MamaDino: A Hybrid Vision Model for Breast Cancer 3-Year Risk Prediction

Ruggiero Santeramo[1] RUGGIERO.SANTERAMO@FHT.ORG
Igor Zubarev[1] IGOR.ZUBAREV@FHT.ORG
Florian Jug[1] FLORIAN.JUG@FHT.ORG
[1] *Fondazione Human Technopole, Milan, Italy*

Editors: Under Review for MIDL 2026

Abstract

Breast cancer screening programmes increasingly seek to move from one-size-fits-all interval to risk-adapted and personalized strategies. The advent of deep learning (DL) gave birth to a wave of image-based risk models, able to provide more accurate short- to medium-term risk (1-5 years), compared with traditional risk models. Existing image-based risk models, such as Mirai, achieve strong discrimination but typically rely on convolutional backbones, ultra-high-resolution inputs and relatively simple multi-view fusion, with limited explicit modelling of contralateral asymmetry. We hypothesised that combining complementary inductive biases (convolutional and transformer-based) with explicit contralateral asymmetry modelling would allow us to match state-of-the-art 3-year risk prediction performance even when operating on substantially lower-resolution mammograms, indicating that using less detailed images in a more structured way can recover state-of-the-art accuracy. In this work, we present a Mammography-Aware Multi-view Attentional DINO-based model: MamaDino. MamaDino is a hybrid network that fuses frozen self-supervised DINOv3 (ViT-S) features with a trainable CNN encoder at 512×512 resolution and aggregates left-right breast information via a BilateralMixer to predict a 3-year breast cancer risk score. We train on 53,883 women from OPTIMAM, a UK cohort, and evaluate on matched 3-year case-control cohorts: an in-distribution test set from four UK screening sites and an external out-of-distribution test set from an unseen site. At breast level granularity MamaDino matched Mirai 3-year risk prediction both on the internal and external test sets while using $\sim 13\times$ fewer input pixels. Adding the BilateralMixer, MamaDino achieved an AUC of 0.736 (*vs.* Mirai's 0.713) on the in-distribution test set and 0.677 (*vs.* 0.666) on the external test set, showing consistent quality results across age, ethnicity, scanner, tumour type, and grade. These findings demonstrate that explicit contralateral modelling and complementary inductive biases enable predictions that match Mirai, despite operating on substantially lower-resolution mammograms.

Keywords: Breast Cancer, Deep Learning, Risk Prediction, DINOv3, Vision Transformers, Hybrid Networks

1. Introduction

Breast cancer is one of the most common cancers and a leading cause of death among women worldwide, despite improvements in systemic treatment and widespread mammography screening (Sung et al., 2021). Population screening has reduced mortality, but most programs still apply uniform interval schedules that do not reflect the large variation in individual risk. Early detection and accurate risk stratification are essential if screening is

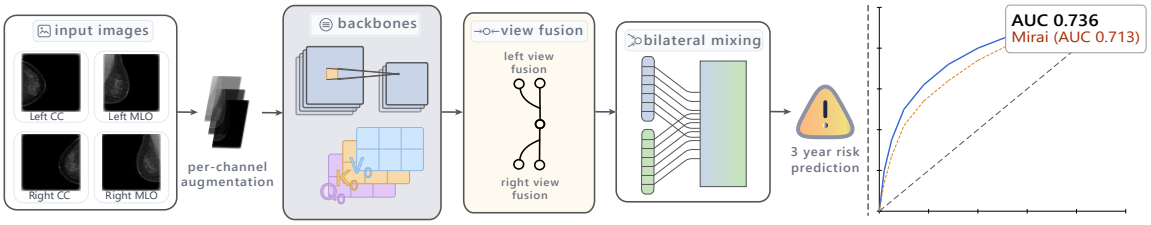

Figure 1: MamaDino overview: four standard mammography views are processed through a hybrid CNN–DINOv3 backbone with per-channel augmentation and bilateral mixing to produce a 3-year breast cancer risk score, achieving higher AUC than Mirai while operating at substantially lower image resolution.

to evolve towards more personalized intervals and preventive strategies (Pashayan et al., 2020; Brentnall et al., 2023).

Risk-adapted screening has traditionally relied on clinical and demographic scores that combine age, family history, reproductive and hormonal factors to estimate long-term risk, but these tools offer only moderate discrimination and largely ignore the rich morphologic and texture information present in screening mammograms. In contrast, deep learning (DL) models that operate directly on full-field digital mammograms (FFDM), such as Mirai (Yala et al., 2021), achieve strong short- to medium-term risk prediction and have been externally validated in multiple settings (Ellis et al., 2024). Recent work on the "Better CAD, Better Risk" paradigm further suggests that high-performing cancer detection systems can be repurposed as accurate fixed-horizon risk models, linking improved detection of early or subtle cancers to better near-term risk estimates (Santeramo et al., 2024).

From a modelling perspective, todays best-performing image-based risk models use rather high-resolution inputs (often >1M pixels per view) in mostly CNN-based backbones. At the same time, self-supervised vision transformers, such as DINOv3 (Siméoni et al., 2025), have recently emerged as powerful general-purpose encoders that have been shown to transfer well to mammography data (Manigrasso et al., 2025). The promise is that DINOv3 features can provide general purpose visual priors without requiring the training of task-specific feature generators.

This raises a central question: is the quality of short-term risk prediction dependent on high-resolution mammograms, or can equivalent performance be achieved by using higher quality feature extractors combined with the right predictive framework? We address this question in the setting of 3-year fixed-horizon breast cancer risk prediction from routine screening FFDM, following the "Better CAD, Better Risk" view that high-quality image-based models, like DINOv3, can provide accurate risk estimates (Santeramo et al., 2024).

