

Explicitly Aligned Multi-View Breast Cancer Risk Prediction from Variable-Length Screening Histories

Solveig Thrun^{a,*}, Zijun Sun^a, Suaiba A. Salahuddin^a, Kristoffer Wickstrøm^a, Elisabeth Wetzer^a,
Stine Hansen^b, Robert Jenssen^{a,c,d}, Michael Kampffmeyer^{a,c}

^a*Department of Physics and Technology, UiT The Arctic University of Norway, Tromsø, 9037, Norway*

^b*SPKI The Norwegian Centre for Clinical Artificial Intelligence, University Hospital of North Norway, Tromsø, 9019, Norway*

^c*Norwegian Computing Center, Oslo, 0373, Norway*

^d*Pioneer Centre for AI, University of Copenhagen, Copenhagen, 1350, Denmark*

Abstract

Accurate breast cancer risk prediction from screening mammography is essential for improved risk stratification and personalized screening. Such prediction can benefit from both temporal changes in breast tissue captured across longitudinal examinations and complementary anatomical information provided by craniocaudal (CC) and mediolateral oblique (MLO) views. Effectively leveraging longitudinal information requires identifying corresponding tissue regions across examinations, yet most existing approaches rely on implicit feature alignment without explicitly enforcing spatial consistency. Recent work has demonstrated the benefits of explicit temporal alignment; however, existing explicit alignment methods are not leveraging multi-view information and instead model each view independently, preventing them from jointly exploiting the complementary information provided by CC and MLO views. Further, explicit approaches have been limited to only leveraging one prior examination. To fill these gaps, we propose FlexLMV-Net¹, a breast-wise longitudinal multi-view framework that jointly models CC and MLO views across screening examinations while accommodating a variable number of prior examinations. FlexLMV-Net explicitly aligns longitudinal feature maps and integrates complementary views through cross-view temporal fusion and a dual-stream attention block. Evaluated on EMBED and CSAW-CC, FlexLMV-Net consistently outperforms existing approaches across breast density, cancer subtypes, and demographic subgroups, highlighting the benefits of explicit longitudinal alignment, multi-view modeling, and flexible use of available screening history. We further introduce BCRBench², an open-source benchmark that unifies state-of-the-art single- and multi-timepoint breast cancer risk prediction models under standardized preprocessing, training, inference, and evaluation protocols, enabling fair and reproducible comparison of existing methods.

Keywords: Breast cancer risk prediction, Screening mammography, Longitudinal analysis, Multi-view learning

¹github.com/sot176/FlexLMV-Net

²github.com/sot176/BCRBench

*Corresponding author

Email address: solveig.thrun@uit.no (Solveig Thrun)

1. Introduction

Breast cancer is the most commonly diagnosed cancer among women worldwide and remains a leading cause of cancer-related mortality (Bray et al., 2024). Population-based mammography screening programs have contributed substantially to reductions in breast cancer mortality by enabling earlier cancer detection and timely treatment (Zielonke et al., 2020). In routine clinical practice, each screening examination consists of two standard mammographic views per breast, the craniocaudal (CC) and mediolateral oblique (MLO) views, which provide complementary anatomical information for lesion detection (Bhushan et al., 2021). These examinations are typically performed at regular intervals, with current screening guidelines recommending annual or biennial screening primarily based on age (Ren et al., 2022). While this population-based approach has proven effective, it also results in millions of negative examinations each year, increasing healthcare costs, unnecessary recalls, overdiagnosis, and patient burden (Habib et al., 2021; Bond et al., 2013; McCann et al., 2002). Conversely, these screening protocols may result in under-screening of women at high risk of breast cancer, delaying the detection of early malignancies and leading to diagnosis at more advanced stages (Rubio et al., 2024). Furthermore, age-based screening intervals do not adequately account for the substantial variation in individual breast cancer risk (Clift et al., 2022).

These limitations have motivated a shift toward personalized screening strategies, in which screening frequency and supplemental imaging are tailored to an individual’s estimated risk of developing breast cancer. Such individualized screening strategies have the potential to improve early cancer detection, reduce screening-related harms, and optimize the use of healthcare resources (Clift et al., 2022; van Winkel et al., 2025).

Recent advances in deep learning have enabled future breast cancer risk to be estimated directly from mammograms, often outperforming traditional clinical risk models (Tyrer et al., 2004) based on demographic and lifestyle factors. Early deep learning approaches primarily relied on the current screening examination alone (Yala et al., 2021; Donnelly et al., 2024), learning imaging biomarkers associated with future cancer development from a single time point. However, radiologists routinely compare the current mammogram with prior examinations to identify subtle tissue changes that may indicate early malignant transformation (Akwo et al., 2024; Varela et al., 2005). Motivated by this clinical workflow, recent models have begun incorporating prior mammograms, allowing them to model temporal changes in breast tissue and improve long-term risk prediction (Lee et al., 2023; Wang et al., 2023; Dadsetan et al., 2022; Karaman et al., 2024; Wang et al., 2024; Sun et al., 2025).

While most existing longitudinal mammography models compare current and prior mammograms using implicit feature alignment (e.g., LoMaR Karaman et al. (2024), VMRA-MaR Sun et al. (2025)), recent work has demonstrated that explicitly aligning mammograms before temporal feature comparison produces more accurate representations of longitudinal tissue changes and improves breast cancer risk prediction (Thrun et al., 2025b). However, existing explicit alignment approaches, including ImgFeatAlign (Thrun et al., 2025a) and OA-BreaCR (Wang et al., 2024), perform longitudinal modeling independently for each mammographic view, thereby not exploiting the complementary information contained in the CC and MLO views during temporal analysis. Moreover, existing longitudinal approaches typically assume a fixed number of prior examinations, despite substantial variation in the availability of prior screening mammograms across women. This limits their applicability to patients with different screening histories and, in particular, to women attending screening for the first time.

To address these limitations, we propose FlexLMV-Net, a breast-wise longitudinal multi-

view model for breast cancer risk prediction that jointly exploits temporal and complementary multi-view information while accommodating a variable number of prior examinations. Our model explicitly aligns feature maps from prior and current mammograms for both CC and MLO views, enabling accurate modeling of longitudinal tissue changes, and integrates the aligned representations through cross-view temporal fusion. By combining explicit temporal alignment, joint multi-view analysis, and flexible handling of longitudinal histories, FlexLMV-Net can leverage the available screening history without requiring a fixed number of prior examinations, providing a more comprehensive representation of breast tissue evolution for breast cancer risk prediction.

The main contributions of this study are summarized as follows:

- We propose FlexLMV-Net, a breast-wise longitudinal multi-view framework that captures temporal changes through explicit feature-space alignment and cross-view fusion while accommodating variable numbers of prior examinations.
- We introduce a multi-view dual-stream attention block that models intra- and inter-view dependencies through parallel self- and cross-attention, enabling temporal fusion across variable-length screening histories.
- We validate our approach on two large-scale public datasets (EMBED and CSAW-CC), demonstrating robust performance across breast density, cancer subtypes, and demographic subgroups.
- We contribute BCRBench³, an open-source benchmark that unifies state-of-the-art models for single- and multi-timepoint breast cancer risk prediction within a common, reproducible framework, supported by comprehensive documentation⁴, to enable fair and reproducible model comparison.

A preliminary version of this work appeared in (Thrun et al., 2026). In this extended study, we substantially enhance our previous model (LMV-Net) through the following key improvements:

1. We extend LMV-Net to FlexLMV-Net, enabling variable numbers of prior examinations through a feature-aware temporal fusion block that integrates available longitudinal information. This allows FlexLMV-Net to also be applied to first-time screening participants, broadening its applicability to the wider screening population.
2. We introduce BCRBench, a structured benchmark for image-based breast cancer risk prediction that provides standardized evaluation protocols and facilitates reproducible comparison of risk prediction models.
3. We deepen the literature review to provide a more comprehensive positioning of our approach relative to existing methods and to clarify how our work extends prior research.
4. We substantially expand the methodological description of FlexLMV-Net, detailing its handling of variable-length longitudinal inputs, explicit feature alignment, and multi-view temporal fusion, while systematically analyzing the contribution of each component to predictive performance.

³github.com/sot176/BCRBench

⁴<https://sot176.github.io/BCRBench>

5. We further assess the generalization of LMV-Net and FlexLMV-Net across datasets and evaluate its robustness across clinically relevant subgroups, demonstrating its applicability across different populations and imaging settings.

The paper is organized as follows. Section 2 provides an overview of relevant related work, focusing on breast cancer risk prediction and multi-view analysis. Section 3 presents the proposed FlexLMV-Net in detail. In Section 4, we describe the experimental setup and the evaluation metrics used to assess model performance. Section 5 presents and discusses the results. Finally, Sections 6 and 7 address the limitations of the proposed approach, outline directions for future research, and summarize the key findings of the study.

2. Related Work

2.1. Breast Cancer Risk Prediction

Traditional breast cancer risk prediction has relied on classical models such as Tyrer-Cuzick (Tyrer et al., 2004), which primarily use manually collected clinical variables and do not leverage imaging information from routine screening. However, more recently end-to-end deep learning approaches have enabled risk prediction directly from full-field mammograms (Yala et al., 2021; Karaman et al., 2024; Lee et al., 2023; Wang et al., 2024; Sun et al., 2025; Yeoh et al., 2026), allowing models to learn predictive imaging biomarkers automatically (Hussain et al., 2024). Mirai (Yala et al., 2021) established a strong baseline using multi-view mammograms and significantly outperformed traditional risk models for long-term risk prediction. AsymMirai (Donnelly et al., 2024) further demonstrated that bilateral asymmetry can serve as a predictive and interpretable imaging biomarker. However, these approaches operate on a single screening timepoint and do not capture temporal changes in breast tissue that may provide additional information about future cancer risk.

Motivated by the longitudinal assessment routinely performed by radiologists, several recent methods incorporate prior screening examinations into risk prediction. PRIME+ (Lee et al., 2023), VMRA-MaR (Sun et al., 2025), LoMaR (Karaman et al., 2024), and TRINet (Yeoh et al., 2026) combine information from current and prior mammograms, demonstrating improved risk stratification compared with single-timepoint approaches. TRINet additionally incorporates time-decay attention and handcrafted radiomic features within a multiple-instance learning framework. However, these methods primarily rely on implicit alignment, comparing or combining representations across timepoints without explicitly enforcing spatial correspondence, which may limit their ability to capture subtle longitudinal tissue changes.

While implicit alignment is computationally efficient, it is sensitive to geometric variations caused by differences in patient positioning and breast compression across visits. Recent evidence indicates that explicit spatial registration can better capture longitudinal tissue changes and improve risk prediction (Thrun et al., 2025b). OA-BreaCR (Wang et al., 2024) learns deformation fields directly in feature space, while ImgFeatAlign (Thrun et al., 2025a,b) estimates image-based deformation fields and applies them to feature representations, enabling more precise spatial correspondence across examinations.

