47% of the exams had a BI-RADS breast density of A or B, rest 53% had BI-RADS C or D. 4% of the patients were Black, 9% Hispanic, 85% White. 75% of the exams had patient age between 40 and 79, both inclusive. The model has a set operating threshold of 0.1 which was selected by Lunit to balance sensitivity and specificity across various populations and cannot be adjusted after regulatory clearance.

4.3 Data

This analysis was conducted using digital breast tomosynthesis (DBT) images from the Emory Breast Imaging Dataset (EMBED) [28]. The relevant images were collected from screening exams in four hospitals in the Emory Healthcare network between January 2013 and December 2020. Women with both average and above-average breast cancer risk were included in the study as the INSIGHT DBT model does not specify an eligible population. Within EMBED, approximately 72.1 percent of screening exams are ComboHD studies that contain both 2D and DBT + synthetic 2D images, whereas the remainder are 2D only. For this analysis, exams that were 2D-only were excluded. No exams contain DBT or C-view only. 99.8 percent of exams are from Hologic scanners.

In total, 229,207 exams for 80,331 patients were evaluated by the model, with 1 exam failing to process. The failure was due to issues in reading or transforming the image data, stemming from missing or malformed DICOM information. Predicted abnormality scores were aggregated to the exam-level by taking the maximum score for each. 167,860 exams for 61,332 patients had unambiguous ground truth data and follow-up available and were considered in the final analysis after excluding 61,368 exams during the label assignment and data handling steps shown in Figure 1.

4.4 Ground Truthing

The original radiology report, inclusive of any addendums, was used as ground truth for imaging features and BI-RADS characteristics. Pathology ground truth was obtained from reports and summarized into six categories based on prior consensus from an expert panel of breast pathologists, oncologists, and radiologists [28]: invasive cancers, DCIS, high-risk lesions, borderline lesions, benign lesions, and non-breast cancers. The most common pathologies assigned to each category are listed in Supplementary Table S2. Demographics were obtained from the EHR.

4.5 Label Assignment

Exams were categorized using a comprehensive set of clinical conditions to determine ground truth labels for evaluation. As all images used in the evaluation were captured at screening, any follow-up diagnostic assessments (recall studies) were matched to their preceding exams. Diagnostic exams for problem evaluations were not considered, except in the case of interval cancers as described below. Descriptions of the five exam labels considered in the analysis are shown in Table 3.

The primary clinical features for label assignment were BI-RADS assessments and pathology results. Because EMBED contains data at the finding level (i.e. a screening exam may have multiple findings in each breast), finding-level assessments were aggregated to the exam-level by selecting the most severe BI-RADS assessment and pathology results from each exam to match the aggregated outputs from the INSIGHT DBT model.

4.5.1 Negative Exams

Three types of 'negative' exams were defined: screen negatives (SN), diagnostic negatives (DN), and biopsy-proven benign (B). Screen negatives were defined as screening exams with only BI-RADS 1 or 2 findings that did not fit the interval cancer definition. Diagnostic negatives were defined as screening exams with at least one BI-RADS 0 finding that prompted a diagnostic follow-up which only had BI-RADS 1, 2, or 3 findings. Biopsy-proven benign were defined as abnormal screening exams that prompted a diagnostic follow-up with a BI-RADS 4 or 5 finding which was biopsied and resulted in a high-risk, borderline, or benign lesion. High-risk lesions that were resected and did not result in a cancer were considered negative; those that were resected and yielded a cancer were considered in the positive group (see below). Non-breast cancers like lymphoma and other primary source of cancers were excluded in this evaluation.

4.5.2 Positive Exams

Positive exams were classified as screen-detected cancer detected via the normal screening pathway: an abnormal screening exam that prompted a diagnostic follow-up with a BI-RADS 4 or 5 finding followed

by an invasive cancer or DCIS on biopsy or resection. This included high-risk lesions on biopsy that were resected and found to be malignant.

4.5.3 Interval Cancers

Interval cancers are defined as a negative screening exam followed by a confirmed cancer diagnosis within 12 months of the screening exam. These were considered as a separate class during evaluation due to the difficulty in categorizing them as either negative or positive. Some interval cancers may be visible on preceding study in retrospect while others were not. However, previous work has shown that AI models were able to detect at least some of these cancers [34–36].

4.5.4 Excluded/Invalid Exams

Because EMBED relies on human entered data, there are rare circumstances of idiosyncratic or missing data. This includes missing or invalid BI-RADS scores based on exam type, abnormal screening studies with no diagnostic follow-up, or BI-RADS 4 and 5 diagnostic studies with no biopsy results. There are also instances in which a patient leaves the healthcare system and therefore cannot be accurately ground-truthed.

Considering this, exams were excluded from the analysis for several reasons. These included: exams with incorrect data, such as screening exams with diagnostic-only BI-RADS assessments or invalid exam accessions; exams with missing data, such as abnormal screening exams with no diagnostic follow-up, screen negatives with no long-term follow-up to confirm their ground truth, or diagnostic BI-RADS 4/5 cases with no biopsy results; exams falling outside the study cohort of interest, such as diagnostic exams and exams containing previously-known breast cancers (assigned a BI-RADS 6) or non-breast cancers. Numbers and types of excluded cases are shown in Figure 1.

4.6 Analysis

4.6.1 Feature Selection for Subgroup Analysis

Model performance was assessed across various demographic, imaging, and pathologic subgroups. Demographic subgroups were considered for patient race, ethnicity, and age. The analysis included the patient race categories: Asian, Black, White, unknown or unavailable, and other patient races. The ethnicity groups considered were: Hispanic or Latino, not Hispanic or Latino, and unknown. Patient age at the time of exam was binned into three groups: under-50, 50-to-75, and over-75.

Imaging features like exam-level presence of masses, asymmetries, architectural distortions, and calcifications were also considered during analysis. Outputs for imaging features stratified by BI-RADS descriptors [6] are included in the supplement for mass shapes and margins (Supplementary Figures S1 and S2), calcification distribution and morphologies (Supplementary Figures S3, S4, and S5), and architectural distortion and asymmetry subtypes (Supplementary Figures S6 and S7).

Lastly, pathological subgroups were defined based on the final diagnosis determined from biopsy or, when available, subsequent surgical resection. For example, if a patient had an abnormal screening exam followed by diagnostic mammogram and biopsy that resulted in ductal carcinoma in situ (DCIS), followed by resection resulting in invasive ductal carcinoma (IDC), the diagnosis attributed to the screening exam would be IDC. Any subsequent post-treatment diagnoses or recurrences were not considered. Results were stratified by major pathology types: invasive cancers, DCIS, high-risk lesions, borderline lesions, and benign lesions/normal tissue. Invasive cancers and DCIS were further decomposed into common subtypes.

4.6.2 Statistics & Reproducibility

Using the ground truth categories determined during label assignment, the model was evaluated as a binary classifier with screen-detected cancers as the positive class (assigned a numeric label of 1) and the negative groups as the negative class (assigned a numeric label of 0). Interval cancers were not considered during metric calculation. Threshold-dependent metrics were evaluated at the commercial operating point for Insight DBT of 0.10. Due to the heavy class imbalance of the data towards then negative class, performance was primarily evaluated with precision, recall, false positive rate (FPR), true negative rate (TNR), false negative rate (FNR), and area under the receiver operating characteristic curve (AUROC).

Confidence intervals for all classification metrics were estimated using non-parametric bootstrapping (5,000 resamples). Statistical comparisons between subgroups were conducted using the Kruskal-Wallis test to bootstrapped AUC distributions to assess for global differences. When significant heterogeneity was observed ($p < 0.05$), post-hoc pairwise comparisons between subgroups were performed using bootstrapped difference tests. Where multiple group comparisons were performed, the Bonferroni correction was applied to control the family-wise type 1 error rate at 5%. When comparing AUC between invasive cancers and DCIS, both groups were compared to the same set of negative cases. Results were described as statistically significant if they met the pre-defined cutoff after correcting for multiple comparisons. All statistical analyses were conducted using Python. No statistical method was used to predetermine sample size since this was a retrospective study.

