Radiology

ORIGINAL RESEARCH • BREAST IMAGING

AsymMirai: Interpretable Mammography-based Deep Learning Model for 1–5-year Breast Cancer Risk Prediction

Jon Donnelly, BS • Luke Moffett, MS • Alina Jade Barnett, MS, PhD • Hari Trivedi, MD • Fides Schwartz, MD • Joseph Lo, PhD • Cynthia Rudin, PhD**

From the Departments of Computer Science (J.D., L.M., A.J.B., C.R.) and Electrical and Computer Engineering (C.R.), Duke University, 308 Research Dr, LSRC Building D101, Duke Box 90129, Durham, NC 27708; Department of Radiology and Imaging Services, Emory University, Atlanta, Ga (H.T.); Department of Radiology, Harvard University, Cambridge, Mass (F.S.); and Department of Radiology, Duke University School of Medicine, Durham, NC (J.L.). Received October 13, 2023; revision requested November 22; revision received January 27, 2024; accepted February 2. **Address correspondence to** J.D. (email: *jon.donnelly@duke.edu*).

Supported by the National Science Foundation (grant HRD-2222336).

* J.L. and C.R. are co-senior authors.

Conflicts of interest are listed at the end of this article.

See also the editorial by Freitas in this issue.

Radiology 2024; 310(3):e232780 • https://doi.org/10.1148/radiol.232780 • Content codes: BR AI

Background: Mirai, a state-of-the-art deep learning–based algorithm for predicting short-term breast cancer risk, outperforms standard clinical risk models. However, Mirai is a black box, risking overreliance on the algorithm and incorrect diagnoses.

Purpose: To identify whether bilateral dissimilarity underpins Mirai's reasoning process; create a simplified, intelligible model, AsymMirai, using bilateral dissimilarity; and determine if AsymMirai may approximate Mirai's performance in 1–5-year breast cancer risk prediction.

Materials and Methods: This retrospective study involved mammograms obtained from patients in the EMory BrEast imaging Dataset, known as EMBED, from January 2013 to December 2020. To approximate 1–5-year breast cancer risk predictions from Mirai, another deep learning–based model, AsymMirai, was built with an interpretable module: local bilateral dissimilarity (localized differences between left and right breast tissue). Pearson correlation coefficients were computed between the risk scores of Mirai and those of AsymMirai. Subgroup analysis was performed in patients for whom AsymMirai's year-over-year reasoning was consistent. AsymMirai and Mirai risk scores were compared using the area under the receiver operating characteristic curve (AUC), and 95% CIs were calculated using the DeLong method.

Results: Screening mammograms (n = 210 067) from 81 824 patients (mean age, 59.4 years ± 11.4 [SD]) were included in the study. Deep learning–extracted bilateral dissimilarity produced similar risk scores to those of Mirai (1-year risk prediction, r = 0.6832; 4–5-year prediction, r = 0.6988) and achieved similar performance as Mirai. For AsymMirai, the 1-year breast cancer risk AUC was 0.79 (95% CI: 0.73, 0.85) (Mirai, 0.84; 95% CI: 0.79, 0.89; P = .002), and the 5-year risk AUC was 0.66 (95% CI: 0.63, 0.69) (Mirai, 0.71; 95% CI: 0.68, 0.74; P < .001). In a subgroup of 183 patients for whom AsymMirai repeatedly highlighted the same tissue over time, AsymMirai achieved a 3-year AUC of 0.92 (95% CI: 0.86, 0.97).

Conclusion: Localized bilateral dissimilarity, an imaging marker for breast cancer risk, approximated the predictive power of Mirai and was a key to Mirai's reasoning.

© RSNA, 2024

Supplemental material is available for this article.

Interpretability is essential for the ethical application of artificial intelligence (AI) to radiology (1), a requirement likely to be enshrined via regulation in the United States (2) and Europe (3). This conflicts with the design of many top-performing radiology AI algorithms, which are black boxes to developers and radiologists alike. This can cause an overreliance on algorithms (4) and incorrect diagnoses (5). Clinical risk prediction models for breast cancer do not consider mammography image data (6–8), despite recent AI studies reporting significantly improved performance when mammography data are used (9). Better risk prediction is an active research goal because it is instrumental for the development of personalized screening strategies aimed at simultaneously reducing the financial and psychologic burdens of screening mammography and justifying the use of targeted advanced imaging (10–12).

The case of recent interest is Mirai, a deep learning neural network trained on screening mammograms from 56 786 patients to predict short-term (up to 5 years) breast cancer risk (13). Mirai results were externally validated on data from seven hospitals across three continents (14). The robust performance suggests that Mirai has captured critical information that may complement existing clinical risk models. However, Mirai's predictions are difficult to interpret because Mirai consists of a convolutional neural network (CNN) and a transformer (15), two distinct, complex architectures. Post hoc explanations of neural networks such as GradCAM and GradCAM++ are not reliable (16–18), and because of the unique architecture of

This copy is for personal use only. To order copies, contact *reprints@rsna.org*

Abbreviations

AI = artificial intelligence, AUC = area under the ROC curve, CNN = convolutional neural network, EMBED = EMory BrEast imaging Dataset, ROC = receiver operating characteristic

Summary

Using bilateral dissimilarity as a mammography marker of near-term breast cancer risk, AsymMirai, a simplified deep learning bilateral dissimilarity–based model, performed similarly to the state-of-the-art black box model, Mirai, for 1–5-year breast cancer risk prediction.

Key Points

- In a retrospective study of 210 067 screening mammograms (81 824 patients), bilateral dissimilarity as measured with AsymMirai, a simplified alternative to Mirai, performed similarly to Mirai (1-year risk, $r = 0.6832$; 2-year risk, $r = 0.6988$).
- For predicting cancer, AsymMirai achieved areas under the receiver operating characteristic curve (AUCs) of 0.79, 0.69, 0.68, 0.67, and 0.66 for 1–5-year horizons.
- In a patient subgroup in which AsymMirai repeatedly highlighted the same tissue over time; its 3-year AUC was 0.92.

Mirai, those methods are not applicable. As a result, to our knowledge, no explanation of Mirai's reasoning process has been provided prior to this work.

Prior work (19,20) has shown that explicit bilateral reasoning can support AI breast imaging models. To evaluate bilateral dissimilarity (differences between corresponding left- and right-laterality views from a patient's mammograms), we propose AsymMirai. AsymMirai is a simplified alternative to Mirai that computes risk by using only localized bilateral dissimilarities. Thus, the aim of our study was to *(a)* identify whether bilateral dissimilarity as a mammography marker underpins the deep learning model Mirai's reasoning process for high-quality predictions; *(b)* use bilateral dissimilarity as an imaging marker to create a simplified model, AsymMirai, with an intelligible reasoning process; and *(c)* evaluate both models to determine if AsymMirai may approximate performance of Mirai in 1–5-year breast cancer risk prediction.

Materials and Methods

Study Design

This retrospective study was compliant with the Health Insurance Portability and Accountability Act and was approved by the institutional review board. The requirement for informed consent was waived by the institutional review board. Mirai was initially applied to public data sets containing unilateral images. Since Mirai required bilateral views, all unilateral views were mirrored, resulting in uniformly low-risk predictions. All images in the bilateral examinations were then mirrored, and it was confirmed that Mirai consistently predicted low risk for mirrored examinations, even those with actionable lesions (Appendix S1). This demonstrated that Mirai relies on bilateral dissimilarities.