Our main hypothesis is that a resolution-efficient hybrid vision model, which combines complementary CNN and transformer inductive biases with explicit bilateral reasoning, can match or exceed the 3-year risk prediction performance of established high-resolution CNN-based approaches such as Mirai, while using substantially fewer pixels and remaining robust across imaging sites and scanner vendors.

2. Related Work

Classical and mammography-based risk models. Classical risk-adapted screening relies on scores such as Gail (Gail et al., 1989), Claus (Claus et al., 1991) and Tyrer–Cuzick/IBIS (Tyrer et al., 2004), which combine age, family history, reproductive and hormonal factors, and sometimes genetic information to estimate long-term risk (Pashayan et al., 2020). These tools do not directly exploit screening mammograms and offer only moderate discrimination. Mammographic density and parenchymal texture partially bridge this gap, with quantitative descriptors improving discrimination over purely clinical models (Brentnall et al., 2015). DL-based mammography risk models operating directly on FFDMs further increase discrimination and calibration compared with Tyrer–Cuzick (Yala et al., 2019, 2021; Omoleye et al., 2023), enabling imaging-enriched and image-only risk stratification.

Deep learning for mammographic risk prediction. Deep neural networks trained on FFDMs now achieve state-of-the-art short- and medium-term risk prediction. Mirai (Yala et al., 2021), a multiview model predicting five-year risk, improves upon Tyrer–Cuzick and earlier DL approaches on several settings. Subsequent work combined image-based scores with clinical covariates and revisited mammographic risk prediction with modern CNNs and larger cohorts (Lauritzen et al., 2023; Ellis et al., 2024). Deep learning CAD systems trained for detection can also act as strong short-term risk predictors, suggesting a route to repurpose detection models for risk stratification (Santeramo et al., 2024). However, these systems typically use very high input resolutions and rely solely on convolutional backbones, narrowing the range of inductive biases they bring to the task, which may leave important global markers of risk under-modelled.

Bilateral and contralateral architectures. Bilateral reasoning is a natural inductive bias for paired organs such as the breasts (Roychoudhuri et al., 2006), and contralateral comparison is central to radiological assessment (Alterson and Plewes, 2003). Recent deep models encode this explicitly: contralateral attention improves detection over single-breast baselines (Mohamed et al., 2022), while DisAsymNet (Wang et al., 2023) and STA-Risk (Zhou et al., 2025) contrast left–right embeddings to capture inter-breast asymmetry for detection and risk prediction. Unlike prior single-scale CNNs with late contralateral fusion, our BilateralMixer (patient-level head) uses hybrid-encoder embeddings that decompose symmetric, asymmetric and interaction terms to capture subtle pre-diagnostic changes over three years.

Hybrid vision models and self-supervised transformers. Hybrid convolution and transformer architectures combine the local inductive biases of CNNs with the global receptive field of vision transformers, offering a favourable bias–capacity trade-off for fine-scale texture tasks compared with pure transformers (Dai et al., 2021; d'Ascoli et al., 2021; Goyal and Bengio, 2022). Channel-wise gating mechanisms such as squeeze-and-excitation and Hadamard-product gating further refine this bias by reweighting feature channels based on global context (Hu et al., 2018b,a; Duke and Taylor, 2018). Self-supervised vision transformers like DINOv3 (Siméoni et al., 2025) learn semantically meaningful representations from large natural-image corpora and transfer well to medical imaging (Gao et al., 2025; Liu et al., 2025), and ViT-based mammography models at reduced resolution can recover much of the performance of high-resolution CNNs at lower computational cost (Manigrasso

et al., 2025). Most prior work, however, either fine-tunes such transformers end-to-end or uses them as single-stream backbones, without combining them with texture-specialised encoders or explicitly modelling bilateral context. Here we keep a DINOv3 branch frozen as a global prior, condition a trainable SE-ResNeXt branch via cross-attentional fusion, and couple both with a bilateral mixing head at 512×512 for three-year risk prediction.

3. Methods

3.1. Dataset and Cohort Selection

Data were sourced from the OPTIMAM Mammography Image Database (Halling-Brown et al., 2020) and comprised women attending routine FFDM screening at four UK services (Jarvis, Leicester, Imperial, St George's) between 2010 and 2021. All exams were standard two-view bilateral screenings (L-CC, R-CC, L-MLO, R-MLO) using only *FOR-PRESENTATION* images. Women aged 40–80 years were included, and exams with missing views or implants were excluded. All data were anonymised.

After filtering, the training cohort comprised **53,883** women. Approximately 500 developed biopsy-confirmed cancer during the screening interval (CI), $\sim$ 1,500 had benign findings (B), $\sim$ 11,000 had malignant or suspicious recalls (M), and the remainder were normal (N). Many women contributed multiple examinations (mean 2.57 episodes per patient); supervision at *episode level* used the most severe bilateral outcome, while lesion-side OPTIMAM annotations provided breast-level labels. During training, episodes with CI/M/B outcomes and their screening exams in the preceding 3 years (CIP, MP) were labelled positive at both episode and breast level using lesion laterality, while normal (N) episodes and contralateral normal breasts in unilateral M episodes were labelled negative.

Imaging was predominantly acquired on Hologic systems ($\sim$ 96%), with the remaining $\sim$ 4% from GE, Siemens and Philips devices. Self-reported ethnicity was available at the patient level but substantially missing and unevenly distributed (Tab. A1), so it was used only for fairness assessment, not as a model input.