5 Data Availability

The clinical data used in this study, the Emory Breast Imaging Dataset (EMBED), is not available due to medical institutional data policies. 20% of the dataset is available under restricted access through the AWS Open Data program for non-commercial research use. Access can be obtained by submitting a request through an online form. All use of EMBED Open Data is subject to the data use agreement (Available: `https://github.com/Emory-HITI/EMBED_Open_Data/blob/main/EMBED_license.md`). Lunit INSIGHT DBT model outputs are considered proprietary and cannot be publicly released. Source data for tables and figures cannot be released publicly due to these restrictions on sharing EMBED data and Lunit INSIGHT DBT model outputs.

6 Code Availability

The code used to perform the data processing and label assignment for this study is hosted on Github (Available: Emory-HITI/Lunit-Model-Evaluation) [37].

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2020," *CA: A Cancer Journal for Clinicians*, vol. 70, no. 1, pp. 7–30, 2020.

[2] P. Gøtzsche and K. Jørgensen, "Screening for breast cancer with mammography," *Cochrane Database of Systematic Reviews*, vol. 2013, no. 6, 2013.

[3] M. Marmot, D. Altman, D. Cameron, J. Dewar, S. Thompson, M. Wilcox, and I. U. P. o. B. C. S. The, "The benefits and harms of breast cancer screening: An independent review," *British Journal of Cancer*, vol. 108, no. 11, pp. 2205–2240, 2013.

[4] M. Broeders, S. Moss, L. Nyström, S. Njor, H. Jonsson, E. Paap, N. Massat, S. Duffy, E. Lynge, and E. Paci, "The Impact of Mammographic Screening on Breast Cancer Mortality in Europe: A Review of Observational Studies," *Journal of Medical Screening*, vol. 19, pp. 14–25, Sept. 2012.

[5] P. Autier, K. J. Jørgensen, M. Smans, and H. Støvring, "Effect of screening mammography on the risk of breast cancer deaths and of all-cause deaths: A systematic review with meta-analysis of cohort studies," *Journal of Clinical Epidemiology*, vol. 172, Aug. 2024.

[6] ACR, *2013 ACR BI-RADS Atlas: Breast Imaging Reporting and Data System*. American College of Radiology, Jan. 2014.

[7] C. D. Lehman, R. F. Arao, B. L. Sprague, J. M. Lee, D. S. M. Buist, K. Kerlikowske, L. M. Henderson, T. Onega, A. N. A. Tosteson, G. H. Rauscher, and D. L. Miglioretti, "National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium," *Radiology*, vol. 283, pp. 49–58, Apr. 2017.

[8] M. P. Cunningham, "The Breast Cancer Detection Demonstration Project 25 years later," *CA: a cancer journal for clinicians*, vol. 47, pp. 131–133, May 1997.

[9] N. Houssami and P. Skaane, "Overview of the evidence on digital breast tomosynthesis in breast cancer detection," *The Breast*, vol. 22, pp. 101–108, Apr. 2013.

[10] E. Dhamija, M. Gulati, S. V. S. Deo, A. Gogia, and S. Hari, "Digital Breast Tomosynthesis: An Overview," *Indian Journal of Surgical Oncology*, vol. 12, pp. 315–329, June 2021.

[11] I. B. Richman, J. R. Hoag, X. Xu, H. P. Forman, R. Hooley, S. H. Busch, and C. P. Gross, "Adoption of Digital Breast Tomosynthesis in Clinical Practice," *JAMA Internal Medicine*, vol. 179, pp. 1292–1295, Sept. 2019.

[12] S. Astley, S. Connor, Y. Lim, C. Tate, H. Entwistle, J. Morris, S. Whiteside, J. Sergeant, M. Wilson, U. Beetles, C. Boggis, and F. Gilbert, "A comparison of image interpretation times in full field digital mammography and digital breast tomosynthesis," in *Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment*, vol. 8673, pp. 189–196, SPIE, Mar. 2013.

[13] M. L. Zuley, A. I. Bandos, G. S. Abrams, C. Cohen, C. M. Hakim, J. H. Sumkin, J. Drescher, H. E. Rockette, and D. Gur, "Time to Diagnosis and Performance Levels during Repeat Interpretations of Digital Breast Tomosynthesis: Preliminary Observations," *Academic Radiology*, vol. 17, pp. 450–455, Apr. 2010.

[14] Y. Chen, E. S. Sudin, G. J. Partridge, A. G. Taib, I. T. Darker, P. Phillips, J. J. James, K. Satchithananda, N. Sharma, and M. J. Michell, "Measuring reader fatigue in the interpretation of screening digital breast tomosynthesis (DBT)," *British Journal of Radiology*, vol. 96, p. 20220629, Mar. 2023.

[15] J. H. Yoon, F. Strand, P. A. T. Baltzer, E. F. Conant, F. J. Gilbert, C. D. Lehman, E. A. Morris, L. A. Mullen, R. M. Nishikawa, N. Sharma, I. Vejborg, L. Moy, and R. M. Mann, "Standalone AI for Breast Cancer Detection at Screening Digital Mammography and Digital Breast Tomosynthesis: A Systematic Review and Meta-Analysis," *Radiology*, vol. 307, p. e222639, June 2023.

[16] M. C. Pinto, A. Rodriguez-Ruiz, K. Pedersen, S. Hofvind, J. Wicklein, S. Kappler, R. M. Mann, and I. Sechopoulos, "Impact of Artificial Intelligence Decision Support Using Deep Learning on Breast Cancer Screening Interpretation with Single-View Wide-Angle Digital Breast Tomosynthesis," *Radiology*, vol. 300, pp. 529–536, Sept. 2021.

[17] S. Romero-Martín, E. Elías-Cabot, J. L. Raya-Povedano, A. Gubern-Mérida, A. Rodríguez-Ruiz, and M. Álvarez-Benito, "Stand-Alone Use of Artificial Intelligence for Digital Mammography and Digital Breast Tomosynthesis Screening: A Retrospective Evaluation," *Radiology*, vol. 302, pp. 535–542, Mar. 2022.

[18] Y. Shoshan, R. Bakalo, F. Gilboa-Solomon, V. Ratner, E. Barkan, M. Ozery-Flato, M. Amit, D. Khapun, E. B. Ambinder, E. T. Oluyemi, B. Panigrahi, P. A. DiCarlo, M. Rosen-Zvi, and L. A. Mullen, "Artificial Intelligence for Reducing Workload in Breast Cancer Screening with Digital Breast Tomosynthesis," *Radiology*, vol. 303, pp. 69–77, Apr. 2022.

[19] E. F. Conant, A. Y. Toledano, S. Periaswamy, S. V. Fotin, J. Go, J. E. Boatsman, and J. W. Hoffmeister, "Improving Accuracy and Efficiency with Concurrent Use of Artificial Intelligence for Digital Breast Tomosynthesis," *Radiology: Artificial Intelligence*, vol. 1, p. e180096, July 2019.

[20] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, and K. Tsaneva-Atanasova, "Artificial intelligence, bias and clinical safety," *BMJ Quality & Safety*, vol. 28, pp. 231–237, Mar. 2019.

[21] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, pp. e745–e750, Nov. 2021.

[22] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, p. 195, Oct. 2019.

[23] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," *Nature Medicine*, vol. 28, pp. 31–38, Jan. 2022.

[24] S. Purkayastha, H. Trivedi, and J. W. Gichoya, "Failures Hiding in Success for Artificial Intelligence in Radiology," *Journal of the American College of Radiology*, vol. 18, pp. 517–519, Mar. 2021.

[25] K. Lång, V. Josefsson, A.-M. Larsson, S. Larsson, C. Högberg, H. Sartor, S. Hofvind, I. Andersson, and A. Rosso, "Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): A clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study," *The Lancet. Oncology*, vol. 24, pp. 936–944, Aug. 2023.

[26] V. Hernström, V. Josefsson, H. Sartor, D. Schmidt, A.-M. Larsson, S. Hofvind, I. Andersson, A. Rosso, O. Hagberg, and K. Lång, "Screening performance and characteristics of breast cancer detected in the Mammography Screening with Artificial Intelligence trial (MASAI): A randomised, controlled, parallel-group, non-inferiority, single-blinded, screening accuracy study," *The Lancet. Digital Health*, pp. S2589–7500(24)00267–X, Feb. 2025.