Using this insight, a neural architecture was developed around bilateral dissimilarity. This study used the EMory BrEast imaging Dataset (EMBED) (21), a retrospective data set containing full-field screening and diagnostic mammograms from 116 890 patients obtained from January 2013 to December 2020 by using Hologic (92%), General Electric (6%), and Fujifilm (2%) machines. EMBED contains self-reported race descriptors for the entire cohort and cohorts for training (70 136 patients), validation (23 382 patients), and testing (23 333 patients). The patient cohort in this study is the same as that in the study by Jeong et al (21), which introduced this public data set. EMBED was chosen because it was included in a 2022 external validation of Mirai (14). We excluded examinations with data abnormalities, examinations without two-dimensional images, examinations without all four screening views, and diagnostic examinations from our study (Fig 1). The code is available at *https://github.com/jdonnelly36/AsymMirai/releases/tag/radiology-1.0*.

Model Architecture

Mirai and AsymMirai both accept as inputs the four standard screening mammography views—left and right mediolateral oblique and left and right craniocaudal—passing them through identical ResNet-18 CNN backbones, extracting features for each view. Mirai passes these extracted features to a transformer, which predicts clinical risk factors and *n*-year breast cancer risk. In contrast, AsymMirai simply computes a localized bilateral dissimilarity between the left and right breast at multiple locations using these features for each view. The maximum dissimilarity across locations—called the prediction window—produces one dissimilarity score for each view; the scores are averaged to produce one bilateral dissimilarity score. By excluding Mirai's transformer, AsymMirai maintains spatial correspondence between the extracted features and the input images. AsymMirai omits nonimaging features, which did not benefit Mirai (13). AsymMirai's architecture allows its outputs to be directly overlayed on the mammogram, highlighting dissimilarities. Figure 2 summarizes the model architecture. AsymMirai is described in detail in Appendix S2.

Model Evaluation: Predictive Power

AsymMirai was evaluated on two fronts. In addition to the mirroring analysis in Appendix S1, the Pearson correlation between the predictions by AsymMirai and Mirai was computed. Second, AsymMirai was evaluated by using dissimilarity to predict risk, enabling comparison with Mirai using Mirai's metrics. How well these scores predict breast cancer was assessed by plotting the 1–5-year risk receiver operating characteristic (ROC) curves and determining the corresponding area under ROC (AUC) (13). A screening examination was included in the *n*-year ROC calculation if *(a)* an *n*-year positive examination had positive pathologic findings within *n* years or *(b)* an *n*-year negative examination had a negative screening follow-up at least *n* years later.

AsymMirai issues one prediction per examination, while Mirai issues five predictions, one for each year into the future (Appendix S3 explains this difference). The same score was used when evaluating the *n*-year risk for AsymMirai, while the distinct *n*-year risk prediction from Mirai was used. Both models were also assessed on subgroups by age and race to determine whether either model was biased along these features.

Model Evaluation: Applications of Bilateral Dissimilarity

The outputs of AsymMirai were visualized by overlaying a heat map of AsymMirai's computed bilateral dissimilarity scores and highlighting the prediction window with a red box. Unlike post hoc saliency maps, these overlays faithfully visualize AsymMirai's computed bilateral dissimilarity. Using these overlays, two post hoc analyses of AsymMirai's reasoning were performed.

First, these visualizations were used as diagnostic criteria to identify confounders. Figure 3 presents examples. For these illustrations, prediction outputs were binned as low (risk score, <0.25), moderate (risk score, 0.25–0.50), or high (risk score, >0.50) risk. Second, using 10 001 patients (26 930 examinations) with multiple screenings, patients' dissimilarity distributions over time were analyzed, quantifying whether the same tissue produced the maximum asymmetry across examinations. AsymMirai's predictions were analyzed as a function of "location consistency," the change in prediction window location between current and previous examinations (Fig 4; formal definition in Appendix S4). A similar analysis was performed for Mirai, using the change in Mirai's risk scores (Appendix S5).

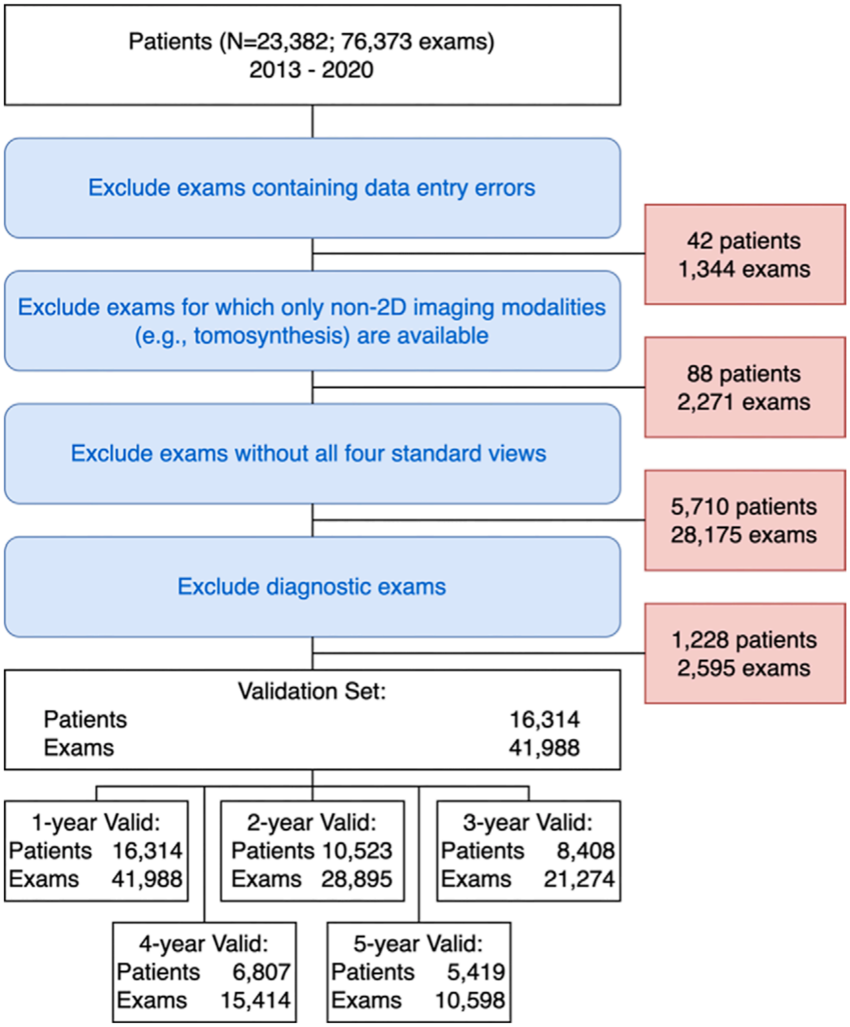

Figure 1: Exclusion flowchart for the validation cohort. The EMory BrEast imaging Dataset (EMBED) validation split included 23 382 patients and 76 373 examinations from 2013 to 2020. Examinations with data abnormalities (42 patients, 1344 examinations), examinations without two-dimensional (2D) images (88 patients, 2271 examinations), examinations without all four screening views (5810 patients, 28 175 examinations), and diagnostic examinations (1228 patients, 2595 examinations) were excluded. The resulting cohort included 16 314 patients with 41 988 examinations. The number of patients and examinations with sufficient follow-up data to evaluate 1-year (16 314 patients, 41 988 examinations), 2-year (10 523 patients, 28 895 examinations), 3-year (8408 patients, 21 274 examinations), 4-year (6807 patients, 15 414 examinations), and 5-year (5419 patients, 10 598 examinations) areas under the receiver operating characteristic curve are at the bottom of the figure.