In-distribution validation cohort. The internal held-out cohort was a 1:2 matched case–control sample drawn from the four UK screening services used for training (Imperial, Jarvis, Leicester, St George's), including women aged 47–73 years (UK screening age being 50-70). Matching was performed at the *patient level* on screening site, age, and imaging manufacturer (Hologic, Siemens, Philips); cases were defined as screening examinations acquired three years before a subsequently diagnosed breast malignancy (including interval cancers), and controls were women with normal outcomes at their subsequent screening. The final internal test set comprised 525 cases and 1,050 matched controls, with detailed distributions of age, self-reported ethnicity, imaging site, scanner type, cancer type (DCIS vs. invasive), and tumour grade reported in Tab. A2.

Out-of-distribution validation cohort. The external cohort consisted of women from the Oxford screening service, which was not used in model development, and used the same age range (47–73 years) as the internal cohort. We constructed a 4:1 matched case–control sample, matching on age and scanner manufacturer (Hologic, GE Medical Systems); the final external test set comprised 376 cases and 1,504 matched controls. Scanner distribution strongly differed from the training and internal cohorts, with GE Medical Systems

accounting for roughly one third of examinations, and racial/ethnic information was largely unavailable, precluding ethnicity-based stratification. Summary characteristics for this cohort, including cancer type and tumour grade for malignancies diagnosed three years after the baseline examination, are reported in Tab. A2.

3.2. Per-channel augmentation

To exploit ImageNet-pretrained backbones with single-channel mammograms, we construct a pseudo-RGB image by applying distinct intensity transforms to a single greyscale view. Given a preprocessed mammogram $I \in \mathbb{R}^{H\times W}$, we define three channels $I^{(1)}, I^{(2)}, I^{(3)} \in \mathbb{R}^{H\times W}$. Channel $I^{(1)}$ undergoes a random brightness jitter to promote robustness to exposure, breast thickness, and global intensity scaling; $I^{(2)}$ receives an independent, random contrast jitter to mitigate vendor- and site-specific windowing while preserving relative tissue and lesion contrast; $I^{(3)}$ is processed with contrast-limited adaptive histogram equalisation (CLAHE) using relatively high clip limits and a fine grid (e.g. 12×12 tiles) to enhance subtle low-contrast structures such as early masses or microcalcifications.

We compare this scheme against naive three-channel replication across input resolutions in Fig. A1.

3.3. Model Architecture

We design a hybrid fusion network that couples a frozen DINOv3 vision transformer with a trainable SE-ResNeXt101 backbone to exploit complementary inductive biases. For each breast $b \in \{\mathrm{L}, \mathrm{R}\}$, the input consists of the two standard views (CC and MLO), stacked into a tensor $x^{(b)} \in \mathbb{R}^{2\times 3\times H\times W}$.

Backbone encoders and cross-attentional fusion The transformer branch is a frozen DINOv3 ViT-S/16+ model that produces a grid of patch tokens for each view, while a view-specific SE-ResNeXt101 encoder (same architecture, untied weights) extracts convolutional texture features. SE blocks modulate channel responses to emphasise discriminative patterns. We resize DINO and ResNeXt feature maps to a common 16×16 grid, project them to a shared latent dimension d (here $d = 512$) with 1×1 convolutions, and apply multi-head cross-attention in which ResNeXt tokens query DINO tokens, followed by residual feed-forward layers. This allows the convolutional stream to integrate globally coherent, semantically rich transformer features while preserving high-frequency CNN information.

BridgeMixer block and breast embedding. On top of the SE-ResNeXt backbone, a SE-based BridgeMixer projects ResNeXt features through a bottleneck to d channels, concatenates this projection with the cross-attention output for each view, and compresses the result back to d channels via a 1×1 convolution. This yields fused view-specific feature maps $F^{\text{view}} \in \mathbb{R}^{d\times 16\times 16}$ for CC and MLO. We concatenate the two views along the channel dimension to obtain a joint breast representation $F^{(b)} \in \mathbb{R}^{2d\times 16\times 16}$, apply adaptive average pooling to 2×2, and flatten the result into an $8d$-dimensional per-breast embedding.

Breast-level classifier and patient-level baseline. For each breast, the embedding is passed through a two-layer MLP with LayerNorm, GELU and dropout to produce a logit $z^{(b)}$, trained with a breast-level binary focal loss. At inference we obtain probabilities $p^{(\mathrm{L})}$

and $p^{(\mathrm{R})}$ and define patient risk as $p_{\text{patient}} = \max\big(p^{(\mathrm{L})}, p^{(\mathrm{R})}\big)$, assuming malignancy in either breast determines patient risk; this max aggregation provides the baseline for the bilateral fusion head.

The BilateralMixer: Patient-level prediction A key ingredient to the method we present is the BilateralMixer at patient level: a fusion module that combines left and right breast embeddings via a shallow transformer, a learned asymmetry gate, and symmetric feature composition to capture bilateral context, asymmetry, and concordance.

Let $\mathbf{e}_L, \mathbf{e}_R \in \mathbb{R}^d$ denote the left and right breast embeddings from the per-side encoder. The BilateralMixer combines them in three stages. First, *relational encoding via a symmetric transformer*: the embeddings are treated as tokens in a compact transformer encoder sequence $[\mathrm{CLS}, \mathbf{e}_L, \mathbf{e}_R]$, where a learnable classification token (CLS) aggregates bilateral context through self-attention and yields a patient-level embedding $\mathbf{c}$. This symmetric design does not privilege either side and is permutation invariant with respect to laterality.

Second, *gated asymmetry weighting*: a small multilayer perceptron computes soft attention coefficients (α_L, α_R) from the concatenation of both embeddings and their absolute difference,

$$[\alpha_L, \alpha_R] = \mathrm{softmax}\big(f_\theta([\mathbf{e}_L, \mathbf{e}_R, |\mathbf{e}_L - \mathbf{e}_R|])\big), \tag{1}$$

so that the model can emphasize the side with stronger pathological evidence while retaining bilateral context.