Despite these advances, existing explicit alignment approaches model mammographic views independently and therefore do not jointly exploit the complementary information provided by CC and MLO projections during longitudinal analysis. Furthermore, existing explicit longitudinal models typically assume a fixed number of prior examinations, limiting their applicability to patients with different screening histories and to examinations where no prior mammogram is

available. FlexLMV-Net addresses these limitations by explicitly aligning longitudinal representations, jointly modeling CC and MLO views, and accommodating a variable number of prior examinations.

2.2. Multi-view mammography analysis

Early deep learning approaches in multi-view mammography typically processed CC and MLO projections through independent convolutional branches, combining them only after global pooling or simple feature concatenation (Wu et al., 2020). While effective, such strategies may discard richer cross-view correlations that radiologists exploit clinically. More recent architectures explicitly model these interactions to capture cross-view relationships (van Tulder et al., 2021). For instance, MV-Swin-T (Sarker et al., 2024) introduces a Multi-headed Dynamic Attention Block, enabling self- and cross-view information fusion at the spatial feature level and outperforming multi-view CNN baselines across multiple benchmark datasets. Similarly, MVPT-NET (Chen and Martel, 2025) leverages visual prompt tuning on a pre-trained single-view backbone to integrate multi-view information efficiently without requiring full retraining.

In longitudinal risk prediction, multi-view information has also been combined with temporal modeling. LRP-NET (Dadsetan et al., 2022) processes CC and MLO projections from multiple consecutive examinations through parallel GRU branches, capturing bilateral spatiotemporal tissue differences for near-term risk estimation. LoMaR (Karaman et al., 2024) extends the single-timepoint Mirai model with a transformer-based visit aggregator and temporal positional encodings, distilling information from multiple visits and all four standard views into a comprehensive risk embedding. VMRA-MaR (Sun et al., 2025) further incorporates bilateral asymmetry through side encoding and temporal attention, capturing fine-grained spatial and longitudinal asymmetric cues across CC and MLO views.

Despite these advances, existing multi-view longitudinal models either do not explicitly register mammographic views across timepoints or assume a fixed number of prior examinations. This limits their ability to jointly exploit spatially aligned multi-view information while accommodating patients with different screening histories. FlexLMV-Net addresses these gaps by combining explicit longitudinal feature alignment, cross-view temporal fusion, and flexible input handling for varying numbers of prior examinations within a unified framework.

3. Methodology

Breast cancer risk prediction estimates the probability of developing breast cancer within a fixed time horizon. Here, we consider five-year risk prediction from screening mammography. Longitudinal data provide valuable information by enabling the detection of subtle tissue changes over time. To jointly exploit longitudinal and multi-view (CC and MLO) information while accommodating variable numbers of prior examinations, we introduce FlexLMV-Net (Figure 1). The framework comprises three components. First, longitudinal feature maps from prior examinations are explicitly aligned with the current examination to establish spatial correspondence and capture temporal changes, then aggregated by the Feature-Aware Temporal Fusion block. Second, a multi-view dual-stream attention block models intra- and inter-view dependencies through self- and cross-attention, integrating complementary information across views and longitudinal time points. Finally, a risk prediction head estimates five-year risk from a fused multi-view representation, with view-specific features used as auxiliary supervision during training. Each component is described in detail in the following subsections.

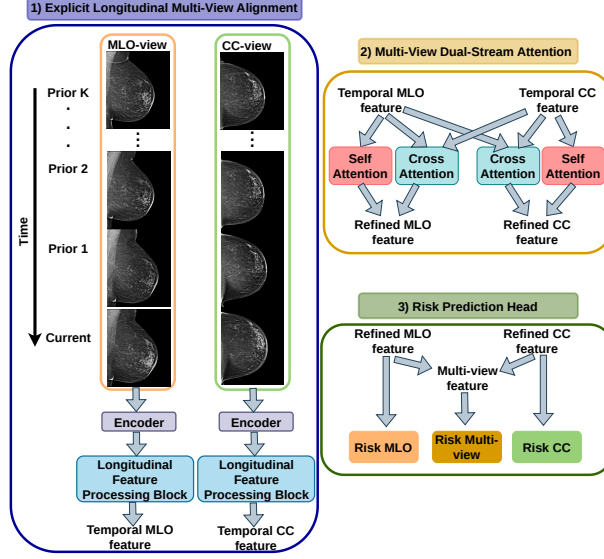


Figure 1: Overview of the proposed FlexLMV-Net architecture. The framework consists of three main components: (1) explicit longitudinal multi-view alignment to establish correspondence between current and prior examinations, (2) a multi-view dual-stream attention block that models intra-view and inter-view dependencies across longitudinal data, and (3) a risk prediction head that leverages fused multi-view representations, with auxiliary view-specific supervision during training, to predict five-year breast cancer risk.

3.1. Explicit Longitudinal Multi-View Alignment

The explicit longitudinal alignment component of FlexLMV-Net, illustrated in Figure 2, consists of a shared image encoder that processes longitudinal multi-view mammographic inputs, together with a longitudinal feature processing block designed to accommodate a variable number of prior examinations.

First, a shared encoder extracts feature maps from the current and available prior mammograms in both MLO and CC views. For a given patient, the current examination is represented by feature maps $\mathbf{f}_{\text{cur}}^{\text{MLO}}, \mathbf{f}_{\text{cur}}^{\text{CC}} \in \mathbb{R}^{f_c \times f_h \times f_w}$, while the (K) available prior examinations are represented by feature maps $\mathbf{f}_{\text{pri-}k}^{\text{MLO}}, \mathbf{f}_{\text{pri-}k}^{\text{CC}} \in \mathbb{R}^{f_c \times f_h \times f_w}$, for $k \in (1, \dots, K)$, where (K) may vary between patients.

Our approach leverages explicit longitudinal feature alignment to establish spatial correspondence between the current examination and each available prior examination, enabling anatomically consistent modeling of temporal changes. Specifically, we use image-derived deformation fields obtained from a pretrained and frozen MammoRegNet (Thrun et al., 2025a) to align each prior examination to the current examination. For each prior examination and view, MammoRegNet produces a deformation field $\phi \in \mathbb{R}^{2 \times H \times W}$, which is downsampled to $\phi \in \mathbb{R}^{2 \times f_h \times f_w}$ to match the feature-map resolution. The corresponding deformation field is then applied to the prior feature map using a Spatial Transformer Block (Meng et al., 2023; Thrun et al., 2025a) with bilinear sampling, yielding aligned prior features $\mathbf{f}_{\text{pri-}k\text{-align}}^{\text{MLO}}, \mathbf{f}_{\text{pri-}k\text{-align}}^{\text{CC}}$ for $k \in (1, \dots, K)$. This process is performed independently for each available prior examination, resulting in a sequence of spatially aligned longitudinal representations for each view.

To capture temporal changes, view-specific difference features are computed between consecutive examinations using the aligned prior features. For the most recent prior examination, the

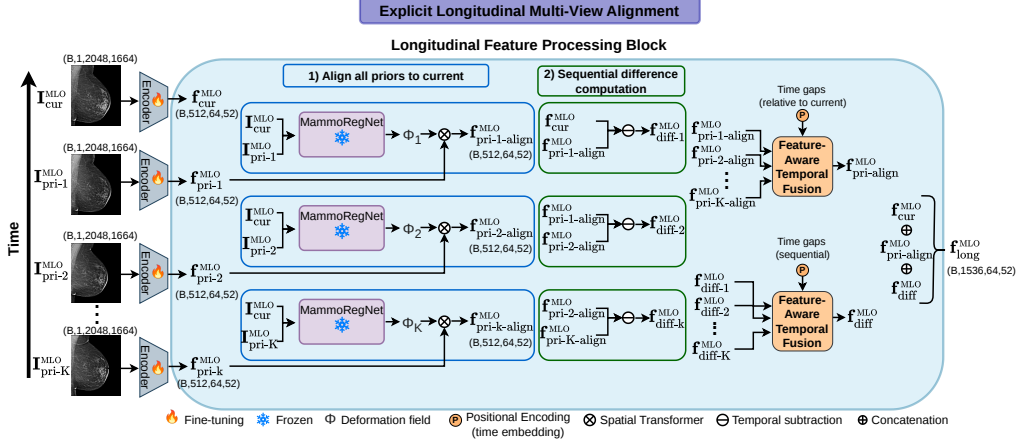


Figure 2: Overview of the explicit longitudinal multi-view feature alignment in FlexLMV-Net. For illustration, the alignment process is shown for the MLO view, where prior feature maps are aligned to the current features using image-derived deformation fields, enabling consistent correspondence and the construction of view-specific longitudinal representations. The CC view is processed in exactly the same manner.

difference is computed with respect to the current examination, while differences between earlier examinations are computed sequentially. Specifically, for each view, the resulting difference features capture changes between temporally adjacent examinations after spatial alignment.

The resulting sequences of aligned prior and difference features can contain a variable number of elements depending on the available longitudinal imaging history. Rather than requiring a fixed number of prior examinations, FlexLMV-Net uses a Feature-Aware Temporal Fusion block to independently aggregate the aligned prior and difference features into compact longitudinal representations. The temporal gaps between screenings are incorporated into this fusion process to account for varying intervals between examinations. The resulting fused representations are then combined with the current features to form the longitudinal CC and MLO representations, which are subsequently processed by the multi-view dual-stream attention block.

The Feature-Aware Temporal Fusion block and its handling of variable-length longitudinal sequences are described in detail in the following subsection.

3.2. Feature-Aware Temporal Fusion

To flexibly aggregate longitudinal information from a variable number of prior examinations, we introduce a Feature-Aware Temporal Fusion block (Figure 3). The block independently fuses the aligned prior features and difference features, allowing the model to adaptively determine the contribution of each available examination based on its feature content, temporal distance, and spatial relevance. Let K denote the number of available prior examinations for a patient, with $k = 1, \dots, K$ indexing the individual prior examinations.