Statistical Analysis

The 95% CIs and *P* values for the AUC were calculated by using the DeLong method (22). *P* value < .05 was considered to indicate a significant difference, and all statistical analyses were computed with SciPy (version 1.7.3, *https://scipy.org/*), ROC (version 0.1, *https://github.com/alistairewj/pyroc*), and Python (version 3.7.3, *https://www.python.org/*) packages. Correlations greater than 0.7 are considered high, as prescribed by Mukaka in 2012 (23). Statistical analysis was performed by two authors (J.D. and L.M.). All data available were used for each analysis.

Results

Patient Demographics

This study considered EMBED, a retrospective breast imaging data set containing images from 116 890 patients (mean age, 58.5 years ± 12.1 [SD]) obtained between 2013 and 2020. AsymMirai was trained on EMBED's training cohort. Patients self-reported their race as African American, American Indian or Alaskan Native, Asian, multiple, Native Hawaiian or Pacific Islander, unknown, or White.

To evaluate performance on the same cohort as Mirai, results are reported on the EMBED validation cohort, which was used in Mirai's external validation; 2.9% (*n* = 679) of patients in this cohort were diagnosed with breast cancer. Appendix S6 reports AsymMirai's performance on the EMBED test cohort. The validation cohort is included in the publicly available EMBED Open Data data set *(https://github.com/Emory-HITI/EMBED_Open_Data)*. In the EMBED validation cohort, we excluded examinations with

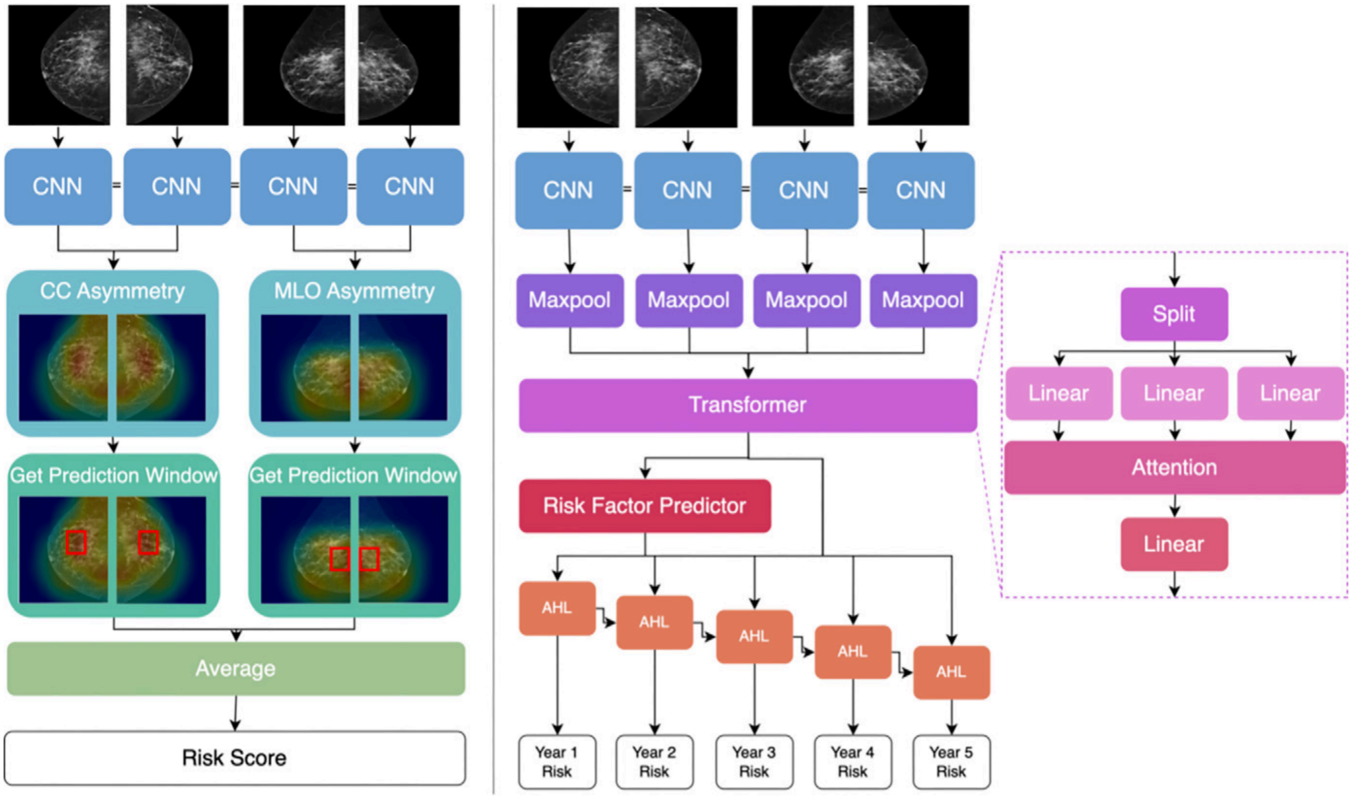

Figure 2: Architecture comparison of AsymMirai (left) and Mirai (right). Both models feed the four screening views into the same convolutional neural network (CNN) layers, but reasoning diverges thereafter. AsymMirai has fewer computational layers and instead calculates differences in the latent features, as shown by heat maps in the craniocaudal (CC) asymmetry and mediolateral oblique (MLO) asymmetry steps. AsymMirai then finds the prediction window containing the highest differences for each view, represented by red boxes in the Get Prediction Window step. The maximum feature differences within these windows are averaged to create a risk score. The Mirai architecture was described by Yala et al (13). AHL = additive hazard layer.

data abnormalities (42 patients and 1344 examinations), examinations without two-dimensional images (88 patients and 2271 examinations), examinations without all four screening views (5810 patients and 28 175 examinations), and diagnostic examinations (1228 patients and 2595 examinations) (Fig 1). Table 1 summarizes the distributions of patient age and race.

Model Evaluation: Predictive Power

AsymMirai achieved an AUC of 0.79 (95% CI: 0.73, 0.85) for 1-year risk prediction task (Mirai AUC: 0.84; 95% CI: 0.79, 0.89; 16 314 patients; $P = .002$), an AUC of 0.68 (95% CI: 0.65, 0.71) for 3-year risk prediction (Mirai AUC: 0.72; 95% CI: 0.69; 0.76; 8408 patients; $P < .001$), and an AUC of 0.66 (95% CI: 0.63, 0.69) for 5-year risk prediction (Mirai AUC: 0.71; 95% CI: 0.68, 0.74; 5419 patients; $P < .001$). The difference between the AUCs for the two models was at most 0.05 for all tasks, although the 95% CIs overlapped in all the cases. Figure 5A and B shows the performance of AsymMirai and Mirai on the EMBED validation set.

Note that Mirai's inclusion criteria on EMBED admitted diagnostic mammograms (14), which may bias a risk model intended for screening. After excluding these data, Mirai's AUC for 5-year risk prediction on the screening-only images decreased by 0.05, from 0.76 to the 0.71 reported here (Fig 5).

Figure 5C shows the correlation between the predictions of the two models for 1-, 3-, and 5-year risk scores. The Pearson correlation coefficients (*r* values) for the *n*-year predictions between AsymMirai and Mirai starting at 1 year were 0.6832, 0.7011, 0.7011, 0.6987, and 0.6987 (95% CIs are shown in Fig 5).

A subgroup performance analysis was performed, and the results are reported in Table 2. There are two main results from this analysis. First, both AsymMirai and Mirai demonstrated lower performance (3-year AUC: AsymMirai, 0.63; Mirai, 0.69) in the African American race subgroup ($n = 6812$) than in the White race subgroup ($n = 6689$; 3-year AUC: AsymMirai, 0.73; Mirai, 0.77). Second, AsymMirai and Mirai showed similar performance for risk prediction in the younger than 50 years age group ($n = 9967$; 3-year AUC: AsymMirai, 0.69; Mirai, 0.71).