[27] K. Dembrower, A. Crippa, E. Colón, M. Eklund, and F. Strand, "Artificial intelligence for breast cancer detection in screening mammography in Sweden: A prospective, population-based, paired-reader, non-inferiority study," *The Lancet Digital Health*, vol. 5, pp. e703–e711, Oct. 2023.

[28] J. J. Jeong, B. L. Vey, A. Bhimireddy, T. Kim, T. Santos, R. Correa, R. Dutt, M. Mosunjac, G. Oprea-Ilies, G. Smith, M. Woo, C. R. McAdams, M. S. Newell, I. Banerjee, J. Gichoya, and H. Trivedi, "The EMory BrEast imaging Dataset (EMBED): A Racially Diverse, Granular Dataset of 3.4 Million Screening and Diagnostic Mammographic Images," *Radiology: Artificial Intelligence*, vol. 5, p. e220047, Jan. 2023.

[29] U.S. Food and Drug Administration, Center for Devices and Radiological Health, "Lunit INSIGHT DBT v1.1 K242652 approval letter," Sept. 2024.

[30] E. K. Park, S. Kwak, W. Lee, J. S. Choi, T. Kooi, and E.-K. Kim, "Impact of AI for Digital Breast Tomosynthesis on Breast Cancer Detection and Interpretation Time," *Radiology. Artificial Intelligence*, vol. 6, p. e230318, May 2024.

[31] M. L. Spangler, M. L. Zuley, J. H. Sumkin, G. Abrams, M. A. Ganott, C. Hakim, R. Perrin, D. M. Chough, R. Shah, and D. Gur, "Detection and Classification of Calcifications on Digital Breast Tomosynthesis and 2D Digital Mammography: A Comparison," *American Journal of Roentgenology*, vol. 196, pp. 320–324, Feb. 2011.

[32] G. E. Park, B. J. Kang, S. H. Kim, and J. Lee, "Retrospective Review of Missed Cancer Detection and Its Mammography Findings with Artificial-Intelligence-Based, Computer-Aided Diagnosis," *Diagnostics*, vol. 12, p. 387, Feb. 2022.

[33] T. Hovda, S. R. Hoff, M. Larsen, L. Romundstad, K. K. Sahlberg, and S. Hofvind, "True and Missed Interval Cancer in Organized Mammographic Screening: A Retrospective Review Study of Diagnostic and Prior Screening Mammograms," *Academic Radiology*, vol. 29, pp. S180–S191, Jan. 2022.

[34] M. Nanaa, V. O. Gupta, S. E. Hickman, I. Allajbeu, N. R. Payne, O. Arponen, R. Black, Y. Huang, A. N. Priest, and F. J. Gilbert, "Accuracy of an Artificial Intelligence System for Interval Breast Cancer Detection at Screening Mammography," *Radiology*, vol. 312, p. e232303, Aug. 2024.

[35] K. Freeman, J. Geppert, C. Stinton, D. Todkill, S. Johnson, A. Clarke, and S. Taylor-Phillips, "Use of artificial intelligence for image analysis in breast cancer screening programmes: Systematic review of test accuracy," *BMJ*, vol. 374, p. n1872, Sept. 2021.

[36] D. Byng, B. Strauch, L. Gnas, C. Leibig, O. Stephan, S. Bunk, and G. Hecht, "AI-based prevention of interval cancers in a national mammography screening program," *European Journal of Radiology*, vol. 152, p. 110321, July 2022.

[37] B. Brown-Mulry and R. Isaac, "Emory-HITI/Lunit-Model-Evaluation: V1.0.0 Release." Zenodo, Sept. 2025.

7 Acknowledgments

This study was funded by Lunit, Inc. and AIM-AHEAD (Grant Number 3OT2OD032581-01S7). We also acknowledge support from the Emory University AI Image Extraction Core Facility (RRID:SCR_026693). J.W.G. declares support from NHLBI Award Number R01HL167811.

8 Author Contributions

The study was conceptualized by H.T., S.H.L., and A.S. Imaging data preparation was conducted by B.B-M. and A.M. Model inference was conducted by K.M., A.S., and S.H.L. Data engineering, analyses, and visualization were conducted by B.B-M. and R.S.I. under the supervision of H.T. and J.W.G. with input from T.D. and F.L. Statistical analysis was performed by R.S.I. with input from M.W. and B.B-M. C.A.F-R. and B.P. were responsible for clinical interpretation. The first manuscript draft was prepared by B.B-M., R.S.I., and H.T. All authors contributed to the editing of the manuscript.

9 Competing Interests

This study was funded by Lunit Inc., however all scientific evaluation and analysis was performed solely by personnel at Emory University. The authors declare no other competing interests.

10 Tables

Table 1: Patient demographics and imaging characteristics by outcome. Positive class was defined as a screen-detected cancer, negative class was defined as a negative screening exam, abnormal screening followed by a negative diagnostic, or an abnormal screening followed by a benign biopsy. Interval cancers were defined as a negative screening exam followed by a cancer within 12 months and were considered separately (n=167,860).

	Overall	Screen Negative	Diagnostic Negative	Biopsy Proven Benign	Interval Cancer	Screen Detected Cancer
n	167860	146903	15557	3927	105	1368
Race (%)						
Black	77306 (46.1)	67442 (45.9)	7100 (45.6)	2068 (52.7)	43 (41.0)	653 (47.7)
White	72498 (43.2)	64113 (43.6)	6335 (40.7)	1378 (35.1)	53 (50.5)	619 (45.2)
Asian	9044 (5.4)	7933 (5.4)	844 (5.4)	201 (5.1)	5 (4.8)	61 (4.5)
Other	1536 (0.9)	1253 (0.9)	222 (1.4)	50 (1.3)	1 (1.0)	10 (0.7)
Unknown	7476 (4.5)	6162 (4.2)	1056 (6.8)	230 (5.9)	3 (2.9)	25 (1.8)
Ethnicity (%)						
Not Hispanic or Latino	139088 (82.9)	122085 (83.1)	12489 (80.3)	3187 (81.2)	96 (91.4)	1231 (90.0)
Hispanic or Latino	4994 (3.0)	4177 (2.8)	664 (4.3)	126 (3.2)	0 (0.0)	27 (2.0)
Unknown	23778 (14.2)	20641 (14.1)	2404 (15.5)	614 (15.6)	9 (8.6)	110 (8.0)
Age (%)						
< 50	39358 (23.4)	31968 (21.8)	5810 (37.3)	1370 (34.9)	28 (26.7)	182 (13.3)
$50 - 75$	113632 (67.7)	101397 (69.0)	8805 (56.6)	2361 (60.1)	71 (67.6)	998 (73.0)
≥ 75	14870 (8.9)	13538 (9.2)	942 (6.1)	196 (5.0)	6 (5.7)	188 (13.7)
Tissue Density (%)						
A	18282 (11.0)	16842 (11.5)	1014 (6.6)	326 (8.5)	0 (0.0)	100 (7.4)
B	68861 (41.3)	60975 (41.7)	5748 (37.5)	1479 (38.4)	26 (25.0)	633 (46.8)
C	70438 (42.2)	60414 (41.3)	7594 (49.6)	1791 (46.5)	71 (68.3)	568 (42.0)
D	9229 (5.5)	7964 (5.4)	954 (6.2)	253 (6.6)	7 (6.7)	51 (3.8)
Screen Detected Pathology (%)						
No Pathology	162565 (96.8)	146903 (100.0)	15557 (100.0)	0 (0.0)	105 (100.0)	0 (0.0)
Benign Lesion	3132 (1.9)	0 (0.0)	0 (0.0)	3132 (79.8)	0 (0.0)	0 (0.0)
Borderline Lesion	35 (0.0)	0 (0.0)	0 (0.0)	35 (0.9)	0 (0.0)	0 (0.0)
High Risk Lesion	760 (0.5)	0 (0.0)	0 (0.0)	760 (19.4)	0 (0.0)	0 (0.0)
Invasive Cancer	914 (0.5)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	914 (66.8)
DCIS	454 (0.3)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	454 (33.2)
Finding Characteristics (%)						
Mass (%)	6472 (3.9)	2780 (1.9)	2766 (17.8)	695 (17.7)	0 (0.0)	231 (16.9)
Asymmetry (%)	13778 (8.2)	3057 (2.1)	8737 (56.2)	1449 (36.9)	3 (2.9)	532 (38.9)
Architectural Distortion (%)	2112 (1.3)	116 (0.1)	1431 (9.2)	352 (9.0)	0 (0.0)	213 (15.6)
Calcification (%)	6735 (4.0)	2173 (1.5)	2196 (14.1)	1741 (44.3)	2 (1.9)	623 (45.5)