For each longitudinal feature $\mathbf{f}_k$, the pooled current feature, pooled longitudinal feature, and absolute temporal gap Δt_k are concatenated and passed through a lightweight multilayer perceptron (MLP) to obtain a temporal importance score:

$$s_k = \text{MLP}([\text{GAP}(\mathbf{f}_{\text{cur}}); \text{GAP}(\mathbf{f}_k); \Delta t_k]), \quad k = 1, \dots, K.$$

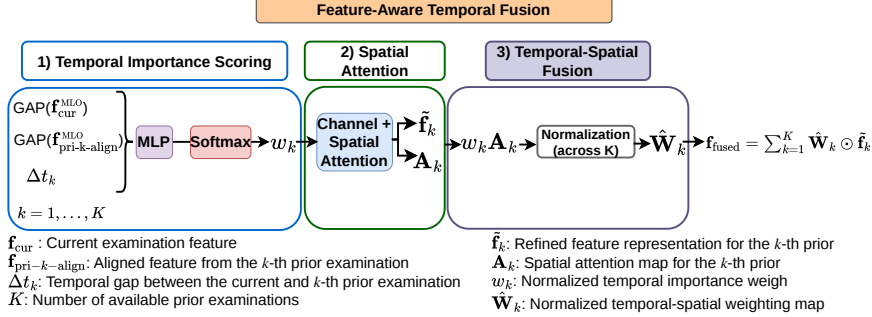


Figure 3: The Feature-aware temporal fusion of FlexLMV-Net adaptively weights longitudinal features based on temporal importance and spatial relevance.

The MLP consists of two linear layers with a ReLU activation and dropout between them, producing a scalar importance score for each prior examination. The resulting scores are normalized across the K available examinations using a softmax:

$$w_k = \frac{\exp(s_k)}{\sum_{j=1}^K \exp(s_j)}, \quad k = 1, \dots, K,$$

where s_k denotes the scalar importance score assigned to the k -th prior examination, and w_k denotes its normalized temporal importance weight. The input to the MLP has dimension $2f_C + 1$, corresponding to the pooled current and longitudinal features together with the temporal gap. Dropout provides regularization of the learned temporal weighting.

In parallel, a spatial attention module applies channel and spatial attention to each longitudinal feature, producing a refined feature representation and a corresponding spatial attention map. The temporal weight is then combined with the spatial attention map to obtain a joint temporal-spatial weighting map:

$$\mathbf{W}_k = w_k \mathbf{A}_k, \quad k = 1, \dots, K.$$

where $\mathbf{A}_k$ denotes the spatial attention map for the k -th prior examination and $\mathbf{W}_k$ denotes the resulting unnormalized temporal-spatial weighting map. These weighting maps are subsequently normalized across the K available examinations independently at each spatial location, yielding $\hat{\mathbf{W}}_k$. The refined longitudinal features are then fused using the resulting spatially varying weights:

$$\mathbf{f}_{\text{fused}} = \sum_{k=1}^K \hat{\mathbf{W}}_k \odot \tilde{\mathbf{f}}_k,$$

where $\tilde{\mathbf{f}}_k$ denotes the refined longitudinal feature representation of the k -th prior examination, $\hat{\mathbf{W}}_k$ denotes its normalized temporal-spatial weighting map, and $\odot$ denotes element-wise multiplication.

This feature-aware fusion enables FlexLMV-Net to selectively emphasize both informative examinations and relevant spatial regions while naturally accommodating a variable number of available priors. The fusion process is applied independently to the aligned prior features and the difference features, which are subsequently combined with the current features to form the longitudinal CC and MLO representations.

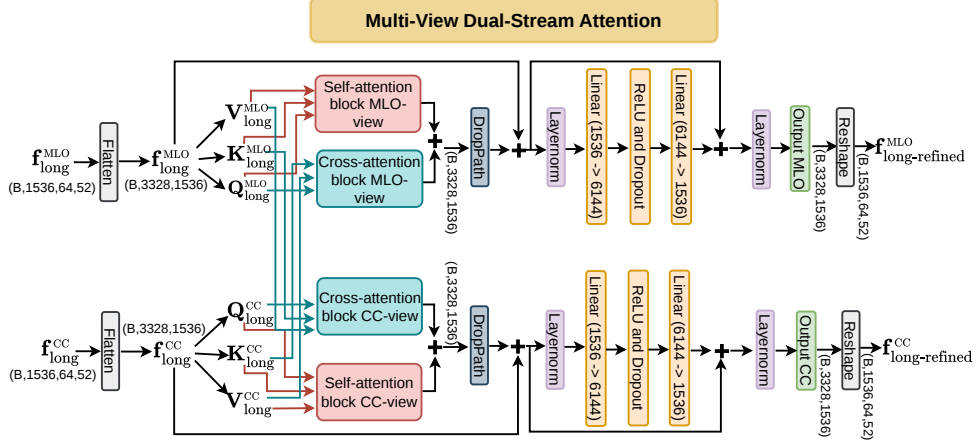


Figure 4: Overview of the Multi-View Dual-Stream Attention block in FlexLMV-Net. Longitudinal features from the MLO and CC views are tokenized and processed through parallel self- and cross-attention streams to model intra- and inter-view dependencies. The resulting features are then refined via residual connections and reshaped back into spatial feature maps, yielding enhanced multi-view representations.

3.3. Multi-View Dual-Stream Attention

We propose a Multi-view Dual-Stream Attention block (see Figure 4) to enhance the modeling of longitudinal multi-view representations by jointly refining the feature maps $\mathbf{f}_{\text{long}}^{\text{MLO}} \in \mathbb{R}^{3fc \times fh \times fw}$ and $\mathbf{f}_{\text{long}}^{\text{CC}} \in \mathbb{R}^{3fc \times fh \times fw}$. The block operates on flattened spatial features and processes each view through two parallel attention streams: a self-attention stream and a cross-attention stream.

For each view, the input feature maps $\mathbf{f}_{\text{long}} \in \mathbb{R}^{3fc \times fh \times fw}$ are first reshaped into a sequence of tokens $\in \mathbb{R}^{(fh \cdot fw) \times 3fc}$. Queries, keys, and values are then obtained via linear projections. In the self-attention stream, queries, keys, and values originate from the same view (e.g., MLO), enabling the model to capture long-range intra-view dependencies and improve spatial coherence. In parallel, the cross-attention stream uses queries from one view and keys and values from the complementary view (e.g., CC), allowing the model to attend to anatomically and semantically corresponding regions across views.

The outputs of the self-attention and cross-attention streams are fused via element-wise summation, producing a unified attention representation for each view. DropPath (Huang et al., 2016) regularization is applied to the fused features to improve generalization. The result is then combined with the original input through a residual connection, followed by layer normalization (Ba et al., 2016). Each view is subsequently processed by a feed-forward network (FFN), consisting of two linear layers with ReLU activation and dropout (Srivastava et al., 2014). The FFN expands the feature dimension by a factor of four before projecting it back to the original size, increasing representational capacity. A second residual connection and layer normalization are applied after the FFN. Finally, the refined token sequences are reshaped back to their original spatial dimensions, yielding updated feature maps $\mathbf{f}_{\text{long-refined}}^{\text{MLO}}, \mathbf{f}_{\text{long-refined}}^{\text{CC}} \in \mathbb{R}^{3fc \times fh \times fw}$ for each view. Through this dual-stream attention design, the model effectively integrates intra-view spatial context and inter-view complementary information, resulting in enriched and discriminative multi-view representations for downstream risk prediction.

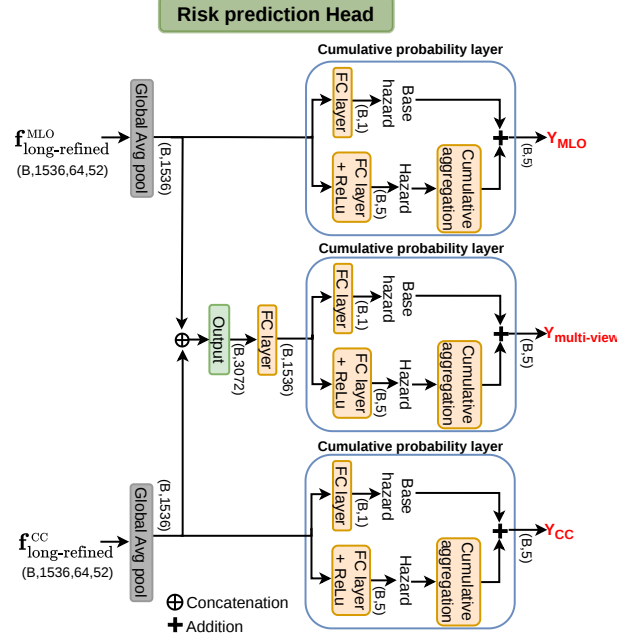


Figure 5: Risk prediction head of FlexLMV-Net. Refined MLO and CC features are globally pooled and fused for risk estimation via a cumulative probability layer, with auxiliary view-specific branches providing additional supervision during training.

3.4. Risk Prediction Head

Figure 5 illustrates the risk prediction head, which takes the refined longitudinal feature maps $\mathbf{f}_{\text{long-refined}}^{\text{MLO}}$ and $\mathbf{f}_{\text{long-refined}}^{\text{CC}}$ as input to estimate time-dependent breast cancer risk. The refined feature maps $\mathbf{f}_{\text{long-refined}}^{\text{MLO}}$ and $\mathbf{f}_{\text{long-refined}}^{\text{CC}}$ are first globally pooled to obtain compact feature vectors for each view. These are then concatenated to form a fused multi-view representation. In addition to the fused representation, view-specific branches are retained to provide auxiliary supervision during training. Risk estimation is performed using a cumulative probability layer (Yala et al., 2021; Lee et al., 2023), a formulation commonly used in survival analysis for time-to-event modeling. This layer consists of two fully connected branches: one predicting hazard values $h_j(\mathbf{f})$ at each future time step $j \in 1, \dots, T_{\max}$, and another estimating a baseline hazard $h_{\text{base}}(\mathbf{f})$. A ReLU activation function constrains the predicted hazard values to be non-negative. The cumulative risk score up to time t is then computed as

$$z(t | \mathbf{f}) = h_{\text{base}}(\mathbf{f}) + \sum_{j=1}^t h_j(\mathbf{f}), \quad (1)$$

which models risk as a sequential accumulation of hazards over time. During training, the cumulative risk scores are passed to the binary cross-entropy loss with logits, which applies a sigmoid transformation to map the scores to probabilities in the range $[0, 1]$. To encourage both global and view-specific learning, we adopt a multilevel joint learning strategy (Wang et al., 2024). The fused multi-view representation is used for the primary risk prediction, while the individual

MLO and CC feature vectors are passed through separate cumulative probability layers to produce auxiliary predictions. This setup promotes discriminative feature learning at the view level while improving the quality of the fused representation used at inference. The model is trained using a time-dependent binary cross-entropy (BCE) loss with logits. The ground-truth $\mathbf{y}$ is a binary vector of length $T_{\max}$ and encodes whether cancer occurs within each time horizon t , with $T_{\max} = 5$ years in our setting. Specifically, $y(t) = 1$ if cancer is diagnosed within t years, and 0 otherwise. The sigmoid transformation inherent to the BCE-with-logits formulation converts the cumulative risk scores into probabilities for optimization. To handle censored data, we employ a masking function $\delta(t)$ (Yala et al., 2021; Wang et al., 2024), which ensures that only valid time points contribute to the loss:

$$\delta(t) = \begin{cases} 1, & \text{if the follow-up period} \geq \min(t, T_{\max}) \\ & \text{or if the patient is diagnosed within } t \text{ years.} \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

The overall loss is computed as

$$\mathcal{L}_{\text{BCE}} = \sum_{t=1}^{T_{\max}} \delta(t) \cdot \ell_{\text{BCE}}(t), \quad (3)$$

where the per-time-step binary cross-entropy is defined as

$$\ell_{\text{BCE}}(t) = -[y(t) \cdot \log(P(t)) + (1 - y(t)) \cdot \log(1 - P(t))] , \quad (4)$$

This formulation ensures that loss contributions are restricted to observed and clinically valid time points, enabling robust learning in the presence of censored longitudinal data.