For completeness, AsymMirai was also evaluated on the EMBED test set (cohorts 9 and 10; recall that the aforementioned results were based on the validation set). The performance of AsymMirai on the EMBED test set was similar to that on the validation set, and the 95% CIs overlapped with the validation set results (Appendix S6). This test set was not used for the evaluation of Mirai and thus was not the focus of this study.

Model Evaluation: Applications of Bilateral Dissimilarity

Examples of patients with moderate- to high-risk scores from AsymMirai are shown in Figure 3, including patients

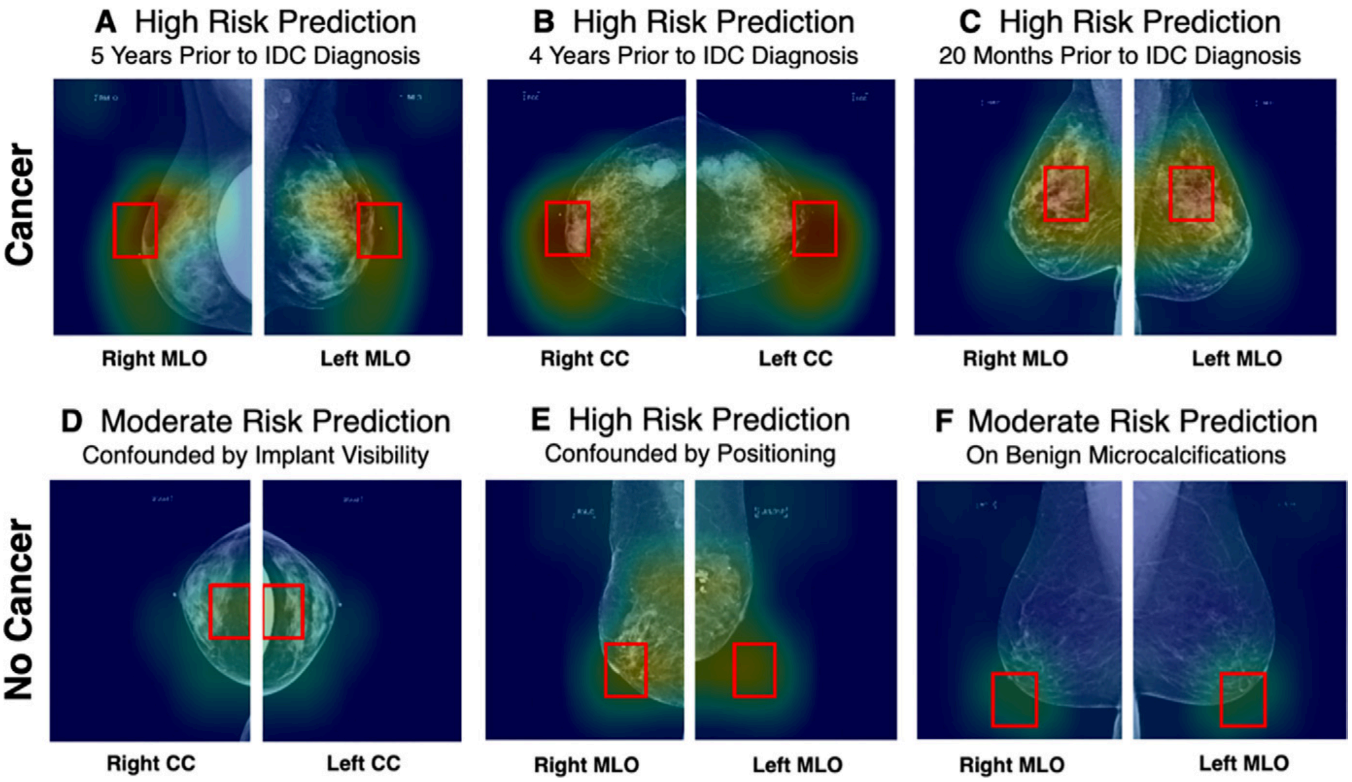
A High Risk Prediction
5 Years Prior to IDC Diagnosis
B High Risk Prediction
4 Years Prior to IDC Diagnosis
C High Risk Prediction
20 Months Prior to IDC Diagnosis
Cancer
Right MLO
Left MLO
Right CC
Left CC
Right MLO
Left MLO
D Moderate Risk Prediction
Confounded by Implant Visibility
E High Risk Prediction
Confounded by Positioning
F Moderate Risk Prediction
On Benign Microcalcifications
No Cancer
Right CC
Left CC
Right MLO
Left MLO
Right MLO
Left MLO

Figure 3: AsymMirai model outputs. Input images are full-field screening mammograms. The two bilateral screening images are overlayed within the heat map, and the prediction window (red box) indicates the area with the highest dissimilarity. The heat map and prediction window are visualizations of AsymMirai's model outputs, not post hoc saliency maps such as GradCAM. Analyzing these outputs provides a deeper understanding of the scores, in these cases distinguishing confounded reasoning from nonconfounded reasoning for patients with macro asymmetries. **(A–C)** Images in patients who developed cancer within 1–5 years. **(A)** In a 49-year-old White woman with unilateral breast augmentation who underwent annual screening, AsymMirai predicted high risk for developing cancer. Biopsy confirmed invasive ductal carcinoma in the right breast 5 years later. The prediction window was not affected by the unilateral implant. **(B)** In a 43-year-old African American woman with initial screening at 42 years old, AsymMirai predicted high risk of developing cancer. The prediction window corresponds to retroareolar asymmetry. Biopsy performed 4 years later confirmed invasive ductal carcinoma in the right breast. Intramammary lymph nodes were correctly ignored. **(C)** In a 50-year-old African American woman with regular screening and coarse heterogenous calcifications at the 12-o'clock position, AsymMirai predicted high risk for developing cancer. Biopsy confirmed bilateral invasive ductal carcinoma 20 months later, with the cancer in the left breast occurring in the 12-o'clock position. **(D–F)** Images in patients who did not develop cancer but had identifiably confounded risk predictions. **(D)** In a 60-year-old White woman with bilateral breast augmentation and regular screening mammograms, AsymMirai predicted moderate risk for developing cancer, confounded by artificial asymmetry caused by the exclusion of the implant from the right craniocaudal view. **(E)** In a 73-year-old White woman with regular screening mammograms and known dystrophic calcifications in the left breast, AsymMirai predicted high risk for developing cancer, confounded by poor positioning in the left mediolateral oblique view and possible distortion in the right mediolateral oblique view. **(F)** In a 65-year-old African American woman with bilateral benign microcalcifications, AsymMirai predicted moderate risk for developing cancer, confounded by the calcifications. Among the patients with no cancer, Mirai correctly identified the patient in **D** as having a low risk for developing cancer (20th percentile risk) but also misclassified patients in **E** and **F** (84th and 95th percentiles, respectively). These examples were chosen without knowledge of Mirai's risk scores. Unlike when reviewing the tissue in AsymMirai prediction window, there is no way to ex ante identify the cases where Mirai was confounded because it produces only a score. CC = craniocaudal, IDC = invasive ductal carcinoma, MLO = mediolateral oblique.

with true-positive results who developed cancer within 5 years, as well as patients with false-positive results who did not develop cancer within 1–5 years. In the examples, which have macro asymmetries, false-positive results were caused by confounding factors such as misalignment or implants.