Third, *symmetric feature composition*: the final representation concatenates four components,

$$\mathbf{z} = [\mathbf{c},\ \alpha_L \mathbf{e}_L + \alpha_R \mathbf{e}_R,\ |\mathbf{e}_L - \mathbf{e}_R|,\ \mathbf{e}_L \odot \mathbf{e}_R], \tag{2}$$

where $\odot$ denotes elementwise multiplication. The absolute difference encodes order-invariant asymmetry, and the product captures bilateral concordance as a feature-wise co-activation signal.

The fused vector $\mathbf{z}$ is passed to a shallow multilayer perceptron that outputs a single patient-level malignancy logit, providing a fully differentiable analogue of the radiologist's comparison of contralateral mammograms.

3.4. Training and Evaluation

Training used two stages: first learning breast-level representations, then patient-level aggregation. In stage one (Fig. 2.a), we trained the hybrid fusion encoder on the large screening cohort using 512×512 mammograms with the per-channel and geometric/photometric augmentations (Sec. 3.2), while validation and test images underwent only deterministic resizing per-channel augmentation and normalisation. The DINOv3 branch was initialised from a self-supervised ViT-S/16+ checkpoint on natural images and kept frozen; the SE-ResNeXt101 branch, initialised from ImageNet, was trained jointly with the cross-attention fusion modules to predict breast-level malignancy. Optimisation used AdamW with a binary focal loss to address class imbalance and early stopping based on breast-level AUC on a held-out validation subset.

In stage two (Fig. 2.a), we froze the fusion encoder and trained only the BilateralMixer (Fig. 2.c) head for patient-level prediction on the same training set, using the same optimiser and loss with early stopping on patient-level AUC. All hyperparameters (learning rates,

regularisation, number of epochs) were tuned on the internal training/validation split and fixed before evaluation on the matched in-distribution and external OOD test cohorts.

For comparison, we trained three baselines under the same preprocessing and optimisation pipeline: (i) a DINO-only model with a linear classifier on frozen transformer features, (ii) a SE-ResNeXt-only model trained directly on 512×512 mammograms, and (iii) a hybrid breast-level fusion model with max aggregation across breasts at inference (patient risk $= \max(p^{(\mathrm{L})}, p^{(\mathrm{R})})$). Mirai predictions were obtained from the released model without retraining, using 3-year risk estimates. Subgroup AUCs were computed for pre-specified strata of age, self-reported ethnicity, tumour type and grade, and imaging manufacturer, and reported only where sample sizes were adequate.

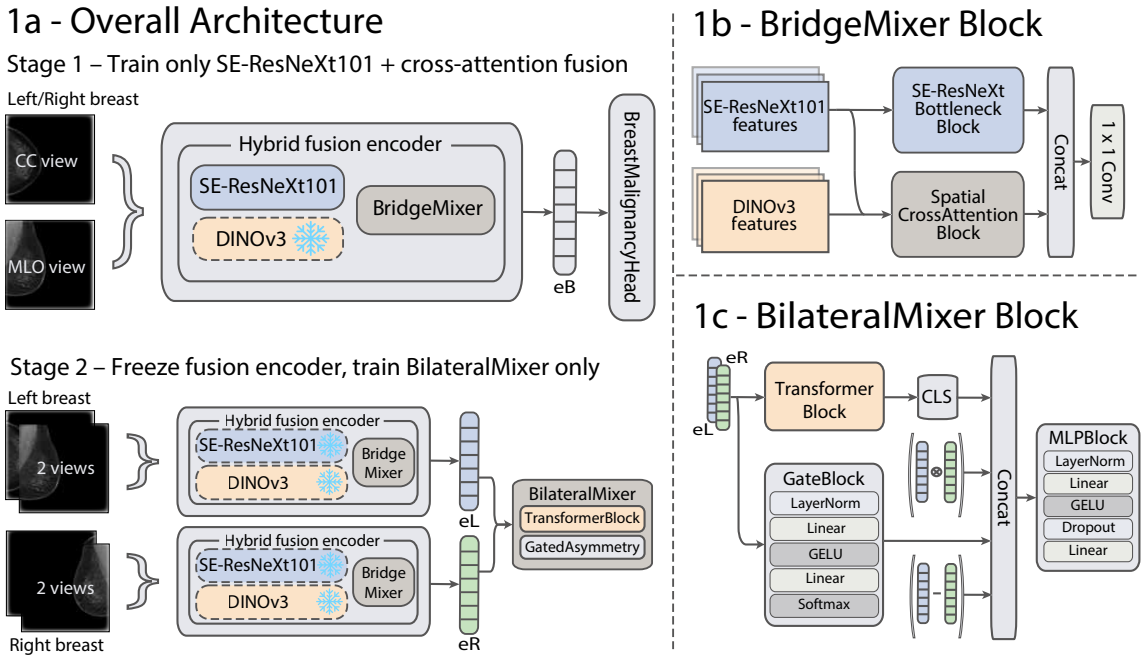

Figure 2: **MamaDino architecture: (a)** Four standard mammography views per exam (at 512×512 resolution) are processed by a hybrid fusion encoder that combines a frozen DINOv3 for global semantics with a trainable SE-ResNeXt101 CNN for local texture, producing per-breast embeddings that are fused for 3-year malignancy risk prediction in **Stage 2** via the BilateralMixer. **(b)** The BridgeMixer block aligns Transformer tokens with convolutional feature maps via spatial cross-attention and 1×1-convolution fusion. **(c)** The BilateralMixer block takes left and right breast embeddings, models bilateral concordance and asymmetry through a transformer, and outputs the final risk score.