Table 2: Model performance overall and stratified by demographic, imaging, and pathologic subgroups. The overall AUROC is 0.91 and sensitivity is 0.73 with robust performance across demographics. In-situ cancers, calcifications, and dense breast tissue are associated with lower performance, while masses and architectural distortions are associated with higher performance. Metrics are presented with 95% confidence intervals. Statistical comparisons used Kruskal–Wallis tests on bootstrapped AUC distributions with Bonferroni-corrected post-hoc pairwise tests. Asterisks indicate statistically significant differences (*p<0.01; **p<0.001) (n=167,755, interval cancers are not included).

Group Value	Total Negatives	Total Positives	Precision	Recall	FPR	TNR	FNR	AUC
Overall	166387 (99.18%)	1368 (0.82%)	0.08 (0.07, 0.08)	0.73 (0.71, 0.76)	0.07 (0.07, 0.07)	0.93 (0.93, 0.93)	0.27 (0.24, 0.29)	0.91 (0.90, 0.92)
Race								
Asian	8978 (99.33%)	61 (0.67%)	0.05 (0.04, 0.07)	0.69 (0.58, 0.81)	0.08 (0.08, 0.09)	0.92 (0.91, 0.92)	0.31 (0.19, 0.42)	0.88 (0.82, 0.93)
Black	76610 (99.15%)	653 (0.85%)	0.08 (0.07, 0.08)	0.71 (0.67, 0.74)	0.07 (0.07, 0.07)	0.93 (0.93, 0.93)	0.29 (0.26, 0.33)	0.91 (0.89, 0.92)
White	71826 (99.15%)	619 (0.85%)	0.09 (0.08, 0.09)	0.76 (0.73, 0.80)	0.07 (0.07, 0.07)	0.93 (0.93, 0.93)	0.24 (0.20, 0.27)	0.92 (0.91, 0.94)
Other	1525 (99.35%)	10 (0.65%)	0.06 (0.02, 0.12)	0.59 (0.25, 0.89)	0.06 (0.05, 0.07)	0.94 (0.93, 0.95)	0.41 (0.11, 0.75)	0.86 (0.74, 0.98)
Unknown	7448 (99.67%)	25 (0.33%)	0.04 (0.02, 0.06)	0.80 (0.63, 0.95)	0.07 (0.06, 0.07)	0.93 (0.93, 0.94)	0.20 (0.05, 0.37)	0.94 (0.88, 0.98)
Ethnicity								
Not Hispanic or Latino	137761 (99.11%)	1231 (0.89%)	0.08 (0.08, 0.09)	0.73 (0.71, 0.76)	0.07 (0.07, 0.07)	0.93 (0.93, 0.93)	0.27 (0.24, 0.29)	0.91 (0.90, 0.92)
Hispanic or Latino	4967 (99.46%)	27 (0.54%)	0.05 (0.03, 0.07)	0.70 (0.53, 0.88)	0.07 (0.07, 0.08)	0.93 (0.92, 0.93)	0.30 (0.12, 0.47)	0.87 (0.77, 0.95)
Unknown	23659 (99.54%)	110 (0.46%)	0.05 (0.04, 0.06)	0.75 (0.66, 0.83)	0.07 (0.06, 0.07)	0.93 (0.93, 0.94)	0.25 (0.17, 0.34)	0.93 (0.90, 0.95)
Tissue Density								
A	18182 (99.45%)	100 (0.55%)	0.08 (0.06, 0.10)	0.74 (0.65, 0.82)	0.05 (0.04, 0.05)	0.95 (0.95, 0.96)	0.26 (0.18, 0.35)	0.95 (0.92, 0.97)*
B	68202 (99.08%)	633 (0.92%)	0.09 (0.08, 0.09)	0.77 (0.73, 0.80)	0.08 (0.07, 0.08)	0.92 (0.92, 0.93)	0.23 (0.20, 0.27)	0.92 (0.91, 0.93)
C	69799 (99.19%)	568 (0.81%)	0.07 (0.07, 0.08)	0.71 (0.67, 0.75)	0.07 (0.07, 0.08)	0.93 (0.92, 0.93)	0.29 (0.25, 0.33)	0.90 (0.88, 0.91)*
D	9171 (99.45%)	51 (0.55%)	0.05 (0.03, 0.07)	0.63 (0.49, 0.76)	0.07 (0.06, 0.07)	0.93 (0.93, 0.94)	0.37 (0.24, 0.51)	0.88 (0.82, 0.92)
Age								
< 50	39148 (99.54%)	182 (0.46%)	0.06 (0.05, 0.07)	0.71 (0.64, 0.77)	0.05 (0.05, 0.05)	0.95 (0.95, 0.95)	0.29 (0.23, 0.36)	0.93 (0.91, 0.95)
$50 - 75$	112563 (99.12%)	998 (0.88%)	0.08 (0.07, 0.09)	0.72 (0.69, 0.75)	0.07 (0.07, 0.07)	0.93 (0.93, 0.93)	0.28 (0.25, 0.31)	0.91 (0.90, 0.92)
≥ 75	14676 (98.74%)	188 (1.26%)	0.08 (0.07, 0.09)	0.81 (0.76, 0.87)	0.12 (0.11, 0.12)	0.88 (0.88, 0.89)	0.19 (0.13, 0.24)	0.91 (0.89, 0.94)
Screen Detected Cancer Type								
Invasive Cancer	166387 (99.45%)	914 (0.55%)	0.06 (0.06, 0.06)	0.83 (0.80, 0.85)	0.07 (0.07, 0.07)	0.93 (0.93, 0.93)	0.17 (0.15, 0.20)	0.94 (0.93, 0.95)**
DCIS	166387 (99.73%)	454 (0.27%)	0.02 (0.02, 0.02)	0.55 (0.50, 0.59)	0.07 (0.07, 0.07)	0.93 (0.93, 0.93)	0.45 (0.41, 0.50)	0.85 (0.83, 0.87)**
Finding Characteristics								
Architectural Distortion	1899 (89.91%)	213 (10.09%)	0.33 (0.29, 0.37)	0.83 (0.78, 0.88)	0.19 (0.18, 0.21)	0.81 (0.79, 0.82)	0.17 (0.12, 0.22)	0.90 (0.88, 0.92)
Asymmetry	13243 (96.14%)	532 (3.86%)	0.25 (0.23, 0.27)	0.78 (0.74, 0.81)	0.09 (0.09, 0.10)	0.91 (0.90, 0.91)	0.22 (0.19, 0.26)	0.92 (0.90, 0.93)
Calcification	6110 (90.75%)	623 (9.25%)	0.27 (0.25, 0.29)	0.66 (0.62, 0.69)	0.18 (0.17, 0.19)	0.82 (0.81, 0.83)	0.34 (0.31, 0.38)	0.80 (0.77, 0.82)**
Mass	6241 (96.43%)	231 (3.57%)	0.20 (0.17, 0.22)	0.85 (0.80, 0.89)	0.13 (0.12, 0.14)	0.87 (0.86, 0.88)	0.15 (0.11, 0.20)	0.93 (0.91, 0.95)

Table 3: Description of the exam-level labels considered during evaluation. 'Screen-detected cancers' were the only positive class and all others were considered negative for calculation of performance metrics. Interval cancers were excluded from performance metric calculations. Screening exams were considered 'normal' if they had exclusively BI-RADS 1 or 2 findings, and 'abnormal' if they had at least one finding with a BI-RADS 0 assignment.