AsymMirai demonstrated superior predictive power for patients for whom the same tissue was highlighted over multiple years (as measured by location consistency) (Fig 4). Location consistency is expressed as a shift relative to the prediction window size. This is a strict criterion because the prediction windows cover only $\frac{1}{25}$ of each view. For example, a 40% prediction window location shift is 20 × 24 mm or less in window distance. This 40% threshold appears to be operative for increasing predictive power (Fig 4).

There were 549 patients who underwent subsequent examinations with a window shift of 40% or less, covering 1154 examinations. Of these, 383 patients had sufficient follow-up data (were *n*-year positive or *n*-year negative) for calculating the 1-year AUC (0.92; 95% CI: 0.88, 0.97); 256 patients, for calculating the 2-year AUC (0.91; 95% CI: 0.85, 0.96); 183 patients, for calculating the 3-year AUC (0.92; 95% CI: 0.86, 0.97); 119 patients, for calculating the 4-year AUC (0.90; 95% CI: 0.82, 0.98); and 59 patients, for calculating the 5-year AUC (0.88; 95% CI: 0.79, 0.98). Sufficient follow-up for location consistency was measured from the second screening examination. Figure 4 shows the ROC curves for the 3-year AUC subgroup at different location consistency thresholds (range, 30%–100%). Figure

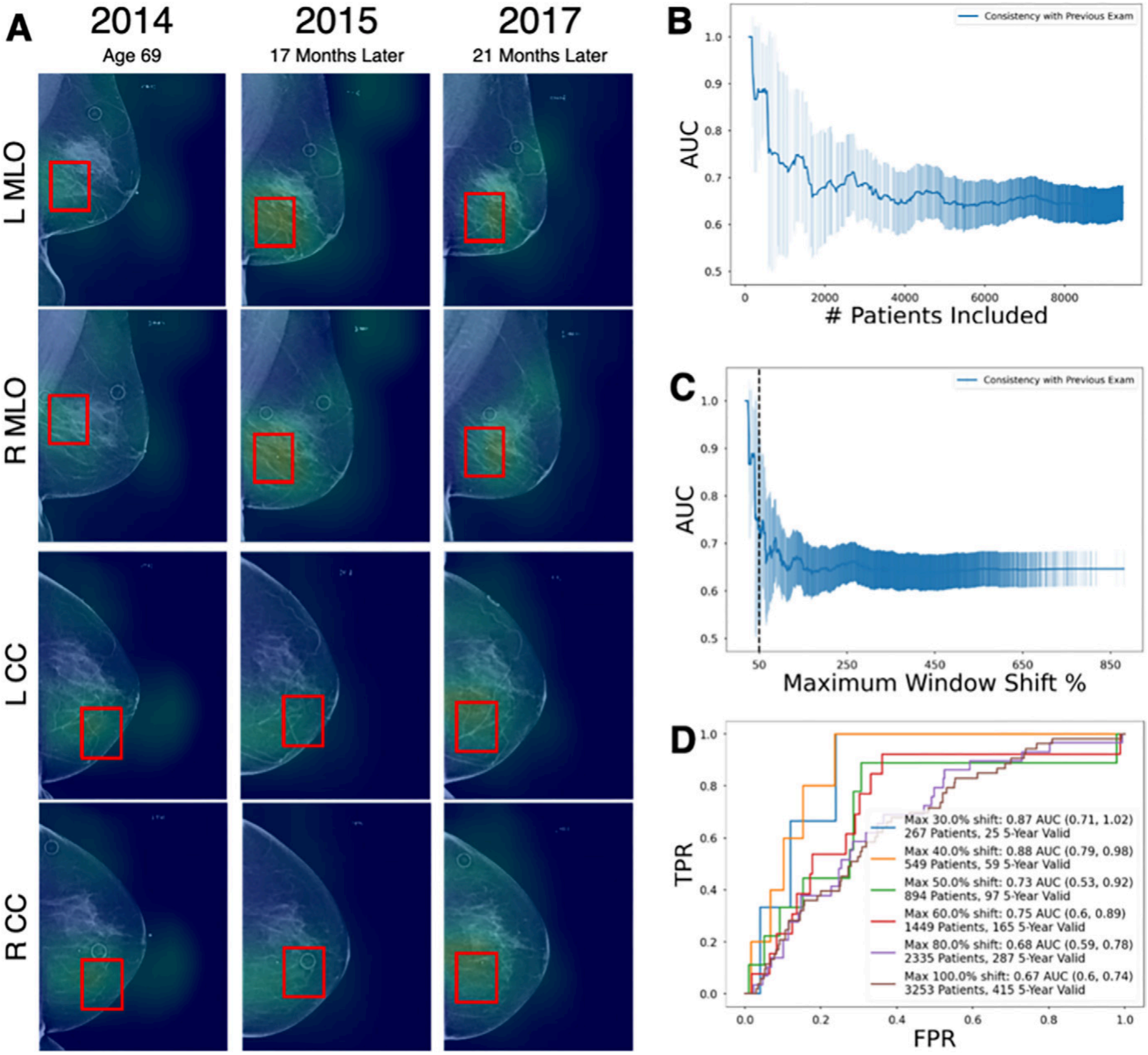

Figure 4: Prediction power of AsymMirai location consistency. **(A)** Full-field screening mammograms obtained at three time points in a White woman. AsymMirai predicted moderate risk for developing cancer, with high location consistency across three screenings. The patient was diagnosed with ductal carcinoma in situ in 2020. The location consistency is defined in Appendix S5. Consistency is expressed as the percentage of the window shift, with a shift of 100% representing no overlap from one year to the next. The red boxes are AsymMirai's prediction windows for each examination. **(B)** Graph of AsymMirai 3-year risk area under the receiver operating characteristic (ROC) curve (AUC) for patient subgroups with increasing location inconsistency. The x-axis is the number of patients included in the subgroup. Model performance is highest for patients with the highest location consistency (left part of the plot), as measured by the shift from the preceding examination's prediction window location. The shaded areas represent the 95% CIs at each threshold. **(C)** Graph of AsymMirai 3-year risk AUC for patient subgroups with increasing location inconsistency. Same as in **B**, except for the x-axis, location consistency is expressed as the window shift percentage. The dotted vertical line indicates a window shift of 50%. **(D)** AsymMirai ROC curves for selected location consistency thresholds as measured by the shift from the previous prediction window location. Model performance improved for patients with high location consistency between examinations, as indicated by lower window shifts. The legend contains the number of patients with an examination satisfying each threshold followed by the number of patients with at least one 3-year valid examination from each subgroup. A 3-year valid examination can include either 3 years of negative screening follow-up or a cancer diagnosis within 3 years. CC = craniocaudal, FPR = false-positive rate, MAX = maximum, MLO = mediolateral oblique, TPR = true-positive rate.

S1 provides the same results for the 1–5-year AUCs, which show that *(a)* an approximately 40% shift threshold is operative for this location consistency metric to yield strong predictions (AUC ≥ 0.88 across risk terms) and *(b)* improved performance persists on all five risk horizons.