4. Experimental Setup and BCRBench

This section outlines the experimental setup for training and evaluating our models. We describe the datasets, label generation and preprocessing (Section 4.1), evaluation metrics (Section 4.2), and model implementation and training details (Section 4.3). Our proposed BCRBench is described in Section 4.4.

4.1. Datasets, Label Generation, and Preprocessing

4.1.1. Datasets

We evaluate our method on two large-scale public available longitudinal mammography datasets: the Emory Breast Imaging Dataset (EMBED)⁵ (Jeong et al., 2023) and the Cohort of Screen-Aged Women Case-Control (CSAW-CC) dataset⁶ (Dembrower et al., 2020). Both datasets consist of two-dimensional (2D) full-field digital mammography (FFDM) images collected over multiple years, enabling comprehensive longitudinal analysis. Synthetic 2D mammograms (C-view images) from digital breast tomosynthesis (DBT) were excluded.

⁵<https://aws.amazon.com/marketplace/pp/prodview-unw4li5rkivs2#overview>

⁶<https://snd.se/en/catalogue/dataset/2021-204-1>

The publicly available EMBED dataset contains 364,564 mammograms from 23,256 patients acquired between 2013 and 2020 using the vendors Hologic, GE and Fujifilm. CSAW-CC includes mammograms from 873 breast cancer cases and 7,850 controls screened between 2008 and 2015, all acquired using Hologic equipment. Both datasets include invasive and non-invasive breast cancer cases. Negative cases are defined as women without a breast cancer diagnosis during the respective follow-up periods (8 years for EMBED and 7 years for CSAW-CC).

To ensure sufficient longitudinal information, only patients with at least five years of follow-up were included, following prior work (Wang et al., 2024). For patients with more than two screening exams, all consecutive screening pairs were used for training.

4.1.2. Time-to-Events label generation

The Time-to-Events label represents the temporal interval between a screening mammogram and the corresponding breast cancer diagnosis. Labels were generated following established protocols (Wang et al., 2024; Karaman et al., 2024; Thrun et al., 2025b), with dataset-specific procedures for identifying positive and negative cases and assigning cancer timing.

EMBED: Labeling follows Wang et al. (2024); Thrun et al. (2025b). Positive cases include mammograms with BI-RADS 6 (biopsy-proven malignancy) or pathological severity scores indicating invasive or non-invasive cancer. Negative cases include BI-RADS 1–2 exams, as well as BI-RADS 0 exams that were subsequently resolved as BI-RADS 1 or 2 in follow-up studies. The cancer diagnosis date is defined as the biopsy confirmation date.

CSAW-CC: Labeling follows Karaman et al. (2024); Thrun et al. (2025b), using the radiological timing variable (`rad_timing`). Positive cases correspond to screen-detected cancers (`rad_timing` = 1) and interval cancers (`rad_timing` = 2), while negative cases consist of exams without any recorded `rad_timing`, indicating no cancer diagnosis during the study period. Since biopsy dates are not available in CSAW-CC, cancer timing is inferred from `rad_timing`. For each patient, the cancer year is determined from the most recent exam: it is set to the exam year for screen-detected cancers and to one year after the exam for interval cancers. For cases in 2016 (the final study year), the cancer year is set to 2016 to accommodate follow-up constraints.

Time-to-Events: For both datasets, the Time-to-Events label is calculated as the difference between the assigned cancer year and the mammogram exam year. Only negative prior screening mammograms are included, as routine screening usually stops once a patient is diagnosed with breast cancer.

4.1.3. Preprocessing

All mammograms were preprocessed following established pipelines (Yala et al., 2021). Raw DICOM images were converted to 16-bit grayscale PNG format, after which background artifacts were removed and the breast region was segmented to exclude non-breast areas. The resulting images were resized to a standardized resolution of 1664×2048 pixels and intensity-normalized to ensure consistency across scanners and acquisition settings. Density information is derived from BI-RADS (Sickles, 2013) categories for EMBED, while for CSAW-CC, LIBRA-based (Keller et al., 2012) percent density estimates are grouped into three categories of equal size corresponding to low, medium and high density. Detailed statistics on the Time-to-Cancer distribution and dataset splits are provided in the Appendix in Table A.4, while breast density distributions are summarized in Table A.5.

4.2. Evaluation Metrics

Following previous studies (Lee et al., 2023; Wang et al., 2024; Yala et al., 2021), we evaluated the performance of the methods on the 1–5 year risk prediction task using the concordance index (C-index) (Uno et al., 2011) and the area under the receiver operating characteristic curve (AUC) (Hanley and McNeil, 1982) across the 5 year risk horizons. 95% confidence intervals (CIs) were estimated via bootstrapping with 1,000 resamples. Statistical significance of improvements over the second-best model was assessed using a paired nonparametric bootstrap test (Kang et al., 2015; Wang et al., 2023), with $p < 0.05$ considered significant.

4.3. Implementation Details

All models were implemented in PyTorch 2.0.1 (Paszke et al., 2019) and trained for up to 30 epochs on eight AMD Instinct MI210 GPUs (64 GB each), using a batch size of 4 per GPU (32 total). We compare FlexLMV-Net with Mirai (Yala et al., 2021), VMRA-MaR (Sun et al., 2025), OA-BreaCR (Wang et al., 2024), ImgFeatAlign (Thrun et al., 2025a,b), and LMV-Net (Thrun et al., 2026). For fair comparison, we use the Mirai encoder (Yala et al., 2021), a ResNet-18 (He et al., 2016) without the final fully connected and global average pooling layers, following prior work (Karaman et al., 2024; Sun et al., 2025; Thrun et al., 2025b). Self- and cross-attention use four heads. Models are optimized with AdamW (Loshchilov and Hutter, 2019) using a learning rate of 5×10^{-5} (5×10^{-6} for encoder fine-tuning) and weight decay of 1×10^{-4} . The learning rate is halved after five epochs without validation C-index improvement, with early stopping after 15 epochs. Augmentation includes random cropping, affine transformations, color jitter, and gamma correction. Comparison methods use their official configurations. LMV-Net is evaluated with both frozen and fine-tuned Mirai encoders, whereas FlexLMV-Net uses the fine-tuned encoder, which yielded improved performance for LMV-Net. In this study, FlexLMV-Net considers up to three prior examinations due to memory limitations.

4.4. BCRBench

As part of this study, we developed BCRBench (<https://github.com/sot176/BCRBench>), an open-source benchmarking framework for image-based breast cancer risk prediction. BCRBench addresses the lack of standardized experimental setups by unifying pre-processing, data handling, training, and evaluation within a reproducible framework, enabling consistent comparisons across models and datasets.

Implemented in PyTorch, BCRBench follows a modular design through which datasets, models, and evaluation protocols can be integrated via a common interface. The current version includes EMBED and CSAW-CC and five risk prediction models: Mirai (Yala et al., 2021), VMRA-MaR (Sun et al., 2025), OA-BreaCR (Wang et al., 2024), ImgFeatAlign (Thrun et al., 2025a,b), LMV-Net (Thrun et al., 2026), and FlexLMV-Net. These models use standard CC and MLO views and cover single-timepoint and longitudinal risk prediction. For Mirai-based models, including VMRA-MaR, ImgFeatAlign, LMV-Net, and FlexLMV-Net, both frozen and end-to-end fine-tuned encoder settings are supported.

BCRBench applies consistent preprocessing, training, and evaluation procedures across methods. Performance is assessed using AUC and C-index across five-year prediction horizons, with support for subgroup analyses based on clinically relevant characteristics such as breast density and cancer subtype. The framework provides the experimental foundation for the comparisons presented in this study and is openly available with comprehensive documentation (<https://sot176.github.io/BCRBench>) to facilitate reproducibility and future benchmarking.

Table 1: 1–5 Year breast cancer risk prediction: C-index and AUC values with $\pm 95\%$ CI. ✱ : frozen encoder, 🔥 : fine-tuned encoder. $\Delta(\%)$ denotes difference between FlexLMV-Net and best performing baseline. * indicates statistically significant improvement of LMV-Net and FlexLMV-Net over the best baseline ($p < 0.05$). Training time in hours (h) and total number of model parameters (#, in millions).