Label	Class	Description
Screen Negative	0	Normal screening exams with at least one follow-up exam (screening or diagnostic) after 12 months.
Diagnostic Negative	0	Abnormal screening exams that were assigned a BI-RADS of 1, 2, or 3 on a follow-up diagnostic within 6 months.
Biopsy-Proven Benign	0	Abnormal screening exams with a biopsy finding on a follow-up diagnostic exam within 6 months indicating a high-risk, borderline, or benign lesion.
Screen-Detected Cancer	1	Abnormal screening exams with a biopsy finding on a follow-up diagnostic exam within 6 months indicating an invasive or DCIS.
Interval Cancer	N/a	Normal screening exams with a biopsy finding on a follow-up diagnostic exam within 12 months indicating an invasive or DCIS cancer.

11 Figure Legends/Captions

Figure 1: Class assignment flowchart. Only the screen-detected cancer label was considered in the positive class, and all others were considered negative with the exception of interval cancers that were considered separately. Abnormal screening exams were divided into confirmed cancers, confirmed benign, and diagnostic negatives. Negative screening exams with at least one follow-up within 1-4 years were considered in the negative class. Exams that did not meet the criteria for any labels were excluded from further analysis. Definitions: HRL – high risk lesion; BLL – borderline lesion

Figure 2: Distribution of model scores across various clinical outcomes: screen negative, diagnostic negative, biopsy proven benign, screen-detected cancer and interval cancer. Model prediction scores range from 0 to 1, with the FDA-cleared model operating point set to 0.1 (dotted line). The figure illustrates the percentage of cases classified as positive or negative based on the model operating point for each clinical outcome at the exam level, with y-axis representing the distribution density of scores. The red hash marks indicate the mean score per clinical outcome. We observe that the proportion of cases classified as positive increases between Screen Negative, Diagnostic Negative, and Biopsy Proven Benign cases (n=167,860).

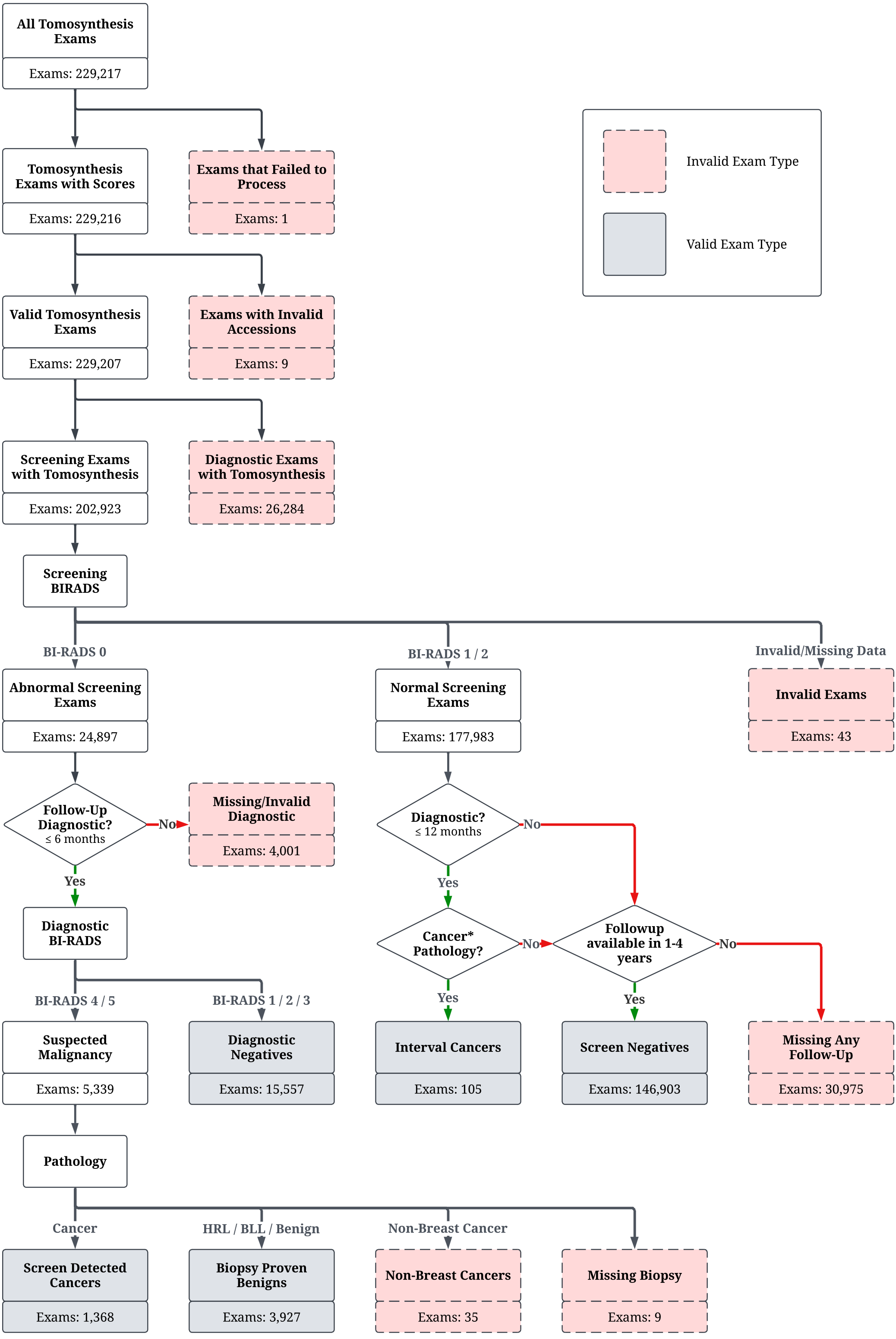
Figure 3: Distribution of model predicted scores for various pathologic outcomes. Each category represents the most severe pathology present for a given exam. Model prediction scores range from 0 to 1, with the FDA-cleared model operating point set to 0.1 (dotted line). The figure illustrates the percentage of cases classified as positive or negative based on the model operating point for each pathologic outcome, with y-axis representing the distribution density of scores. The red hash marks indicate the mean score per clinical outcome. We observe that the proportion of cases classified as positive increases with pathologic subtype severity (n=5,295).