For the subgroup of 59 patients with 40% or lower location consistency and 5 years of follow-up from the second examination, one classification threshold yielded 100% sensitivity (five of five examinations) and negative predictive value (45 of 45 examinations) with a specificity of 76% (45 of 59 examinations). This is because only five patients (0.9%) in this group developed cancer, whereas 2.6% of those in the entire EMBED developed cancer. Of those five patients, four had prediction windows on or adjacent to the

Table 1: Descriptive Statistics of Patients Included in the Validation Data Set

Patient Group	All Patients	Patients in Validation Data Set
No. of patients	81 824 [116 890] (1301)	16 314 [23 382] (236)
No. of examinations	210 067 [383 379]	41 988 [76 373]
Age at examination (y)*	59.5 ± 11.4 [58.9 ± 11.9]	59.6 ± 11.4 [58.9 ± 12.0]
Age group (y)†		
<40	2352 [11 478] (18)	508 [2355] (4)
40–49	48 027 [89 667] (208)	9459 [17 660] (65)
50–59	59 329 [104 405] (372)	11 732 [20 710] (118)
60–69	58 515 [101 458] (475)	11 737 [20 180] (142)
70–79	33 140 [59 359] (345)	6844 [12 135] (133)
≥80	8054 [15 400] (55)	1591[2992] (9)
Unknown	470 [1612] (0)	117 [341] (0)
Race		
African American	34 369 [48 452] (591)	6812 [9653] (104)
American Indian or Alaskan Native	195 [310] (2)	41 [67] (0)
Asian	5279 [7615] (45)	1060 [1566] (7)
Native Hawaiian or Pacific Islander	736 [1138] (10)	150 [218] (2)
Multiple	310 [516] (2)	68 [103] (1)
White	33 352 [45 328] (626)	6689 [9089] (119)
Unknown	7583 [13 531] (24)	1494 [2686] (3)

Note.—The full and validation patient data sets were constructed from the EMory BrEast imaging Dataset (EMBED). The validation data set was the EMBED validation set, which was used for Mirai external validation. Except where indicated, the number of patients satisfying the selection criteria described in Figure 1 in each subgroup is reported, with the number of patients prior to exclusion criteria in brackets and the number of patients who eventually developed cancer in parentheses. Age and race data were collected from the electronic health records used to construct EMBED.

* Data are means ± SDs. Data in brackets are for patients before exclusion.

† Data are numbers of examinations for which the patient was within the given subgroup. One patient may appear in multiple age groups over time. Patients were included if they had at least one valid examination.

area where biopsy-confirmed cancer would later be identified in the examinations preceding diagnosis. The consistent examinations for the remaining patient occurred 6 and 4 years before biopsy confirmed cancer.

There is no way to measure location consistency for Mirai since its reasoning process is opaque. In lieu of a reasoning consistency metric, consistency in Mirai's final risk predictions for identifying useful subgroups was evaluated, but that analysis failed to reliably enhance the confidence in its risk scores (Appendix S5).

Discussion

Although artificial intelligence algorithms, particularly Mirai, show promise in near-term breast cancer risk prediction, most methods are black boxes. We reduced the opacity of this black box by introducing AsymMirai. We determined a key factor on which Mirai depends—bilateral dissimilarity. Using the existing Mirai front-end convolutional neural network for feature extraction, our approach calculates differences in the latent space, providing the location of the dissimilarity, which is visually intuitive. This score approximates that of Mirai ($r > 0.6832$ for 1–5-year risk prediction), with only a slight reduction in 1–5-year risk prediction performance. The relative results are consistent across different prediction horizons (approximately 0.05 decrease in area under the receiver operating characteristic curve [AUC] with overlapping 95% CIs for each of the 1–5-year intervals), although the validity of the 1-year risk prediction AUC (inherited from Mirai) is debatable given that the cancer may already exist and neither Mirai nor AsymMirai is intended for diagnosis.

We demonstrated two possible uses of localized bilateral dissimilarity not available to black box models. We identified confounded model predictions from erroneously placed prediction windows. We further showed that, when the prediction window was in the same location over multiple years, AsymMirai exhibited superior predictive power. We originally expected that this would be the case because AsymMirai would find abnormalities in the tissue before the development of the actual lesion. While this does occur, most patients with location consistency of 40% or less showed little change from prior examinations and thus corresponded to a very low-risk group; they had a cancer rate of only 0.9%, compared with the 2.6% rate for the overall EMBED.

Our study focused on breast cancer risk prediction based on bilateral dissimilarity, which is related but not equivalent to the concept of breast asymmetry used in the Breast Imaging Reporting and Data System, or BI-RADS. Conventional studies in image-based risk prediction relied on a handcrafted approach that used unilateral or bilateral computer vision features to train machine learning models such as a support vector machine (24), a method successfully deployed with bilateral dissimilarity features in 2013 (25). In contrast, AsymMirai uses the CNN front end of Mirai, which leverages the learned latent features of that powerful model.

External validation of Mirai on 62 185 patients at seven sites, including the EMBED, was completed in 2022 and found that Mirai generalized well (14). However, Mirai's reported EMBED external validation did not exclude diagnostic mammograms (14). This led to a small difference between our results for Mirai and those from the study by Yala et al (14). A recent study used a large, enriched screening cohort to evaluate several AI risk prediction models (9). The performance of Mirai was lower on the private data set used in that study (AUC range, 0.67–0.69) than on EMBED