4. Results

4.1. Prediction of 3-year Cancer Risk

In Tab. 1, we report 3-years fixed-horizon risk discrimination results on the matched in-distribution and external OOD cohorts. On the in-distribution test set, single-stream baselines (DINO-only and SEResNeXt-only) were less discriminative than any hybrid variant.

Table 1: **Overall AUC for 3-year breast cancer risk prediction** on in-distribution and out-of-distribution test sets, comparing single-stream baselines, Mirai and MamaDino variants (breast and patient level).

		In-Distribution Test Set		OOD Test Set	
Model	Resolution	AUC	(95% CI)	AUC	(95% CI)
DINO-only*	512×512	0.621	–	–	–
SEResNeXt-only*	512×512	0.668	–	–	–
Mirai† (Yala et al., 2021)	1664×2048	0.713	(0.686–0.740)	0.676	(0.646-0.707)
MamaDino * (ours)	512×512	0.727	(0.701–0.754)	0.666	(0.635-0.696)
MamaDino † (ours)	512×512	**0.736**	(0.710–0.762)	**0.677**	(0.647-0.707)

* Predictions were at breast-level, and patient risk corresponded to the max of the two breast scores.

† Predictions were at patient-level.

Mirai achieved an in-distribution AUC of 0.713 and an OOD AUC of 0.666 at its native resolution of 1664 × 2048 pixels. In contrast, the breast-level MamaDino, (using max aggregation) reached an in-distribution AUC of 0.727 and an OOD AUC of 0.666 at 512 × 512, matching Mirai's OOD performance while using ∼13× fewer input pixels. Adding the BilateralMixer patient head further increased discrimination to an in-distribution AUC of 0.736 and an external OOD AUC of 0.677, slightly outperforming Mirai on both cohorts. Confidence intervals for MamaDino and Mirai overlapped on the external cohort, showing robustness to device manufacturer's shift.

4.2. Subgroup analysis

Tab. 2 reports patient-level AUCs stratified by age, ethnicity, imaging manufacturer, cancer type and tumour grade. On the in-distribution test cohort, the hybrid model showed stable discrimination across age, with AUCs between 0.70 and 0.75 and consistently, though modestly, higher values than Mirai. On the out-of-distribution (OOD) test cohort, both models achieved AUCs around 0.66–0.69 across age strata: Mirai performed better in women younger than 60 years, while the hybrid model matched or exceeded Mirai in the older groups.

Across ethnic groups in the in-distribution cohort, MamaDino and Mirai performed similarly in White women, with larger but less precisely estimated differences in Asian and Black women, where MamaDino and Mirai respectively achieved higher AUCs. Stratification by scanner vendor showed that, on Hologic systems, which contributed the majority of training exams, MamaDino slightly outperformed Mirai, while performance on Siemens devices was comparable. On GE scanners, MamaDino was consistently superior to Mirai on both the in-distribution and OOD test cohorts.

When stratified by cancer type, both models showed higher discrimination for invasive cancers than for DCIS. MamaDino improved over Mirai for invasive disease on both cohorts, with comparable performance for DCIS on the in-distribution test set and slightly worse performance on the OOD test cohort. By tumour grade, the hybrid model achieved AUCs that were generally equal to or higher than Mirai, with the largest gains for low-

Table 2: **Subgroup AUCs for 3-year breast cancer risk prediction** of MamaDino and Mirai on in-distribution and out-of-distribution (Oxford) test cohorts, stratified by age, ethnicity, scanner, cancer type and tumour grade. Values are AUC estimates with corresponding 95% confidence intervals; dashes indicate subgroups not represented or not evaluable.

Test Cohorts	**In-Distribution Test**		**OOD Test Set**	
	MamaDino	Mirai	MamaDino	Mirai
Age (y)				
< 60	**0.746** (0.710-0.783)	0.727 (0.690-0.764)	0.668 (0.626-0.710)	**0.693** (0.651-0.735)
60–65	**0.705** (0.650-0.760)	0.656 (0.597-0.715)	**0.675** (0.618-0.732)	0.657 (0.596-0.717)
65+	**0.747** (0.696-0.797)	0.743 (0.691-0.794)	**0.691** (0.624-0.758)	0.672 (0.606-0.737)
Ethnicity				
White	0.723 (0.689-0.758)	**0.726** (0.691-0.760)	–	–
Asian	**0.801** (0.723-0.880)	0.660 (0.556-0.764)	–	–
Black/African	0.770 (0.617-0.924)	**0.839** (0.728-0.950)	–	–
Scanner				
GE Med. S.	**0.725** (0.555-0.895)	0.627 (0.443-0.811)	**0.669** (0.618-0.721)	0.622 (0.566-0.677)
Hologic, Inc.	**0.740** (0.713-0.766)	0.722 (0.694-0.749)	0.679 (0.642-0.716)	**0.705** (0.669-0.741)
Siemens	**0.463** (0.217-0.709)	0.457 (0.222-0.691)	–	–
Cancer type**				
DCIS	0.696 (0.636-0.757)	**0.711** (0.650-0.771)	0.659 (0.586-0.731)	**0.689** (0.617-0.761)
Invasive	**0.748** (0.718-0.778)	0.713 (0.682-0.745)	**0.680** (0.647-0.714)	0.673 (0.640-0.707)
Tumour grade**				
G1	**0.755** (0.684-0.826)	0.686 (0.613-0.758)	**0.683** (0.621-0.745)	0.679 (0.614-0.745)
G2	**0.787** (0.751-0.824)	0.756 (0.717-0.796)	**0.695** (0.645-0.745)	0.660 (0.609-0.711)
G3	**0.641** (0.572-0.709)	0.633 (0.561-0.705)	0.664 (0.600-0.728)	**0.686** (0.623-0.749)

* Data are medians, with IQRs in parentheses.
** Cancer data from 3-year follow-up.

and intermediate-grade tumours and similar performance between models for high-grade disease.