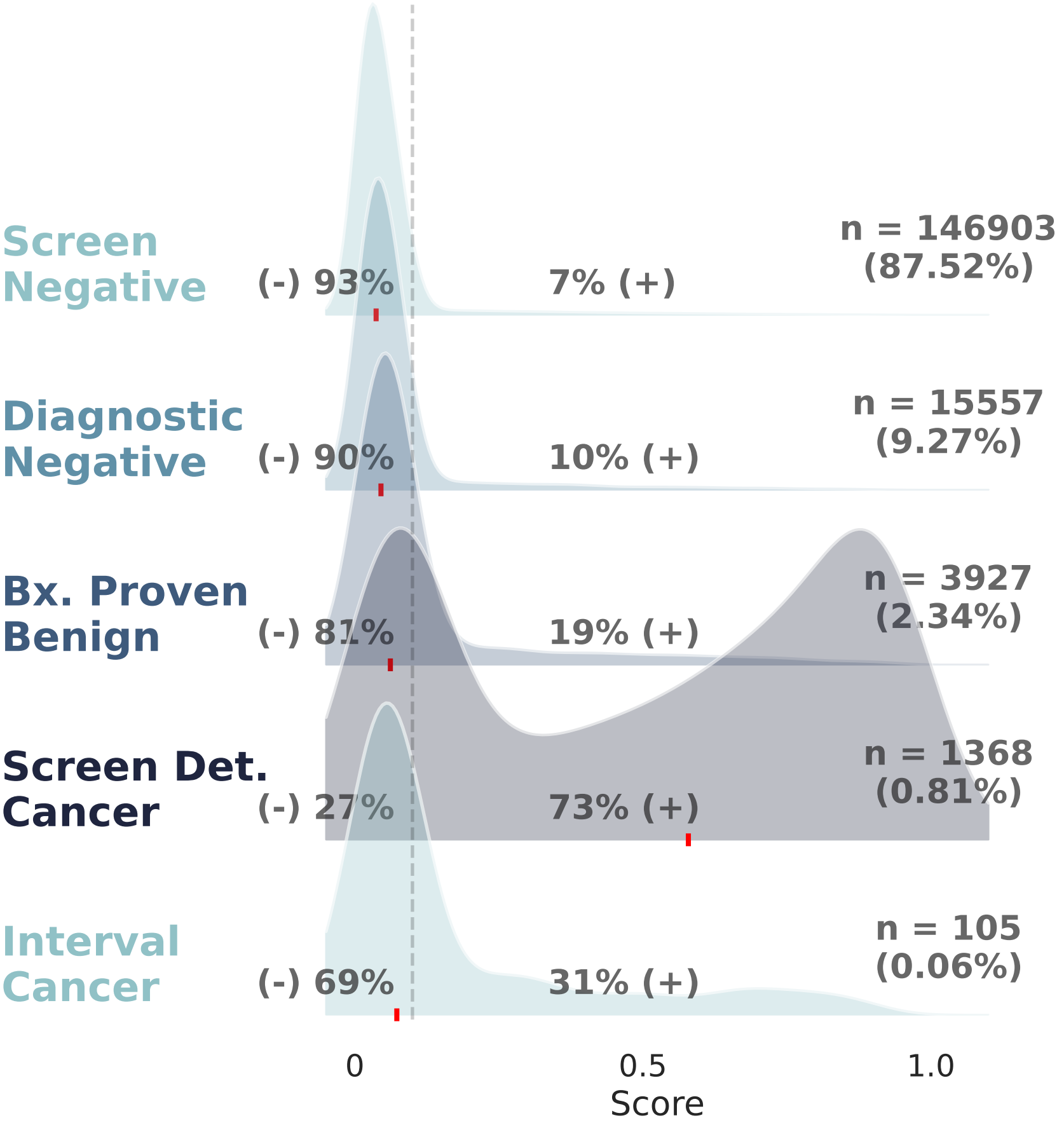
Figure 4: Distribution of model predicted scores for various common invasive cancer subtypes. The model operating point of 0.1 is marked by a horizontal line with the percentage of exams above and below the threshold indicated per pathology. IDC has many subtypes with varying imaging features and prognosis, and are represented by the adjacent panel. Model performance for invasive cancers is generally good, with lower proportion of positive cases within certain subtypes although low number of samples precludes any conclusions regarding subtypes (n=734).

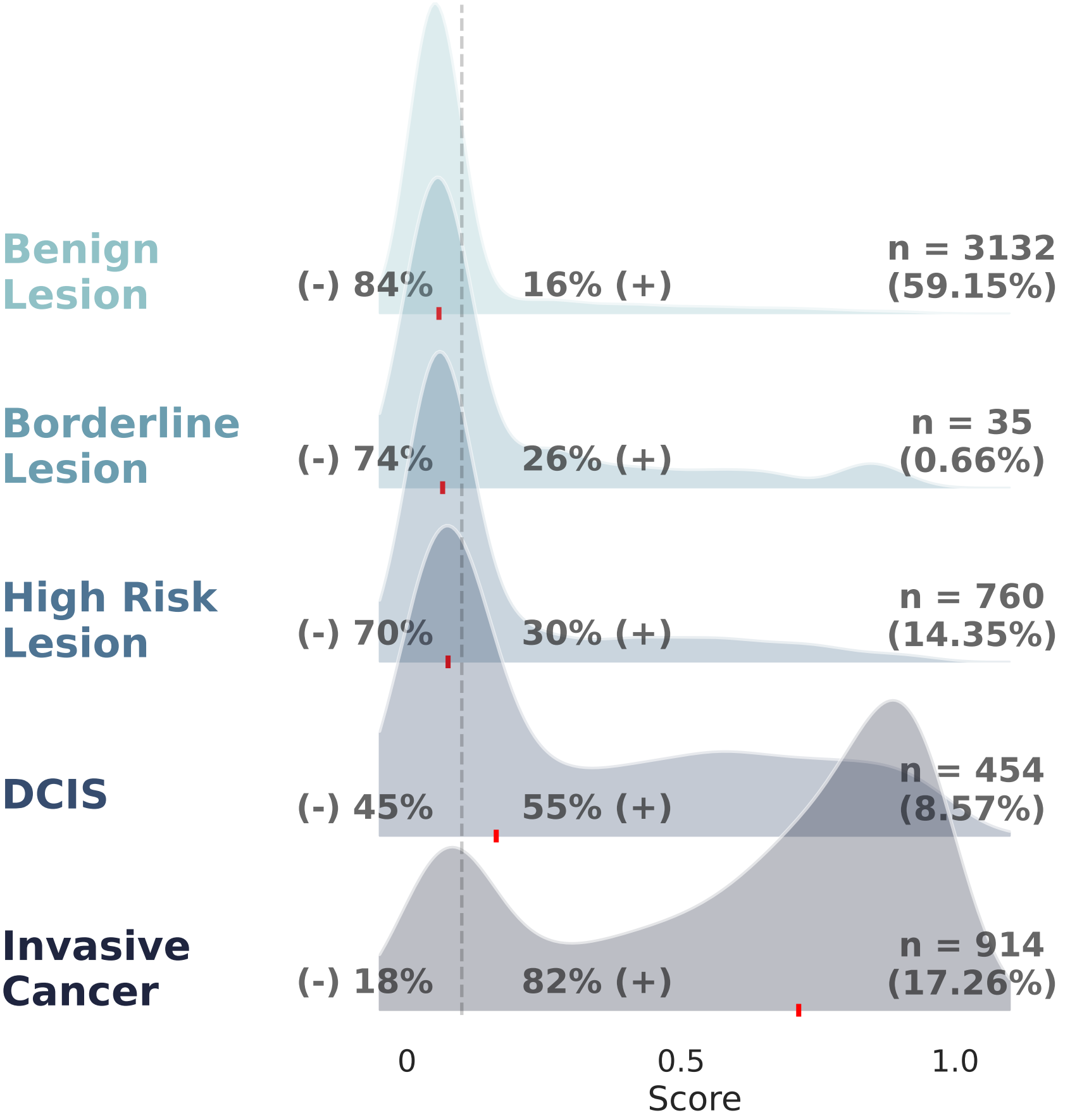
Figure 5: Distribution of model predicted scores for various DCIS subtypes. The model operating point of 0.1 is marked by a horizontal line with the percentage of exams above and below the threshold indicated per pathology. DCIS grades and subtypes are represented in the adjacent panel. Model performance for DCIS was significantly lower for all subtypes compared to invasive cancers (n=361).