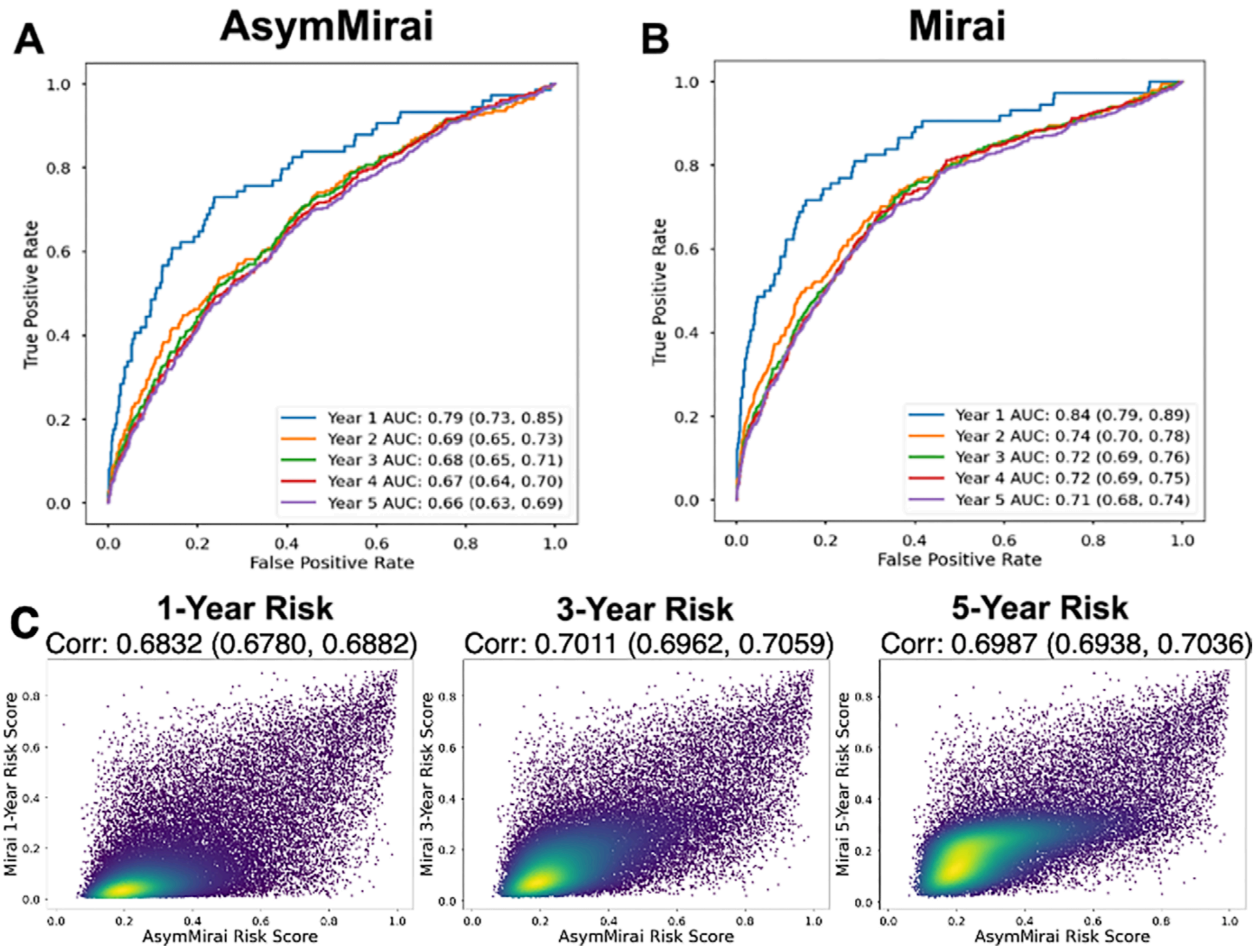

Figure 5: Comparison of the performance of Mirai and AsymMirai on EMory BrEast imaging Dataset (EMBED) validation screening mammograms. **(A)** AsymMirai 1–5-year breast cancer risk prediction receiver operating characteristic (ROC) curves and area under the curve (AUC) values, with 95% CIs in parentheses. **(B)** Mirai 1–5-year breast cancer risk prediction ROC curves and AUC values, with 95% CIs in parentheses. The AUC CIs for AsymMirai and Mirai overlap for each year. **(C)** Density plots show prediction correlation for AsymMirai and Mirai with 1-, 3-, and 5-year risk. The Pearson correlation coefficients were 0.6832 (95% CI: 0.6780, 0.6882), 0.7011 (95% CI: 0.6962, 0.7059), and 0.6987 (95% CI: 0.6938, 0.7036) for 1-, 3-, and 5-year risk, respectively. The 2- and 4-year risks are omitted because the predictions are the same as those for the 3- and 5-year risks, respectively.

(AUC range, 0.71–0.84), although Mirai maintained state-of-the-art performance.

In addition to the limitations inherent to any retrospective study, our study had four important limitations. First, despite the importance of bilateral dissimilarity, Mirai does not exclusively reason using this feature. Mirai's transformer is capable of arbitrary function approximation. When model predictions differ, we cannot explain Mirai's decision, except that these are the cases in which AsymMirai's localized bilateral dissimilarity is not the entire explanation. Figure 3 provides illustrative, confounded cases. Second, AsymMirai does not perform equally well across race subgroups. This limitation is inherited from Mirai because of the reuse of Mirai's backbone, which was trained on a data set for which only 3.75% of the data were from African American patients (*n* = 1204), while 42.0% of the data within EMBED were from African American patients (*n* = 34 369). This could be improved by retraining Mirai and/or AsymMirai on a more diverse data set. Third, our study did not consider observed examples of confounding errors caused by major misalignment, such as from poor patient positioning. While further study of this marker will require registration, no registration was performed during preprocessing for Mirai, and doing so in our study would have confounded the comparison. AsymMirai's interpretable results clarified the importance of alignment as a confounder. Finally, when analyzing location consistency, patients needed two consecutive examinations followed by *n* years of follow-up, which reduced the size of our patient sets with known outcomes. For instance, only 10.7% of patients with a 40% or lower window shift had 5 years of subsequent follow-up. Future work could address this issue while also evaluating the generalizability of AsymMirai by using other institutional data sets.

In conclusion, localized bilateral dissimilarity, an imaging marker for breast cancer risk, approximated the predictive power of Mirai and was a key to Mirai's reasoning.

Table 2: AsymMirai and Mirai Subgroup Performance Analysis

Parameter and Model	1-year AUC	2-year AUC	3-year AUC	4-year AUC	5-year AUC
All patients, AM	0.79 (0.73, 0.85)	0.69 (0.65, 0.73)	0.68 (0.65, 0.71)	0.67 (0.64, 0.70)	0.66 (0.63, 0.69)
All patients, Mirai	0.84 (0.79, 0.89)	0.74 (0.70, 0.78)	0.72 (0.69, 0.76)	0.72 (0.69, 0.75)	0.71 (0.68, 0.74)
P value*	.002	<.001	<.001	<.001	<.001
Age at examination (y)					
< 50 AM	0.81 (0.68, 0.94)	0.68 (0.57, 0.79)	0.69 (0.60, 0.78)	0.67 (0.59, 0.76)	0.65 (0.58, 0.73)
< 50 Mirai	0.85 (0.73, 0.96)	0.73 (0.63, 0.83)	0.71 (0.61, 0.80)	0.68 (0.59, 0.77)	0.66 (0.58, 0.73)
P value*	.16	.10	.67	.88	.96
50–70 AM	0.76 (0.67, 0.84)	0.68 (0.63, 0.74)	0.67 (0.62, 0.71)	0.66 (0.62, 0.70)	0.65 (0.61, 0.69)
50–70 Mirai	0.84 (0.76, 0.91)	0.75 (0.70, 0.80)	0.73 (0.69, 0.77)	0.73 (0.69, 0.76)	0.72 (0.68, 0.75)
P value*	.002	<.001	<.001	<.001	<.001
> 70 AM	0.83 (0.77, 0.90)	0.66 (0.59, 0.73)	0.64 (0.59, 0.70)	0.61 (0.56, 0.66)	0.62 (0.57, 0.66)
> 70 Mirai	0.83 (0.75, 0.92)	0.68 (0.60, 0.75)	0.67 (0.62, 0.73)	0.67 (0.62, 0.72)	0.67 (0.62, 0.72)
P value*	.95	.43	.11	.002	.002
Race					
African American, AM	0.73 (0.64, 0.83)	0.64 (0.58, 0.70)	0.63 (0.58, 0.68)	0.63 (0.58, 0.67)	0.61 (0.57, 0.66)
African American, Mirai	0.82 (0.74, 0.89)	0.70 (0.64, 0.76)	0.69 (0.64, 0.74)	0.69 (0.64, 0.73)	0.68 (0.64, 0.72)
P value*	<.001	.004	.007	<.001	.002
White, AM	0.84 (0.77, 0.92)	0.73 (0.68, 0.78)	0.73 (0.69, 0.77)	0.71 (0.68, 0.75)	0.70 (0.67, 0.74)
White, Mirai	0.89 (0.83, 0.95)	0.78 (0.73, 0.83)	0.77 (0.73, 0.81)	0.76 (0.72, 0.80)	0.74 (0.71, 0.78)
P value*	.07	.004	.007	<.001	.002
Other, AM	0.75 (0.48, 1.02)	0.64 (0.46, 0.82)	0.59 (0.45, 0.73)	0.59 (0.45, 0.73)	0.57 (0.44, 0.70)
Other, Mirai	0.60 (0.22, 0.97)	0.64 (0.44, 0.83)	0.60 (0.45, 0.76)	0.61 (0.46, 0.76)	0.62 (0.49, 0.76)
P value*	.09	.94	.79	.61	.25

Note.—Data in parentheses are 95% CIs. AM = AsymMirai, AUC = area under the receiver operating characteristic curve. AUCs were calculated against the validation data set. Subgroups were chosen to match Mirai's reported subgroup performance in Yala et al (14). Age and race data were collected from the electronic health records used to construct the EMory BrEast imaging Dataset (EMBED). The "other" subgroup of "race" included 2763 patients who had a reported race of American Indian or Alaskan Native, Asian, multiple, Native Hawaiian or Pacific Islander, or unknown.

* *P* values for each pairwise comparison.