5. Discussion

MamaDino achieved 3-year risk discrimination comparable to, and in some settings exceeding, Mirai despite operating at 512×512 resolution, *i.e.* using roughly 13× fewer input pixels. At breast level, the hybrid encoder with max aggregation matched Mirai's performance on the external cohort, and adding the BilateralMixer further improved patient-level AUC on both in-distribution and OOD test sets.

Clinically, these results suggest that short-term mammographic risk prediction does not intrinsically require high-resolution inputs if information is organised through stronger priors. MamaDino integrates global semantics from a frozen DINOv3 branch with fine-grained mammographic texture from a trainable SE-ResNeXt101 branch, fuses left–right embeddings via a bilateral mixing head, and uses per-channel augmentation to construct pseudo-RGB inputs from single-channel mammograms. This configuration appears sufficient to recover state-of-the-art performance at lower resolution. In the context of existing DL risk models such as Mirai and CNN-only bilateral architectures, our findings support the view that model design and fusion strategy can compensate for reduced pixel count.

The behaviour of MamaDino relative to Mirai across in-distribution and OOD cohorts highlights robustness to scanner shifts, particularly between Hologic and GE systems, consistent with the hypothesis that self-supervised transformers provide stable domain priors. Discrimination for DCIS remained lower than for invasive cancers, echoing prior reports and suggesting that early in situ changes yield weaker or more diffuse signatures at this resolution. Ethnicity-stratified analyses and Siemens-specific results were limited by sample size, and wide confidence intervals in these strata are more consistent with data scarcity than systematic failure of the model.

Limitations of this study include its restriction to UK screening services and a case–control design, which limits assessment of calibration under true population incidence and generalisability to settings with different screening intervals, prevalence or demographics. Siemens scanners and non-White groups were under-represented, constraining conclusions on device- and ancestry-specific performance. We also evaluated only a single 3-year horizon and did not retrain Mirai-like architectures at 512×512, so we cannot fully disentangle the respective contributions of resolution and architecture.

These limitations motivate future work on self-supervised pretraining directly on large-scale mammography or tomosynthesis, extension of the BilateralMixer to longitudinal and 3D data, evaluation in multinational cohorts with prospective impact assessment, and systematic study of ensembles combining MamaDino with Mirai or other image, clinical and genomic risk scores for personalised, risk-adapted screening.

6. Conclusion

We introduced MamaDino, a hybrid model that fuses frozen DINOv3 features, trainable SE-ResNeXt encoders and a bilateral fusion head for 3-year breast cancer risk prediction. At 512×512 resolution, using roughly 13× fewer pixels than Mirai, MamaDino matched or slightly exceeded its discrimination on in-distribution and OOD UK cohorts, with stable performance across demographics, tumour characteristics and scanner vendors. This suggests that complementary inductive biases and explicit contralateral modelling can reduce reliance on high-resolution mammograms.

These findings have broader implications for AI-enabled screening. Our findings argue that progress in mammographic risk prediction may hinge less on ever higher input resolution and more on how global context, local texture and bilateral relationships are represented and fused. MamaDino combines self-supervised transformer priors, convolutional texture encoders and a dedicated bilateral mixing head, showing that appropriately structured architectures can recover state-of-the-art performance even when the pixel budget is constrained. This shifts the emphasis from brute-force detail towards the design of task-aware, multi-view, hybrid models.

In summary, MamaDino shows that "using less detail in a more structured way" can recover state-of-the-art short-term risk prediction in mammography, pointing toward hybrid, and bilaterally aware architectures as a promising foundation for future AI systems for personalised breast cancer screening.

Data Availability

The images and data used in this publication are derived from the OPTIMAM imaging database (Halling-Brown et al., 2020), we would like to acknowledge the OPTIMAM project team and staff at the Royal Surrey NHS Foundation Trust who developed the OPTIMAM database, and Cancer Research UK who funded the creation and maintenance of the database.

References

Robert Alterson and Donald B Plewes. Bilateral symmetry analysis of breast mri. *Physics in Medicine & Biology*, 48(20):3431, 2003.

Adam R Brentnall, Elaine F Harkness, Susan M Astley, Louise S Donnelly, Paula Stavrinos, Sarah Sampson, Lynne Fox, Jamie C Sergeant, Michelle N Harvie, Mary Wilson, et al. Mammographic density adds accuracy to both the tyrer-cuzick and gail breast cancer risk models in a prospective uk screening cohort. *Breast Cancer Research*, 17(1):147, 2015.

Adam R Brentnall, Emma C Atakpa, Harry Hill, Ruggiero Santeramo, Celeste Damiani, Jack Cuzick, Giovanni Montana, and Stephen W Duffy. An optimization framework to guide the choice of thresholds for risk-based cancer screening. *NPJ Digital Medicine*, 6 (1):223, 2023.

Elisabeth B Claus, Neil Risch, and W Douglas Thompson. Genetic analysis of breast cancer in the cancer and steroid hormone study. *American journal of human genetics*, 48(2): 232, 1991.

Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. CoAtNet: Marrying convolution and attention for all data sizes. *arXiv preprint arXiv:2106.04803*, 2021. doi: 10.48550/arXiv.2106.04803. URL https://arxiv.org/abs/2106.04803.