Dataset	Method	C-index (%) $\uparrow$	Follow-up year AUC (%) $\uparrow$					h	#
			1-yr	2-yr	3-yr	4-yr	5-yr		
EMBED	Mirai(✱)	59.8 $\pm$ 8.7	56.7 $\pm$ 20.7	52.6 $\pm$ 15.3	54.0 $\pm$ 12.6	49.3 $\pm$ 12.5	49.2 $\pm$ 11.6	6	28
	Mirai(🔥)	69.1 $\pm$ 10.6	71.0 $\pm$ 15.6	70.9 $\pm$ 15.5	70.6 $\pm$ 11.7	65.7 $\pm$ 11.3	65.9 $\pm$ 11.6	7	28
	VMRA-MaR (🔥)	72.4 $\pm$ 12.5	77.9 $\pm$ 4.5	75.3 $\pm$ 9.7	73.1 $\pm$ 12.7	68.9 $\pm$ 14.3	67.1 $\pm$ 11.8	45	2802
	OA-BreaCR (🔥)	72.3 $\pm$ 2.0	72.9 $\pm$ 2.8	72.3 $\pm$ 4.9	72.4 $\pm$ 2.5	72.3 $\pm$ 2.4	73.2 $\pm$ 2.4	9	15
	ImgFeatAlign (✱)	74.7 $\pm$ 2.4	75.0 $\pm$ 2.9	75.5 $\pm$ 2.4	75.3 $\pm$ 2.1	75.9 $\pm$ 2.3	72.5 $\pm$ 3.2	12	52
	ImgFeatAlign (🔥)	72.2 $\pm$ 2.7	72.1 $\pm$ 3.0	73.3 $\pm$ 2.6	72.2 $\pm$ 2.5	71.3 $\pm$ 2.4	72.6 $\pm$ 3.3	14	52
	LMV-Net (✱)	78.3 $\pm$ 3.0	80.8 $\pm$ 3.7	78.3 $\pm$ 3.5	77.5 $\pm$ 3.3	76.6 $\pm$ 3.2	75.1 $\pm$ 4.9	15	65
	LMV-Net (🔥)	81.4$\pm$2.8	83.9$\pm$3.4	82.1$\pm$3.0	80.8$\pm$2.9	80.0$\pm$2.8	77.2 $\pm$ 4.7	16	65
	FlexLMV-Net (🔥)	81.2$\pm$2.9	84.1$\pm$3.2	82.0$\pm$3.0	79.6$\pm$3.1	79.3$\pm$3.1	79.0$\pm$4.4	22	65.14
	$\Delta(\%) \uparrow$	+ 6.5	+ 6.2	+ 6.5	+ 4.3	+ 3.4	+ 5.8		
CSAW-CC	Mirai(✱)	71.4 $\pm$ 2.7	56.8 $\pm$ 10.6	54.7 $\pm$ 5.8	56.4 $\pm$ 5.1	56.1 $\pm$ 4.2	55.4 $\pm$ 4.0	7	28
	Mirai (🔥)	72.3 $\pm$ 2.9	69.9 $\pm$ 3.3	69.8 $\pm$ 3.9	69.2 $\pm$ 3.8	67.0 $\pm$ 3.4	66.9 $\pm$ 3.4	8	28
	VMRA-MaR (🔥)	72.7 $\pm$ 2.6	70.9 $\pm$ 5.5	70.3 $\pm$ 5.6	69.5 $\pm$ 4.4	70.9 $\pm$ 4.3	64.5 $\pm$ 6.0	46	2802
	OA-BreaCR (🔥)	61.6 $\pm$ 4.2	63.5 $\pm$ 4.4	61.7 $\pm$ 2.1	65.0 $\pm$ 2.0	67.0 $\pm$ 1.9	67.7 $\pm$ 2.2	10	15
	ImgFeatAlign (✱)	70.4 $\pm$ 1.9	72.0 $\pm$ 2.4	71.5 $\pm$ 1.9	72.6 $\pm$ 1.9	72.0 $\pm$ 2.1	75.2 $\pm$ 2.3	13	52
	ImgFeatAlign (🔥)	71.9 $\pm$ 2.0	74.1 $\pm$ 2.6	72.3 $\pm$ 2.0	73.1 $\pm$ 1.9	72.1 $\pm$ 2.0	74.5 $\pm$ 2.3	16	52
	LMV-Net (✱)	73.4 $\pm$ 2.5	77.5 $\pm$ 3.3	73.5 $\pm$ 2.6	74.0 $\pm$ 2.5	73.1 $\pm$ 2.6	75.9$\pm$2.8	16	65
	LMV-Net (🔥)	74.4 $\pm$ 2.7	78.7 $\pm$ 3.6	74.1$\pm$2.9	74.3$\pm$2.7	73.5$\pm$2.8	75.0 $\pm$ 3.1	17	65
	FlexLMV-Net (🔥)	77.1$\pm$2.5	82.5$\pm$3.0	74.6$\pm$3.0	73.7$\pm$2.9	71.9 $\pm$ 3.1	72.3 $\pm$ 3.3	26	65.14
	$\Delta(\%) \uparrow$	+ 4.4	+ 10.5	+ 3.1	+ 1.1	-0.1	-2.9		

5. Results

This section presents the quantitative and qualitative results of FlexLMV-Net and its special case LMV-Net, where $K = 1$, for five-year breast cancer risk prediction. Section 5.1 compares both models with SOTA methods, while Sections 5.2–5.4 evaluate performance across breast density, cancer type, and demographic subgroups. Section 5.5 assesses in-domain and cross-dataset generalization, and Section 5.6 analyzes the interpretability of the dual-stream attention block. Section 5.7 examines the effect of varying longitudinal history length, while Section 5.8 presents the ablation study.

5.1. Comparison with State-of-the-Arts (SOTA)

We compare the proposed FlexLMV-Net and LMV-Net with state-of-the-art (SOTA) methods, including Mirai (Yala et al., 2021) (Sci. Transl. Med.’21), OA-BreaCR (Wang et al., 2024) (MICCAI’24), VMRA-MaR (Sun et al., 2025) (MICCAI’25), and ImgFeatAlign (Thrun et al., 2025b) (MICCAI’25), on both datasets. Table 1 summarizes the results for 1–5 year breast cancer risk prediction. Both LMV-Net and FlexLMV-Net consistently outperform the competing methods across the two datasets. On EMBED, which encompasses diverse scanner types and population demographics, LMV-Net achieves substantial improvements over SOTA methods, demonstrating robustness to heterogeneous data. Fine-tuning the Mirai encoder further improves LMV-Net performance, highlighting the benefit of adapting the pretrained representation to the risk prediction task. FlexLMV-Net further improves upon LMV-Net at several prediction horizons while maintaining strong performance across the full 1–5 year range. Similarly, on the CSAW-CC dataset, which is less diverse and primarily derived from a single scanner type and a

predominantly Scandinavian population, LMV-Net achieves consistent performance gains. Overall, LMV-Net establishes strong predictive performance, while FlexLMV-Net extends the framework to accommodate a variable number of prior examinations and achieves the strongest overall C-index, with particularly strong short-term risk prediction. Moreover, both LMV-Net (🔥) and FlexLMV-Net (🔥) exhibit relatively narrow confidence intervals, indicating reliable and stable performance across prediction horizons. The improvements over the best baseline are statistically significant ($p < 0.05$).

5.2. Performance across density categories

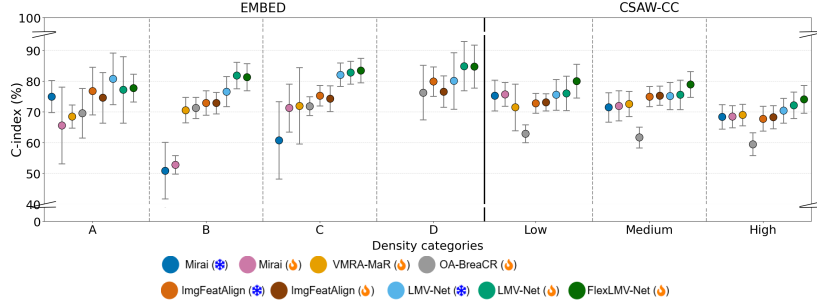


Figure 6: C-index across breast density categories for the proposed LMV-Net and state-of-the-art methods on the EMBED and CSAW-CC datasets (in-domain evaluation). Points show the mean C-index and error bars represent the 95% confidence interval.

The C-index results (Figure 6) show strong and consistent performance of LMV-Net (🔥) and FlexLMV-Net (🔥) across breast density categories on both EMBED and CSAW-CC. On EMBED, LMV-Net performs particularly well across density groups, while FlexLMV-Net achieves comparable or higher performance in the highest-density categories. On CSAW-CC, performance differences are smaller, with both longitudinal models remaining competitive. The year-specific AUC results (Figure 7 and Figure 8) show a similar trend, with both models maintaining strong performance across density categories and follow-up years. FlexLMV-Net is particularly strong for short-term prediction. Overall, these results demonstrate robust risk stratification across density subgroups.

5.3. Performance across cancer types

Figure 9 presents the C-index performance of LMV-Net, FlexLMV-Net, and SOTA models across cancer types on EMBED and CSAW-CC. Results are reported for non-invasive and invasive cancers, with invasive cancers in CSAW-CC further stratified by tumor size ≤ 15 mm and > 15 mm. Across both datasets, FlexLMV-Net achieves the highest C-index across the cancer-type subgroups. On EMBED, it performs best for both non-invasive and invasive cancers, while on CSAW-CC it achieves the highest performance across all subgroups. Notably, the comparison between LMV-Net and FlexLMV-Net shows a substantial improvement when incorporating a larger number of prior examinations, demonstrating the benefit of FlexLMV-Net’s flexible longitudinal aggregation. LMV-Net also remains competitive, generally outperforming the SOTA methods, with encoder fine-tuning providing further improvements. Overall, these results highlight the benefit of jointly modeling longitudinal and multi-view information, particularly when leveraging more extensive imaging histories, for robust risk prediction across cancer types and tumor characteristics.

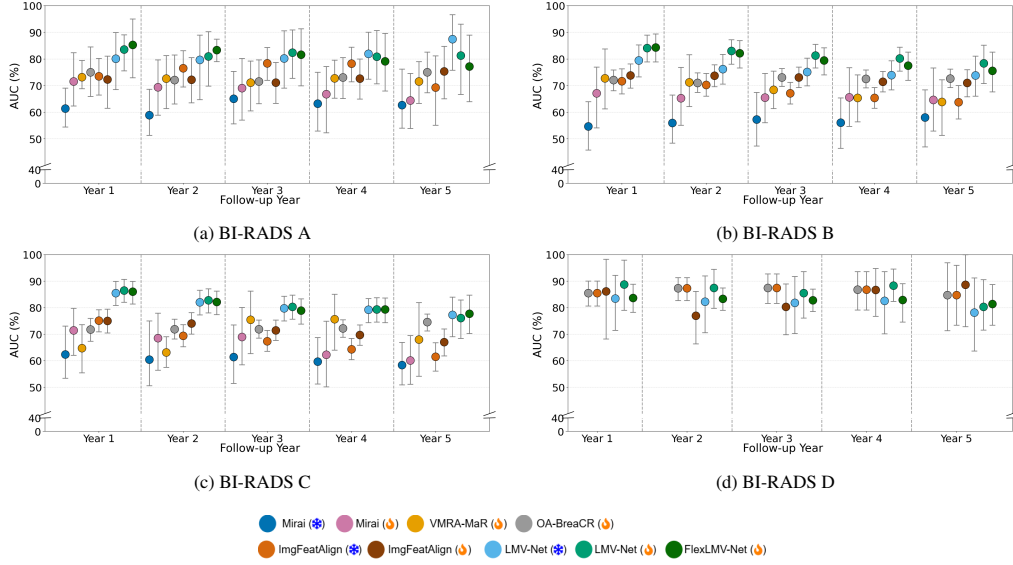


Figure 7: Forest plots showing AUC across five follow-up years for all models stratified by breast density categories (BI-RADS A-D) on the EMBED dataset. Points represent mean AUC and error bars represent 95% confidence intervals.