AsymMirai, a simplified deep learning bilateral dissimilarity-based model, performed similarly to the state-of-the-art black box model, Mirai, for 1–5-year breast cancer risk prediction. This observation agrees with the clinical importance of asymmetry and, as a result, highlights the potential of bilateral dissimilarity as a future imaging marker for breast cancer risk.

Acknowledgment: We are grateful to the Mirai team for releasing their code and model to the public, making this work possible.

Author contributions: Guarantors of integrity of entire study, **J.D., L.M., J.L.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **J.D., L.M., A.J.B., J.L., C.R.**; clinical studies, **F.S.**; experimental studies, **J.D., L.M., A.J.B., J.L.**; statistical analysis, **J.D., L.M., A.J.B., J.L., C.R.**; and manuscript editing, all authors

Disclosures of conflicts of interest: J.D. Duke University received NSF grant FW-HTF-R: Interpretable Machine Learning for Human-Machine Collaboration in High Stakes Decisions in Mammography 09/01/2022-08/31/2026; NSF HRD-2222336 Center for Virtual Imaging Trials. NIH/NIBIB 4/01/21-12/31/25; MedX: Interpretable Deep Learning Models for Better Clinician-AI Communication in Clinical Mammography 11/1/2021-10/31/2022 (internal Duke Award); dataset from iCAD for another project. **L.M.** Duke University received NSF grant FW-HTF-R: Interpretable Machine Learning for Human-Machine Collaboration in High Stakes Decisions in Mammography 09/01/2022-08/31/2026; NSF HRD-2222336 Center for Virtual Imaging Trials. NIH/NIBIB 4/01/21-12/31/25; MedX: Interpretable Deep Learning Models for Better Clinician-AI Communication in Clinical Mammography 11/1/2021-10/31/2022 (internal Duke Award); dataset from iCAD. **A.J.B.** Duke University received SF grant FW-HTF-R: Interpretable Machine Learning for Human-Machine Collaboration in High Stakes Decisions in Mammography 09/01/2022-08/31/2026; NSF HRD-2222336 Center for Virtual Imaging Trials. NIH/NIBIB 4/01/21-12/31/25; MedX: Interpretable Deep Learning Models for Better Clinician-AI Communication in Clinical Mammography 11/1/2021-10/31/2022 (internal Duke Award); payment or honoraria for lectures from Duke University; stock options from Nvidia, Google. **H.T.** No relevant relationships. **F.S.** Duke University received NSF grant FW-HTF-R: Interpretable Machine Learning for Human-Machine Collaboration in High Stakes Decisions in Mammography 09/01/2022-08/31/2026; NSF HRD-2222336 Center for Virtual Imaging Trials. NIH/NIBIB 4/01/21-12/31/25; MedX: Interpretable Deep Learning Models for Better Clinician-AI Communication in Clinical Mammography 11/1/2021-10/31/2022 (internal Duke Award); iCAD provided image data for training/testing. **J.L.** Duke University received NSF grant FW-HTF-R: Interpretable Machine Learning for Human-Machine Collaboration in High Stakes Decisions in Mammography 09/01/2022-08/31/2026; NSF HRD-2222336 Center for Virtual Imaging Trials. NIH/NIBIB 4/01/21-12/31/25; MedX: Interpretable Deep Learning Models for Better Clinician-AI Communication in Clinical Mammography 11/1/2021-10/31/2022 (internal Duke Award); royalties/license from iCAD; patent pending for mammography CAD. **C.R.** Duke University received NSF grant FW-HTF-R: Interpretable Machine Learning for Human-Machine Collaboration in High Stakes Decisions in Mammography 09/01/2022-08/31/2026; NSF HRD-2222336 Center for Virtual Imaging Trials. NIH/NIBIB 4/01/21-12/31/25; MedX: Interpretable Deep Learning Models for Better Clinician-AI Communication in Clinical Mammography 11/1/2021-10/31/2022 (internal Duke Award); dataset from iCAD.

References

1. Geis JR, Brady AP, Wu CC, et al. Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. Can Assoc Radiol J 2019;70(4):329–334.
2. Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. U.S. Food and Drug Administration Web site. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device. Published September 22, 2021. Accessed June 30, 2023.
3. Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Office for Official Publications of the European Communities Luxembourg Web site. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206. Published April 21, 2021. Accessed June 30, 2023.
4. Chen V, Liao QV, Vaughan JW, Bansal G. Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. In: Proceedings of the ACM on Human-Computer Interaction. 2023;7(CSCW2). https://dl.acm.org/doi/pdf/10.1145/3610219. Published October 4, 2023. Accessed October 10, 2023.
5. Antun V, Renna F, Poon C, Adcock B, Hansen AC. On instabilities of deep learning in image reconstruction and the potential costs of AI. Proc Natl Acad Sci USA 2020;117(48):30088–30095.
6. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989;81(24):1879–1886.
7. Lee A, Mavaddat N, Wilcox AN, et al. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. Genet Med 2019;21(8):1708–1718 [Published correction appears in Genet Med 2019;21(6):1462.].
8. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Stat Med 2004;23(7):1111–1130.
9. Arasu VA, Habel LA, Achacoso NS, et al. Comparison of Mammography AI Algorithms with a Clinical Risk Model for 5-year Breast Cancer Risk Prediction: An Observational Study. Radiology 2023;307(5):e222733.
10. Tosteson AN, Stout NK, Fryback DG, et al. Cost-effectiveness of digital mammography breast cancer screening. Ann Intern Med 2008;148(1):1–10.
11. Salz T, Richman AR, Brewer NT. Meta-analyses of the effect of false-positive mammograms on generic and specific psychosocial outcomes. Psychooncology 2010;19(10):1026–1034.
12. Bond M, Pavey T, Welch K, et al. Systematic review of the psychological consequences of false-positive screening mammograms. Health Technol Assess 2013;17(13):1–170, v–vi.
13. Yala A, Mikhael PG, Strand F, et al. Toward robust mammography-based models for breast cancer risk. Sci Transl Med 2021;13(578):eaba4373.
14. Yala A, Mikhael PG, Strand F, et al. Multi-institutional validation of a mammography-based breast cancer risk model. J Clin Oncol 2022;40(16):1732–1740.
15. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
16. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1(5):206–215.
17. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. In: Advances in Neural Information Processing Systems 31 (NeurIPS 2018). https://papers.nips.cc/paper_files/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.
18. Arun N, Gaw N, Singh P, et al. Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. Radiol Artif Intell 2021;3(6):e200267.
19. Mohamed A, Fakhry S, Basha T. Bilateral Analysis Boosts the Performance of Mammography-based Deep Learning Models in Breast Cancer Risk Prediction. Annu Int Conf IEEE Eng Med Biol Soc 2022;2022:1440–1443.
20. Tan M, Zheng B, Leader JK, Gur D. Association Between Changes in Mammographic Image Features and Risk for Near-Term Breast Cancer Development. IEEE Trans Med Imaging 2016;35(7):1719–1728.
21. Jeong JJ, Vey BL, Bhimireddy A, et al. The EMory BrEast imaging Dataset (EMBED): A racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. Radiol Artif Intell 2023;5(1):e220047.
22. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–845.
23. Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. Malawi Med J 2012;24(3):69–71.
24. Tan M, Zheng B, Ramalingam P, Gur D. Prediction of near-term breast cancer risk based on bilateral mammographic feature asymmetry. Acad Radiol 2013;20(12):1542–1550.
25. Sun W, Zheng B, Lure F, et al. Prediction of near-term risk of developing breast cancer using computerized features from bilateral mammograms. Comput Med Imaging Graph 2014;38(5):348–357.