Stéphane d'Ascoli, Hugo Touvron, Matthew Leavitt, Ari Morcos, Giulio Biroli, and Levent Sagun. ConViT: Improving vision transformers with soft convolutional inductive biases. *arXiv preprint arXiv:2103.10697*, 2021. doi: 10.48550/arXiv.2103.10697. URL https://arxiv.org/abs/2103.10697.

Brendan Duke and Graham W. Taylor. Generalized Hadamard-product fusion operators for visual question answering. In *15th Conference on Computer and Robot Vision (CRV)*, pages 39–46. IEEE, 2018. doi: 10.1109/CRV.2018.00016. URL https://doi.org/10.1109/CRV.2018.00016.

Sam Ellis, Sandra Gomes, Matthew Trumble, Mark D Halling-Brown, Kenneth C Young, Nouman S Chaudhry, Peter Harris, and Lucy M Warren. Deep learning for breast cancer risk prediction: application to a large representative uk screening cohort. *Radiology: Artificial Intelligence*, 6(4):e230431, 2024.

Mitchell H Gail, Louise A Brinton, David P Byar, Donald K Corle, Sylvan B Green, Catherine Schairer, and John J Mulvihill. Projecting individualized probabilities of developing

breast cancer for white females who are being examined annually. *JNCI: Journal of the National Cancer Institute*, 81(24):1879–1886, 1989.

Yifan Gao, Haoyue Li, Feng Yuan, Xiaosong Wang, and Xin Gao. Dino u-net: Exploiting high-fidelity dense features from foundation models for medical image segmentation. *arXiv preprint arXiv:2508.20909*, 2025.

Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266):20210068, 2022.

Mark D Halling-Brown, Lucy M Warren, Dominic Ward, Emma Lewis, Alistair Mackenzie, Matthew G Wallis, Louise S Wilkinson, Rosalind M Given-Wilson, Rita McAvinchey, and Kenneth C Young. Optimam mammography image database: a large-scale resource of mammography images and clinical data. *Radiology: Artificial Intelligence*, 3(1):e200103, 2020.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018a. doi: 10.1109/CVPR.2018.00745. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018b.

Andreas D Lauritzen, My C von Euler-Chelpin, Elsebeth Lynge, Ilse Vejborg, Mads Nielsen, Nico Karssemeijer, and Martin Lillholm. Assessing breast cancer risk by combining ai for lesion detection and mammographic texture. *Radiology*, 308(2):e230227, 2023.

Che Liu, Yinda Chen, Haoyuan Shi, Jinpeng Lu, Bailiang Jian, Jiazhen Pan, Linghan Cai, Jiayi Wang, Yundi Zhang, Jun Li, et al. Does dinov3 set a new medical vision standard? *arXiv preprint arXiv:2509.06467*, 2025.

Francesco Manigrasso, Rosario Milazzo, Alessandro Sebastian Russo, Fabrizio Lamberti, Fredrik Strand, Andrea Pagnani, and Lia Morra. Mammography classification with multi-view deep learning techniques: Investigating graph and transformer-based architectures. *Medical Image Analysis*, 99:103320, 2025.

Alaa Mohamed, Sherihan Fakhry, and Tamer Basha. Bilateral analysis boosts the performance of mammography-based deep learning models in breast cancer risk prediction. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1440–1443. IEEE, 2022.

Olasubomi J Omoleye, Anna E Woodard, Frederick M Howard, Fangyuan Zhao, Toshio F Yoshimatsu, Yonglan Zheng, Alexander T Pearson, Maksim Levental, Benjamin S Aribisala, Kirti Kulkarni, et al. External evaluation of a mammography-based deep learning model for predicting breast cancer in an ethnically diverse population. *Radiology: Artificial Intelligence*, 5(6):e220299, 2023.

Nora Pashayan, Antonis C Antoniou, Urska Ivanus, Laura J Esserman, Douglas F Easton, David French, Gaby Sroczynski, Per Hall, Jack Cuzick, D Gareth Evans, et al. Personalized early detection and prevention of breast cancer: Envision consensus statement. *Nature reviews Clinical oncology*, 17(11):687–705, 2020.

Rahul Roychoudhuri, Venkata Putcha, and Henrik Møller. Cancer and laterality: a study of the five major paired organs (uk). *Cancer causes & control*, 17(5):655–662, 2006.

Ruggiero Santeramo, Celeste Damiani, Jiefei Wei, Giovanni Montana, and Adam R Brentnall. Are better ai algorithms for breast cancer detection also better at predicting risk? a paired case–control study. *Breast Cancer Research*, 26(1):25, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

Jonathan Tyrer, Stephen W Duffy, and Jack Cuzick. A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in medicine*, 23(7):1111–1130, 2004.

Xin Wang, Tao Tan, Yuan Gao, Luyi Han, Tianyu Zhang, Chunyao Lu, Regina Beets-Tan, Ruisheng Su, and Ritse Mann. Disasymnet: Disentanglement of asymmetrical abnormality on bilateral mammograms using self-adversarial learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 57–67. Springer, 2023.

Adam Yala, Constance Lehman, Tal Schuster, Tally Portnoi, and Regina Barzilay. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*, 292(1):60–66, 2019.

Adam Yala, Peter G. Mikhael, Fredrik Strand, Gigin Lin, Kevin Smith, Yung-Liang Wan, Leslie Lamb, Kevin Hughes, Constance Lehman, and Regina Barzilay. Toward robust mammography-based models for breast cancer risk. *Science Translational Medicine*, 13 (578):eaba4373, 2021. doi: 10.1126/scitranslmed.aba4373. URL https://www.science.org/doi/10.1126/scitranslmed.aba4373. This paper introduces the Mirai model for multi-timepoint breast cancer risk prediction.