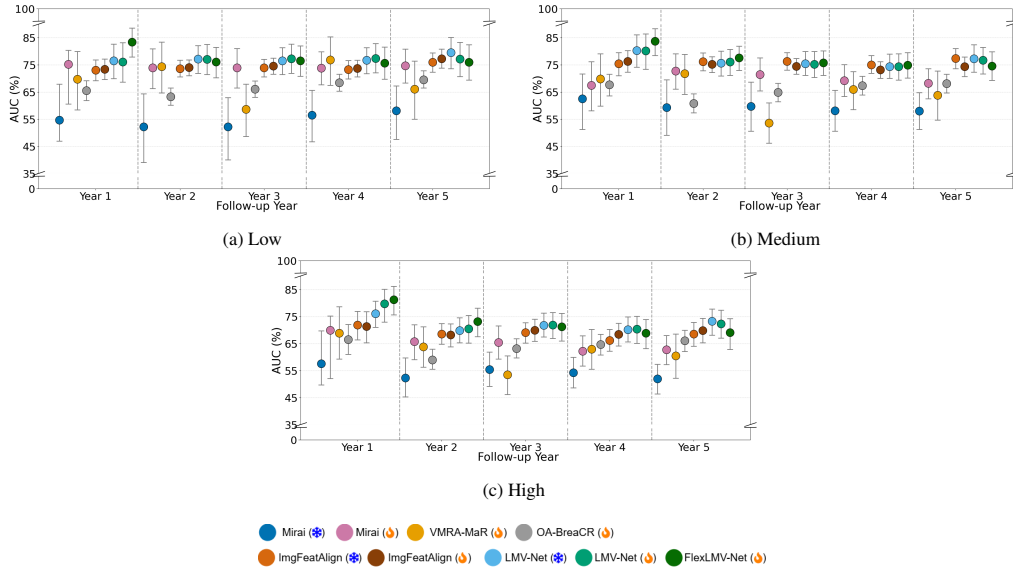


Figure 8: Forest plots showing AUC across five follow-up years for all models stratified by breast density categories (Low, Medium, High) on the CSAW-CC dataset. Points represent mean AUC and error bars represent 95% confidence intervals.

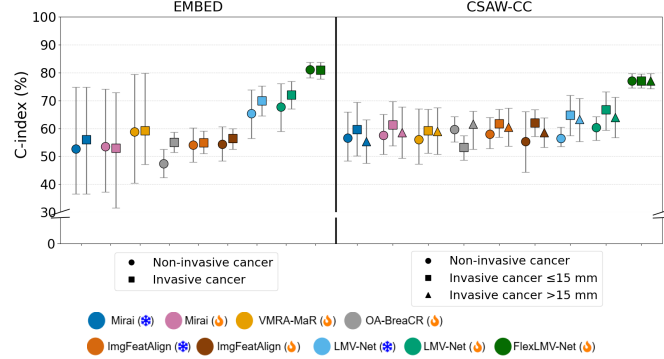


Figure 9: Forest plots showing C-index for all models stratified by cancer types on the EMBED and CSAW-CC dataset. Points represent the mean C-index, and error bars indicate 95% confidence intervals.

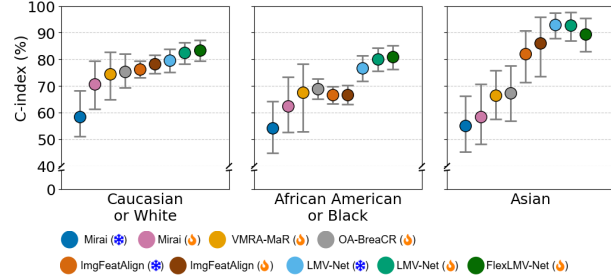


Figure 10: C-index values with 95% confidence intervals for different race categories on the EMBED dataset.

5.4. Performance across demographic subgroups

Figure 10 presents the C-index values with 95% confidence intervals across race categories (Caucasian or White, African American or Black, and Asian) on the EMBED dataset, comparing LMV-Net, FlexLMV-Net, and SOTA models. Across all race categories, the proposed longitudinal models generally achieve higher C-index values than the SOTA methods. FlexLMV-Net achieves the highest or near-highest performance across the evaluated groups. LMV-Net also performs competitively across all race categories, generally outperforming the SOTA methods. These results suggest that incorporating longitudinal information together with multi-view representations can improve risk prediction across diverse populations. Performance differences between models are more pronounced for the African American or Black and Asian groups than for the Caucasian or White group.

5.5. Generalization performance

We evaluate the generalization performance of LMV-Net, FlexLMV-Net, and the best baseline, ImgFeatAlign. Models are trained on either EMBED or CSAW-CC and evaluated both in-domain and cross-dataset to assess robustness to domain shift. In addition to overall C-index and AUC for the 1–5 year risk prediction tasks (Table 2), we evaluate C-index across breast density, cancer subtype, and race categories (Figures 11a–12). Table 2 shows that LMV-Net (🔥) maintains the strongest cross-dataset performance, particularly when trained on EMBED and evaluated on

Table 2: 1–5 year breast cancer risk prediction performance measured by C-index and time-dependent AUC (mean $\pm$ 95% CI). Models were trained on either the EMBED or CSAW-CC dataset and evaluated both on the same dataset (in-domain) and on the other dataset (cross-dataset) to assess generalization performance. Additionally, we report the cross-dataset performance drop in C-index (Δ (%) $\downarrow$), defined as the difference between in-domain and cross-dataset performance, to quantify the degradation in predictive accuracy under domain shift. ✱ : frozen encoder; ✴ : fine-tuned encoder.

Dataset		Method	C-index (%) $\uparrow$	Δ (%) $\downarrow$	Follow-up year AUC (%) $\uparrow$				
Training	Test				1-Y	2-Y	3-Y	4-Y	5-Y
EMBED	EMBED	ImgFeatAlign (✱)	74.7 $\pm$ 2.4		75.0 $\pm$ 2.9	75.5 $\pm$ 2.4	75.3 $\pm$ 2.1	75.9 $\pm$ 2.3	72.5 $\pm$ 3.2
		ImgFeatAlign (✴)	72.2 $\pm$ 2.7		72.1 $\pm$ 3.0	73.3 $\pm$ 2.6	72.2 $\pm$ 2.5	71.3 $\pm$ 2.4	72.6 $\pm$ 3.3
		LMV-Net (✱)	78.3 $\pm$ 3.0		80.8 $\pm$ 3.7	78.3 $\pm$ 3.5	77.5 $\pm$ 3.3	76.6 $\pm$ 3.2	75.1 $\pm$ 4.9
		LMV-Net (✴)	81.4 $\pm$ 2.8		83.9 $\pm$ 3.4	82.1 $\pm$ 3.0	80.8 $\pm$ 2.9	80.6 $\pm$ 2.8	77.2 $\pm$ 4.7
		FlexLMV-Net (✴)	81.2 $\pm$ 2.9		84.1 $\pm$ 3.2	82.0 $\pm$ 3.0	79.6 $\pm$ 3.1	79.3 $\pm$ 3.1	79.0 $\pm$ 4.4
CSAW-CC	EMBED	ImgFeatAlign (✱)	64.7 $\pm$ 2.7	-10.0	66.3 $\pm$ 3.0	64.8 $\pm$ 2.8	65.1 $\pm$ 2.5	66.3 $\pm$ 2.5	64.1 $\pm$ 4.0
		ImgFeatAlign (✴)	61.7 $\pm$ 2.5	-10.5	64.0 $\pm$ 2.8	62.9 $\pm$ 2.6	63.2 $\pm$ 2.5	64.1 $\pm$ 2.6	65.3 $\pm$ 3.8
		LMV-Net (✱)	69.2 $\pm$ 3.3	-9.1	71.8 $\pm$ 3.7	69.6 $\pm$ 3.6	69.2 $\pm$ 3.3	70.5 $\pm$ 3.4	72.0 $\pm$ 2.7
		LMV-Net (✴)	73.7 $\pm$ 3.3	-7.7	76.7 $\pm$ 3.9	74.4 $\pm$ 3.6	73.8 $\pm$ 3.3	74.0 $\pm$ 3.3	75.5 $\pm$ 2.8
		FlexLMV-Net (✴)	72.9 $\pm$ 3.8	-8.3	75.5 $\pm$ 4.6	73.2 $\pm$ 4.0	72.4 $\pm$ 3.5	71.1 $\pm$ 3.7	70.3 $\pm$ 4.6
CSAW-CC	CSAW-CC	ImgFeatAlign (✱)	70.4 $\pm$ 1.9		72.0 $\pm$ 2.4	71.5 $\pm$ 1.9	72.6 $\pm$ 1.9	72.0 $\pm$ 2.1	75.2 $\pm$ 2.3
		ImgFeatAlign (✴)	71.9 $\pm$ 2.0		74.1 $\pm$ 2.6	72.3 $\pm$ 2.0	73.1 $\pm$ 1.9	72.1 $\pm$ 2.0	74.5 $\pm$ 2.3
		LMV-Net (✱)	73.4 $\pm$ 2.5		77.5 $\pm$ 3.3	73.5 $\pm$ 2.6	74.0 $\pm$ 2.5	73.1 $\pm$ 2.6	75.9 $\pm$ 2.8
		LMV-Net (✴)	74.4 $\pm$ 2.7		78.7 $\pm$ 3.6	74.1 $\pm$ 2.9	74.3 $\pm$ 2.7	73.5 $\pm$ 2.8	75.0 $\pm$ 3.1
		FlexLMV-Net (✴)	77.1 $\pm$ 2.5		82.5 $\pm$ 3.0	74.6 $\pm$ 3.0	73.7 $\pm$ 2.9	71.9 $\pm$ 3.1	72.3 $\pm$ 3.3
EMBED	CSAW-CC	ImgFeatAlign (✱)	63.6 $\pm$ 2.1	-6.8	64.9 $\pm$ 2.5	64.7 $\pm$ 2.1	65.2 $\pm$ 2.0	64.8 $\pm$ 2.1	66.5 $\pm$ 2.4
		ImgFeatAlign (✴)	64.7 $\pm$ 2.1	-7.2	66.8 $\pm$ 2.7	64.5 $\pm$ 2.1	65.0 $\pm$ 2.1	64.6 $\pm$ 2.2	65.3 $\pm$ 2.5
		LMV-Net (✱)	70.7 $\pm$ 2.9	-2.7	73.8 $\pm$ 3.5	71.0 $\pm$ 2.9	70.9 $\pm$ 2.9	70.7 $\pm$ 3.0	72.6 $\pm$ 2.9
		LMV-Net (✴)	72.9 $\pm$ 2.9	-1.5	75.7 $\pm$ 3.8	73.1 $\pm$ 3.2	72.4 $\pm$ 2.9	71.5 $\pm$ 2.9	73.0 $\pm$ 2.3
		FlexLMV-Net (✴)	72.5 $\pm$ 4.1	-4.6	75.0 $\pm$ 4.0	72.1 $\pm$ 3.3	71.9 $\pm$ 3.1	70.9 $\pm$ 3.2	71.8 $\pm$ 3.1

CSAW-CC. FlexLMV-Net achieves competitive cross-dataset performance and consistently improves in-domain performance over LMV-Net, highlighting the benefit of incorporating more prior examinations through flexible longitudinal aggregation. However, this additional history does not consistently improve cross-dataset performance, suggesting that domain shift can limit its benefits. ImgFeatAlign generally shows a larger degradation under cross-dataset evaluation. Unlike ImgFeatAlign, which operates at the image level, LMV-Net and FlexLMV-Net model breast-level representations while explicitly integrating multi-view and longitudinal information, potentially providing greater robustness to domain shift. Similar trends are observed across density, cancer subtype, and race subgroups. Fine-tuning generally improves cross-dataset performance, while models trained on EMBED tend to generalize more consistently to CSAW-CC than in the reverse direction. Overall, LMV-Net and FlexLMV-Net demonstrate competitive cross-dataset generalization and outperform the baseline in several settings, with the broader diversity of EMBED potentially contributing to greater robustness to the CSAW-CC domain.