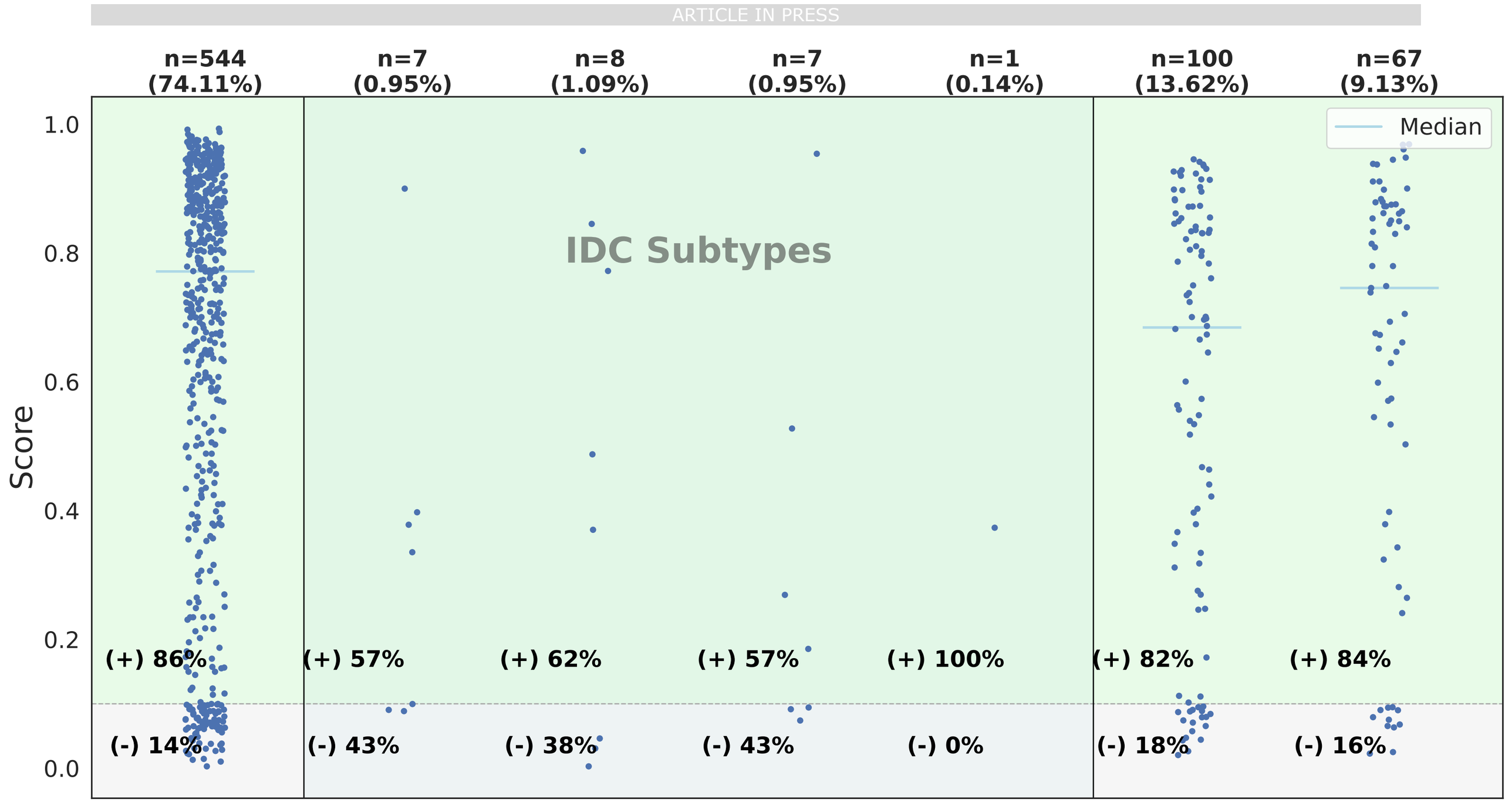
Figure 6: Distribution of model prediction scores at the exam level for cases containing various imaging features, structured by clinical outcome. For mass, asymmetry, and calcification, we observe generally that images containing these features are correctly predicted as negative in 60-81% of exams. However, architectural distortions are much less likely to be predicted as malignant regardless of pathologic outcome. Interestingly, architectural distortions from screen negative exams had a strong bimodal distribution of high and low scares, yielding many false positives (Masses: n=6,472, Asymmetries: n=13,775, Architectural Distortions: n=2,112, Calcifications: n=6,733, groups overlap wheres exams had multiple finding types).

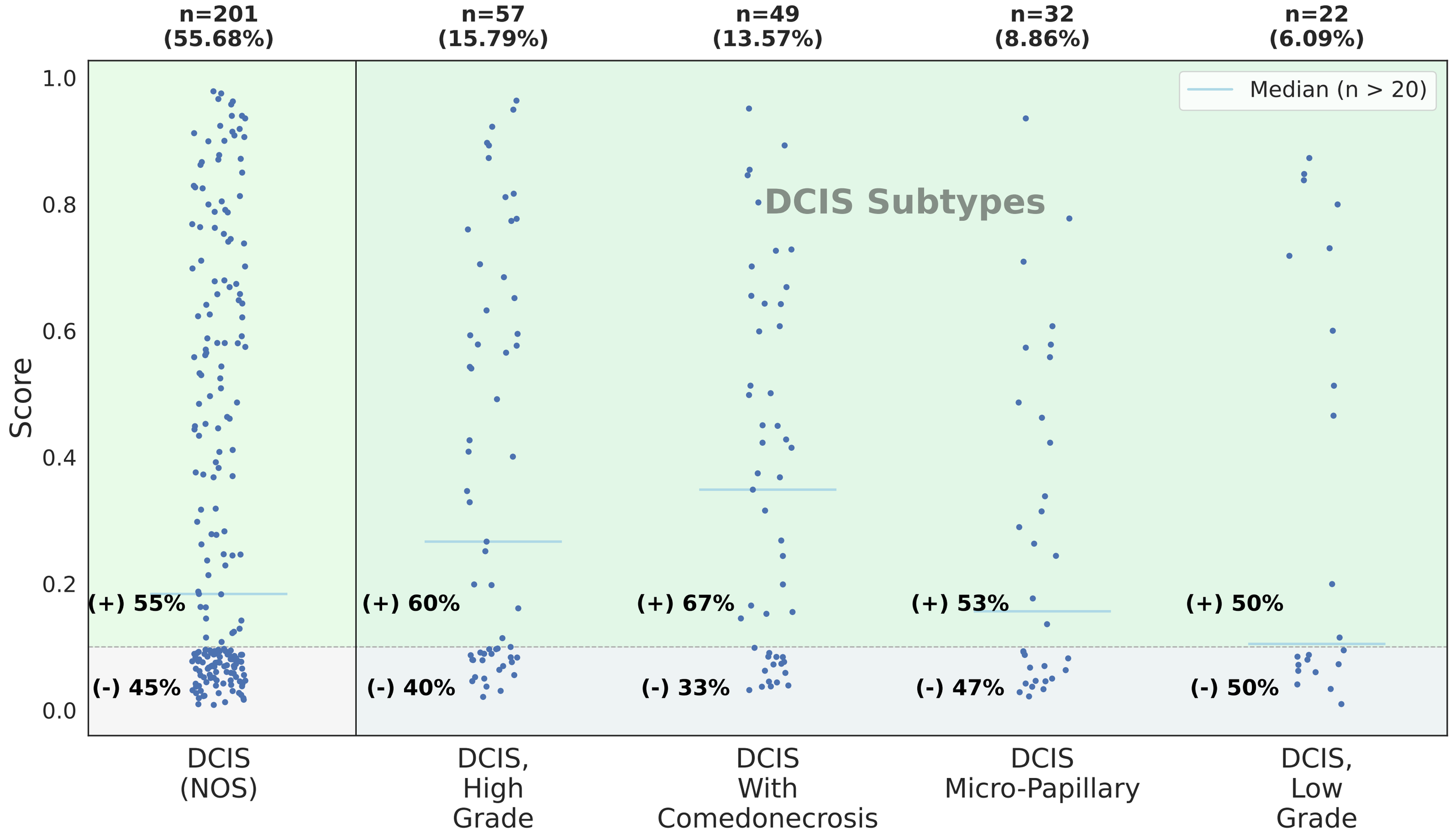
Figure 7: Examples of model predictions across imaging features of mass, asymmetry, architectural distortion, and calcifications. True positive lesions are from exams containing cancer that were correctly identified by the model. False positive lesions are from cases in which the model predicted cancer for exams that were either deemed negative at diagnostic imaging or benign following biopsy. False negative lesions are from exams that were malignant on biopsy, however predicted as negative by the model. Note that the model returns only breast-level and exam-level predictions and not regions of interest and these lesions are highlighted (green box) for demonstration purposes.