Zhengbo Zhou, Dooman Arefan, Margarita Zuley, Jules Sumkin, and Shandong Wu. Starisk: A deep dive of spatio-temporal asymmetries for breast cancer risk prediction. *arXiv preprint arXiv:2505.21699*, 2025.

Appendix A. Appendix A

Table A1: **Ethnic composition of the training cohort.** Self-reported ethnic backgrounds among women in the UK OPTIMAM training cohort, highlighting the predominance of White participants and the large proportion of missing ethnicity records.

Ethnicity	Number of Women
White	30,682
Asian	3,808
Black	1,851
Mixed	601
Other	1,107
Missing	15,834
Total	53,883

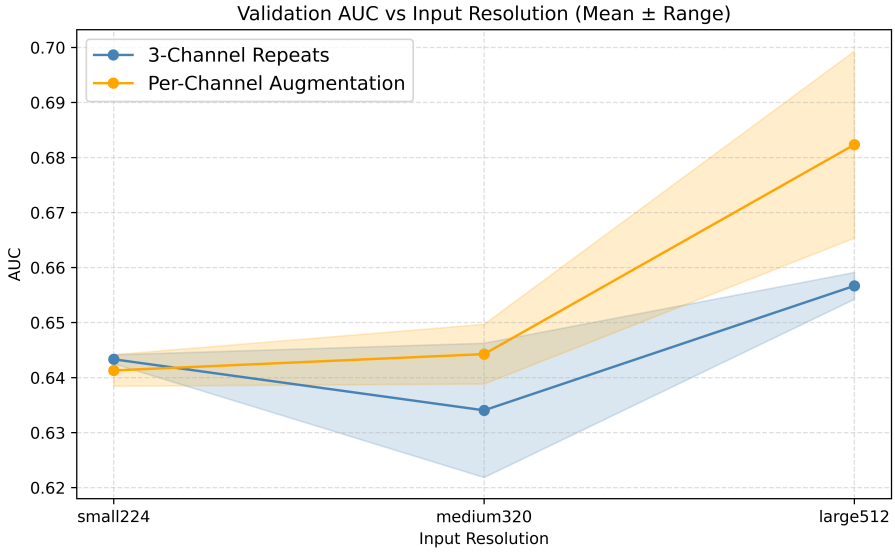

Figure A1: **Per-channel augmentation *vs*. simple replication ablation.** Validation AUC of the breast-level MamaDino encoder on a held-out OPTIMAM validation set as a function of input resolution (224, 320, 512). Curves show mean AUC across random initialisations (2 seed per setup), with shaded regions indicating the range across runs. The blue line ("3-Channel Repeats") corresponds to the standard practice of replicating the grayscale image into three identical channels before augmentation. The orange line ("Per-Channel Augmentation") corresponds to our proposed strategy, where each channel is independently perturbed (brightness/contrast jitter and CLAHE) before being recombined. Across all resolutions, per-channel augmentation yields consistently higher AUC, with the largest gain at 512×512.

Table A2: **Characteristics of the matched internal (in-distribution) and external (out-of-distribution) case–control test cohorts.** Values are counts with column-wise percentages. Age is reported both categorically and as median (IQR). Cancer type and tumour grade refer to the malignancy diagnosed three years after the baseline screening examination. The Screening site and Scanner rows highlight the distributional shift between the different cohorts: internal data come from four services dominated by Hologic systems, whereas the external Oxford cohort includes a much higher proportion of GE scanners.

Test Cohorts	**In-Distribution Test Set**		**OOD Test Set**	
	Control (n=1050)	Case (n=525)	Control (n=1504)	Case (n=376)
Age (y)				
< 60	532 (50.7%)	266 (50.7%)	728 (48.5%)	182 (48.4%)
60–65	260 (24.8%)	130 (24.8%)	419 (27.8%)	105 (27.9%)
65+	258 (24.6%)	129 (24.6%)	357 (23.8%)	89 (23.7%)
Median*	59 (54–64)	59 (54–64)	60 (55–64)	60 (55–64)
Ethnicity				
White	623 (59.3%)	306 (58.3%)	2 (0.1%)	0 (0.0%)
Asian	88 (8.4%)	43 (8.2%)		
Black/African	30 (2.9%)	18 (3.4%)		
Other/Not Stated	309 (29.4%)	158 (30.1%)	1502 (99.9%)	376 (100.0%)
Screening site				
Imperial	246 (23.4%)	123 (23.4%)		
Jarvis Breast Centre	280 (26.7%)	140 (26.7%)		
Leicester	242 (23.0%)	121 (23.0%)		
St. George's	282 (26.9%)	141 (26.9%)		
Oxford			1504 (100.0%)	376 (100.0%)
Scanner				
GE Medical Systems	26 (2.5%)	13 (2.5%)	484 (32.1%)	121 (32.2%)
Hologic, Inc.	1006 (95.8%)	503 (95.8%)	1020 (67.9%)	255 (67.8%)
Siemens	18 (1.7%)	9 (1.7%)		
Cancer type**				
DCIS		102 (19.4%)		62 (16.5%)
Invasive		400 (76.2%)		314 (83.5%)
Unknown		23 (4.4%)		0 (0.0%)
Tumour grade**				
G1		78 (14.9%)		79 (21.0%)
G2		215 (41.0%)		143 (38.0%)
G3		92 (17.5%)		88 (23.4%)
N/A		140 (26.7%)		66 (17.6%)

* Data are medians, with IQRs in parentheses.

** Cancer data from 3-year follow-up.