5.6. Attention map analysis

Figure 13 presents representative examples from four patients, showing mammographic images (CC and MLO views) alongside the corresponding self- and cross-attention maps produced by LMV-Net. The attention maps are overlaid on the images to highlight the regions where the proposed Dual Stream Attention block focuses. Across all cases, the model consistently assigns higher attention to suspicious areas, with particular emphasis on tumor regions (indicated by red bounding boxes provided by clinicians). Self-attention captures intra-view dependencies by concentrating on the lesion and its surrounding tissue within each view, while cross-attention leverages complementary information between CC and MLO views, attending to anatomically corresponding regions in the complementary view. Notably, this behavior is observed for tumors

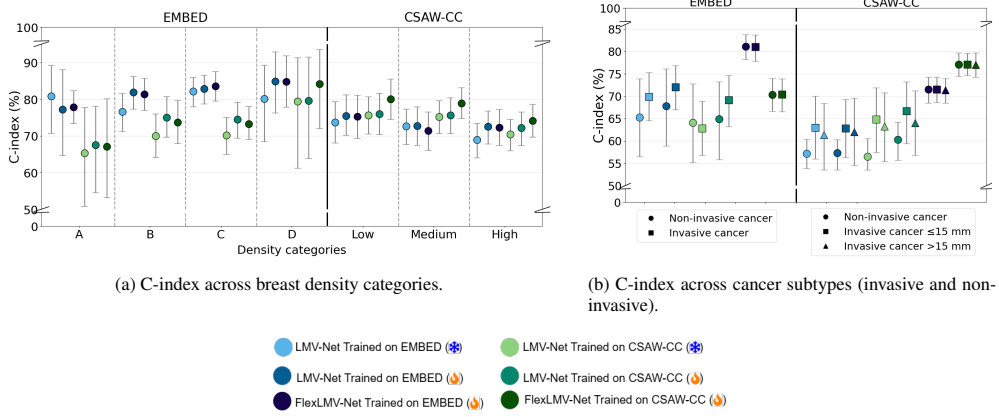


Figure 11: Generalization performance of models across (a) density categories and (b) cancer subtypes. Models are evaluated both in-domain (trained and tested on the same dataset) and cross-dataset (trained on one dataset and tested on the other) to assess generalization. Points show mean C-index and error bars indicate 95% confidence intervals.

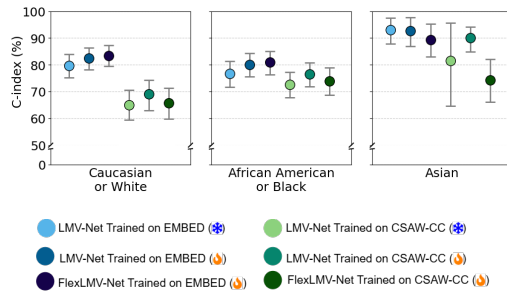


Figure 12: C-index across race categories for models evaluated on the EMBED dataset. Models were trained either on EMBED (in-domain) or CSAW-CC (cross-dataset). Points represent the mean C-index, and error bars indicate the 95% confidence interval.

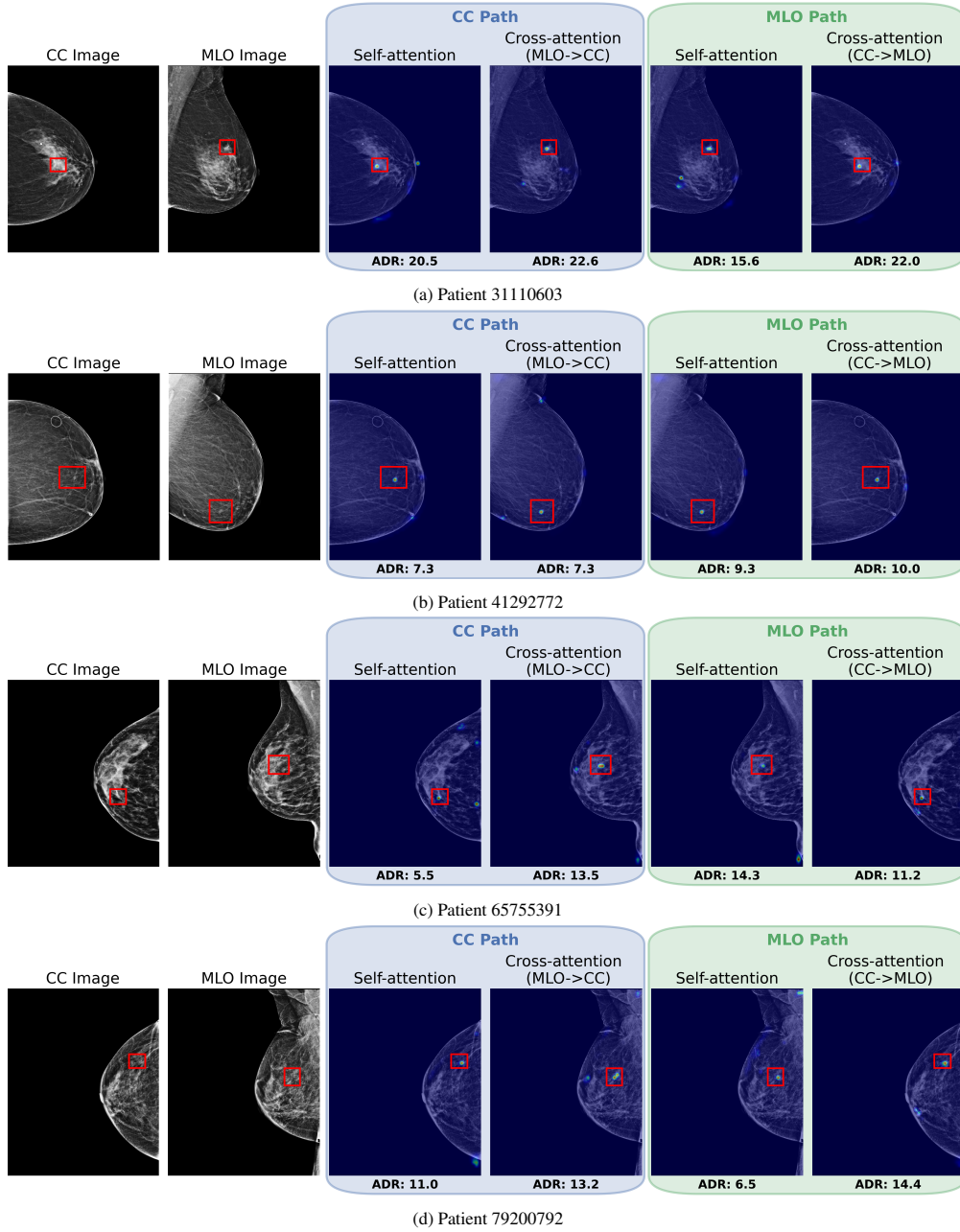


Figure 13: Attention maps across multiple cases showing that FlexLMV-Net correctly focuses on tumor regions (highlighted by red boxes).

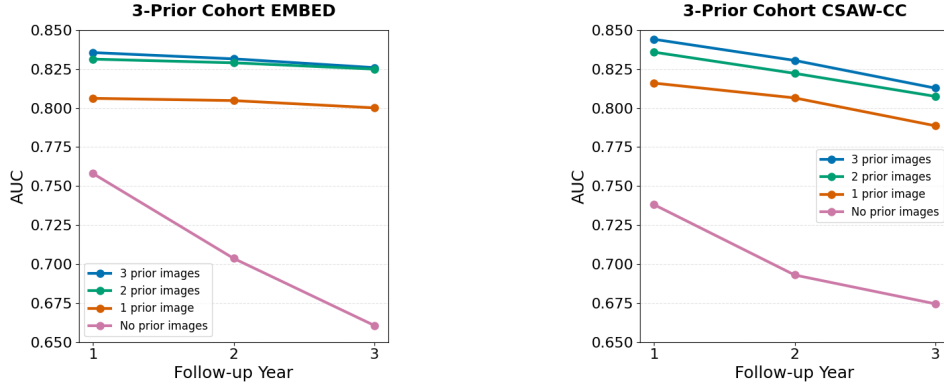


Figure 14: Effect of longitudinal imaging history on breast cancer risk prediction. AUC is shown across 1-, 2-, and 3-year follow-up horizons using up to three prior mammographic examinations.

in both the left and right breasts (two patients each), suggesting that the model robustly identifies clinically relevant regions regardless of laterality.

We further compute the ratio of the attention density within the tumor bounding box to that within the entire breast region, referred to as the Attention Density Ratio (ADR). Inspired by the Relevance Mass Accuracy metric of (Arras et al., 2022), ADR quantitatively measures the concentration of model attention within the tumor region. On average, the attention density within the tumor region is 8–12 $\times$ higher than in the surrounding breast tissue. Across the four examples, the ADR is generally higher for cross-attention than for self-attention. This finding suggests that the proposed Dual Stream Attention block more effectively concentrates attention on anatomically corresponding tumor regions by leveraging information from the complementary view.

5.7. Effect of longitudinal history length on risk prediction

To assess the value of longitudinal imaging history, we evaluated the model in patients with at least three prior mammographic examinations while varying the number of prior images provided as input. As shown in Figure 14, incorporating longitudinal imaging history consistently improved breast cancer risk prediction compared with using no prior images. The benefit was particularly pronounced at longer follow-up horizons, where performance declined substantially when prior imaging was omitted. Performance largely plateaued when using two or three prior examinations, suggesting that even a limited longitudinal history provides substantial value for breast cancer risk prediction.