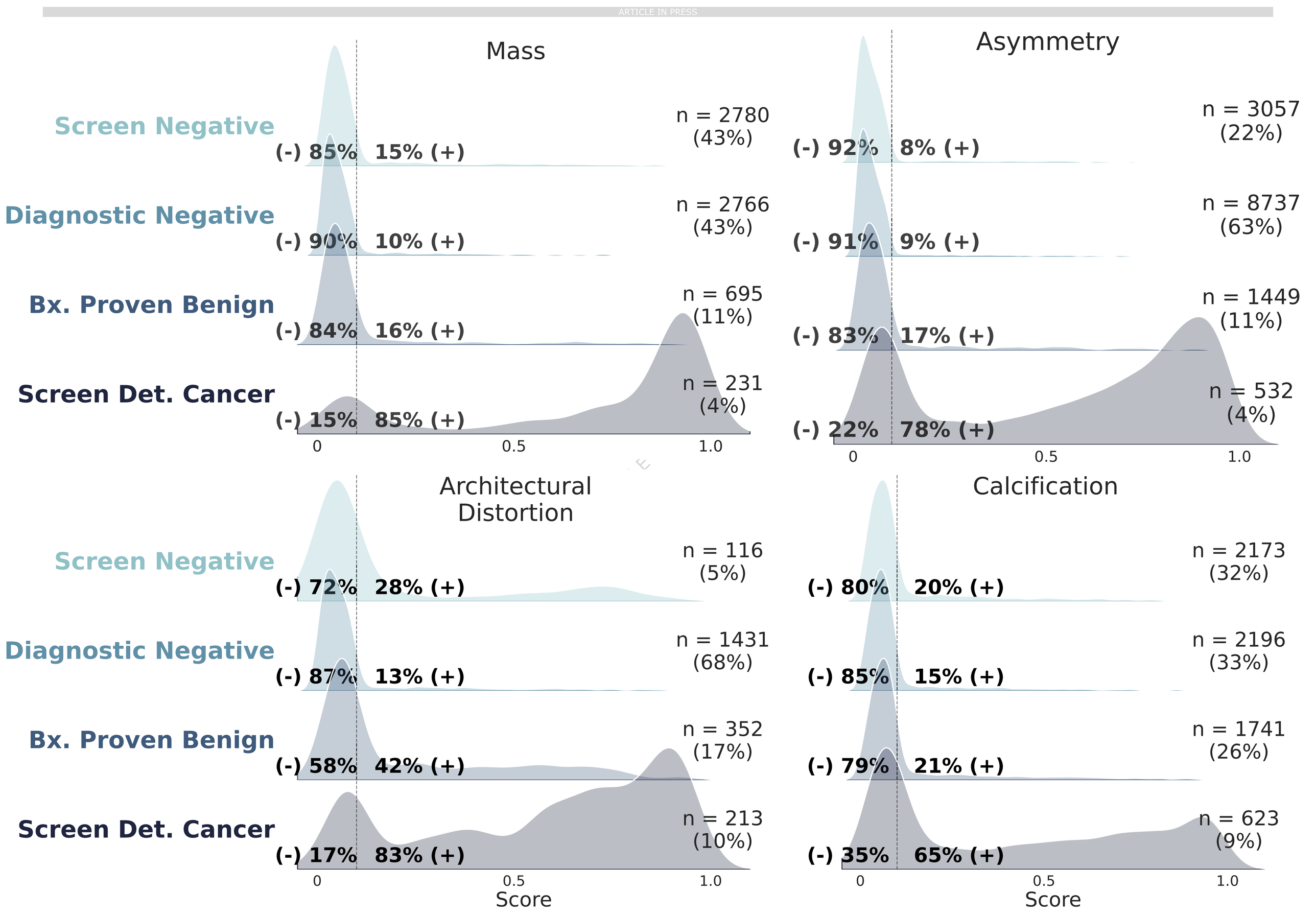

All Tomosynthesis Exams
Exams: 229,217
Exams that Failed to Process
Exams: 1
Tomosynthesis Exams with Scores
Exams: 229,216
Exams with Invalid Accessions
Exams: 9
Valid Tomosynthesis Exams
Exams: 229,207
Diagnostic Exams with Tomosynthesis
Exams: 26,284
Screening Exams with Tomosynthesis
Exams: 202,923
Screening BIRADS
BI-RADS 0
BI-RADS 1 / 2
Invalid/Missing Data
Abnormal Screening Exams
Exams: 24,897
Normal Screening Exams
Exams: 177,983
Invalid Exams
Exams: 43
Follow-Up Diagnostic? ≤ 6 months
No
Missing/Invalid Diagnostic
Exams: 4,001
Diagnostic? ≤ 12 months
No
Yes
Yes
Diagnostic BI-RADS
Cancer* Pathology?
No
Followup available in 1-4 years
No
BI-RADS 4 / 5
BI-RADS 1 / 2 / 3
Yes
Yes
Suspected Malignancy
Exams: 5,339
Diagnostic Negatives
Exams: 15,557
Interval Cancers
Exams: 105
Screen Negatives
Exams: 146,903
Missing Any Follow-Up
Exams: 30,975
Pathology
Cancer
HRL / BLL / Benign
Non-Breast Cancer
Screen Detected Cancers
Exams: 1,368
Biopsy Proven Benigns
Exams: 3,927
Non-Breast Cancers
Exams: 35
Missing Biopsy
Exams: 9
Invalid Exam Type
Valid Exam Type


Screen Negative
(-) 93%
7% (+)
n = 146903 (87.52%)
Diagnostic Negative
(-) 90%
10% (+)
n = 15557 (9.27%)
Bx. Proven Benign
(-) 81%
19% (+)
n = 3927 (2.34%)
Screen Det. Cancer
(-) 27%
73% (+)
n = 1368 (0.81%)
Interval Cancer
(-) 69%
31% (+)
n = 105 (0.06%)
0
0.5
1.0
Score


Benign Lesion
(-) 84%
16% (+)
n = 3132 (59.15%)
Borderline Lesion
(-) 74%
26% (+)
n = 35 (0.66%)
High Risk Lesion
(-) 70%
30% (+)
n = 760 (14.35%)
DCIS
(-) 45%
55% (+)
n = 454 (8.57%)
Invasive Cancer
(-) 18%
82% (+)
n = 914 (17.26%)
0
0.5
1.0
Score


n=544
(74.11%)
n=7
(0.95%)
n=8
(1.09%)
n=7
(0.95%)
n=1
(0.14%)
n=100
(13.62%)
n=67
(9.13%)
Median
IDC Subtypes
Score
1.0
0.8
0.6
0.4
0.2
0.0
(+) 86%
(-) 14%
(+) 57%
(-) 43%
(+) 62%
(-) 38%
(+) 57%
(-) 43%
(+) 100%
(-) 0%
(+) 82%
(-) 18%
(+) 84%
(-) 16%
Invasive Ductal Carcinoma (NOS)
Tubular Carcinoma
Invasive Papillary Carcinoma
Colloid (Mucinous) Carcinoma
Medullary Carcinoma
Invasive Lobular Carcinoma
Invasive Mammary Carcinoma


n=201
(55.68%)
n=57
(15.79%)
n=49
(13.57%)
n=32
(8.86%)
n=22
(6.09%)
Median (n > 20)
DCIS Subtypes
Score
1.0
0.8
0.6
0.4
0.2
0.0
(+) 55%
(-) 45%
(+) 60%
(-) 40%
(+) 67%
(-) 33%
(+) 53%
(-) 47%
(+) 50%
(-) 50%
DCIS (NOS)
DCIS, High Grade
DCIS With Comedonecrosis
DCIS Micro-Papillary
DCIS, Low Grade


Mass
Asymmetry
Architectural Distortion
Calcification
Screen Negative
Diagnostic Negative
Bx. Proven Benign
Screen Det. Cancer
(-) 85% 15% (+)
n = 2780 (43%)
(-) 92% 8% (+)
n = 3057 (22%)
(-) 90% 10% (+)
n = 2766 (43%)
(-) 91% 9% (+)
n = 8737 (63%)
(-) 84% 16% (+)
n = 695 (11%)
(-) 83% 17% (+)
n = 1449 (11%)
(-) 15% 85% (+)
n = 231 (4%)
(-) 22% 78% (+)
n = 532 (4%)
(-) 72% 28% (+)
n = 116 (5%)
(-) 80% 20% (+)
n = 2173 (32%)
(-) 87% 13% (+)
n = 1431 (68%)
(-) 85% 15% (+)
n = 2196 (33%)
(-) 58% 42% (+)
n = 352 (17%)
(-) 79% 21% (+)
n = 1741 (26%)
(-) 17% 83% (+)
n = 213 (10%)
(-) 35% 65% (+)
n = 623 (9%)
0
0.5
1.0
Score