5.8. Ablation study

We conducted ablation studies (Table 3) to assess the contribution of key components of FlexLMV-Net. The following configurations were tested: (i) *Repeating current features*: temporal alignment and subtraction are removed, and the current features are concatenated three times; (ii) *Concatenating with zeros*: temporal alignment and subtraction are removed and two zero tensors are concatenated with the current features; (iii) *Single-view modeling*: one view branch is removed and cross-attention is replaced with self-attention; (iv) *Duplicated-view modeling*: the same view is used in both branches; (v) *Dual-stream attention removed*: the dual-stream attention

Table 3: Ablation study for 1–5 year risk: mean C-index and AUC on EMBED and CSAW-CC. MP: Using multiple prior images (up to 3), MV: Multi-view, DV: Different views in both branches, FT: Encoder fine-tuning, DSA: Dual Stream Attention.

Variant	Prior	MP	MV	DV	FT	DSA	Y_{MLO} Y_{CC}	C-index (%) $\uparrow$	Follow-up year AUC (%) $\uparrow$				
									1-Y	2-Y	3-Y	4-Y	5-Y
EMBED	(i) No prior (current repeated)	$\times$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	77.8	80.3	78.1	76.9	76.3	73.6
	(ii) No prior (zero-padded)	$\times$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	77.5	80.2	78.3	76.4	76.2	73.8
	(iii) Only one view	$\checkmark$	$\times$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	70.8	72.2	70.8	68.9	65.9	61.9
	(iv) Same view both branches	$\checkmark$	$\times$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	77.4	78.8	78.1	77.7	76.2	74.7
	(v) Without DSA	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	72.8	73.9	73.7	72.8	70.3	65.8
	(vi) Use only $Y_{multi-view}$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	80.1	83.0	80.4	79.4	78.8	76.5
	(vii) No finetuning	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	78.3	80.8	78.3	77.5	76.6	75.1
	(viii) One Prior	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	81.4	83.9	82.1	80.8	80.0	77.2
	(ix) Multiple Prior	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	81.2	84.1	82.0	79.6	79.3	79.0
CSAW-CC	(i) No prior (current repeated)	$\times$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	73.0	75.3	73.8	73.2	72.3	73.2
	(ii) No prior (zero-padded)	$\times$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	72.2	75.2	73.0	72.6	71.9	72.4
	(iii) Only one view	$\checkmark$	$\times$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	73.3	73.8	73.1	73.5	73.1	73.9
	(iv) Same view both branches	$\checkmark$	$\times$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	72.2	75.2	72.6	72.5	71.1	73.0
	(v) Without DSA	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\times$	$\checkmark$	69.5	70.5	70.0	70.6	69.3	71.4
	(vi) Use only $Y_{multi-view}$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\times$	73.3	77.6	73.3	73.4	72.6	74.4
	(vii) No finetuning	$\checkmark$	$\times$	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	73.4	77.5	73.5	74.0	73.1	75.9
	(viii) One Prior	$\checkmark$	$\times$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	74.4	78.7	74.1	74.3	73.5	75.0
	(ix) Multiple Prior	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	77.1	82.5	74.6	73.7	71.9	72.3

block is removed; (vi) *Multi-view learning*: only $Y_{multi-view}$ is used; (vii) *Encoder fine-tuning*: the pretrained Mirai backbone is frozen; and (viii) *Single prior*: only one prior examination is used, corresponding to LMV-Net where we don’t have the feature-aware temporal fusion block; and (ix) *Multiple priors*: the full FlexLMV-Net configuration uses up to three prior examinations.

The results demonstrate that removing longitudinal information, multi-view modeling, dual-stream attention, or encoder fine-tuning consistently reduces performance. Using a single prior also performs worse than the full FlexLMV-Net, highlighting the benefit of incorporating multiple prior examinations. Overall, the full model achieves the best performance on both EMBED and CSAW-CC, demonstrating the complementary contribution of its key components.

6. Limitations and future works

While LMV-Net and FlexLMV-Net demonstrate promising results, several limitations remain. First, further validation on independent multi-center and prospective cohorts is needed to assess generalizability across populations, imaging protocols, and clinical settings. Second, the explicit feature-space alignment depends on accurate spatial correspondence, and alignment errors may propagate to risk prediction. Future work could explore more robust, anatomy-aware, or jointly learned alignment strategies. Third, although FlexLMV-Net supports a variable number of prior examinations, future work could investigate longer screening histories and strategies for optimally selecting and weighting prior examinations. Incorporating bilateral breast information is another promising direction, as between-breast asymmetries may provide complementary risk-related signals. Finally, extending the framework to digital breast tomosynthesis (DBT) could leverage its three-dimensional structural information and enable investigation of its complementary value alongside conventional mammography.

7. Conclusion

In this study, we introduced FlexLMV-Net, a longitudinal multi-view framework that supports a variable number of prior examinations through explicit longitudinal feature alignment and

feature-aware temporal fusion of CC and MLO views for breast cancer risk prediction. LMV-Net can be viewed as a special case of FlexLMV-Net with $K = 1$ prior examination, providing a fixed-length formulation of the framework. Both frameworks demonstrated strong performance across the EMBED and CSAW-CC datasets and diverse subgroups, while additional longitudinal history further improved risk prediction. In addition, we introduced BCRBench, an open-source benchmark repository that unifies breast cancer risk prediction models within a standardized framework to facilitate fair and reproducible comparisons. Together, these contributions demonstrate the potential of longitudinal multi-view modeling for breast cancer risk stratification and provide a foundation for future research toward more personalized screening and earlier detection.

CRedit authorship contribution statement

Solveig Thrun: Conceptualization, Software, Methodology, Formal analysis, Investigation, Visualization, Writing - original draft, Writing - review & editing. **Zijun Sun:** Methodology, Writing - review & editing. **Suaiba A. Salahuddin:** Methodology, Writing - review & editing. **Kristoffer Wickstrøm:** Conceptualization, Methodology, Writing - review & editing, Supervision. **Elisabeth Wetzer:** Conceptualization, Methodology, Writing - review & editing, Supervision. **Stine Hansen:** Conceptualization, Methodology, Writing - review & editing, Supervision. **Robert Jenssen:** Conceptualization, Methodology, Writing - review & editing, Supervision, Funding acquisition. **Michael Kampffmeyer:** Conceptualization, Methodology, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The source code of FlexLMV-Net is publicly available at <https://github.com/sot176/FlexLMV-Net> and the proposed BCRBench benchmarking repository, which includes SOTA breast cancer risk prediction models, is available at <https://github.com/sot176/BCRBench>. All datasets used in this work are publicly accessible and links to request access are provided in the Experimental Setup section.

Acknowledgements

This work was supported by The Research Council of Norway (Visual Intelligence, grant no. 309439 as well as FRIPRO grant no. 315029 and 360068 and IKTPLUSS grant no. 303514). We acknowledge Sigma2 – the National Infrastructure for High Performance Computing and Data Storage in Norway, for awarding this project access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium through Sigma2 [grant no. NN8106K]. We would like to thank the Research Software Engineering (RSE) Group at UiT The Arctic University of Norway for their valuable support and advice in designing and implementing the unified GitHub repository BCRBench.

Appendix A. Dataset Details and Dataset Splits

Tables A.4 and A.5 summarize the Time-to-Cancer label distribution, dataset split, and breast density distribution for the EMBED and CSAW-CC datasets. For EMBED, breast density labels are unavailable for some cases. Therefore, the number of cases reported in the density distribution table does not match the total number of cases reported in the Time-to-Cancer table.

Table A.4: Time-to-Events label distribution and dataset splits for breast-level CC+MLO exam pairs. Percentages are reported relative to the total number of cases per split.

Dataset	Split	Positive: Time-to-Events					Negative	
		1Y	2Y	3Y	4Y	5Y	All	>5Y
EMBED	Train	370 (2.9%)	96 (0.8%)	83 (0.6%)	46 (0.4%)	34 (0.3%)	629 (4.9%)	12262 (95.1%)
	Val	146 (2.9%)	32 (0.6%)	29 (0.6%)	20 (0.4%)	14 (0.3%)	241 (4.7%)	4907 (95.3%)
	Test	220 (2.9%)	76 (1.0%)	54 (0.7%)	37 (0.5%)	29 (0.4%)	416 (5.4%)	7341 (94.6%)
	Total	736 (2.9%)	204 (0.8%)	166 (0.7%)	103 (0.4%)	77 (0.3%)	1286 (5.0%)	24510 (95.0%)
CSAW-CC	Train	352 (2.1%)	350 (2.1%)	94 (0.6%)	224 (1.3%)	88 (0.5%)	1108 (6.6%)	15686 (93.4%)
	Val	140 (2.1%)	142 (2.1%)	48 (0.7%)	92 (1.4%)	28 (0.4%)	450 (6.8%)	6188 (93.2%)
	Test	208 (2.1%)	214 (2.1%)	76 (0.8%)	124 (1.2%)	54 (0.5%)	676 (6.7%)	9402 (93.3%)
	Total	700 (2.1%)	706 (2.1%)	218 (0.7%)	440 (1.3%)	170 (0.5%)	2234 (6.6%)	31276 (93.4%)

Table A.5: Breast density distribution across the EMBED and CSAW-CC datasets for breast-level CC+MLO exam pairs. Percentages are reported relative to the total number of cases per split.

Split	EMBED (BI-RADS)				CSAW-CC		
	A	B	C	D	Low	Medium	High
Train	1435 (11.1%)	5631 (43.7%)	5214 (40.4%)	569 (4.4%)	5776 (34.4%)	5538 (33.0%)	5480 (32.6%)
Val	587 (11.4%)	2164 (42.0%)	2119 (41.2%)	265 (5.1%)	2130 (32.1%)	2266 (34.1%)	2242 (33.8%)
Test	943 (12.2%)	3279 (42.3%)	3174 (40.9%)	344 (4.4%)	3354 (33.3%)	3480 (34.5%)	3244 (32.2%)
Total	2965	11074	10507	1178	11260	11284	10966

References

- Akwo, J.D., Trieu, P.D.Y., Barron, M.L., Reynolds, T., Lewis, S.J., 2024. Access to prior screening mammograms affects the specificity but not sensitivity of radiologists’ performance. *Clinical Radiology* 79(12).
- Arras, L., Osman, A., Samek, W., 2022. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion* 81, 14–40. doi: <https://doi.org/10.1016/j.inffus.2021.11.008>.
- Ba, J.L., Kiros, J.R., and Hinton, G.E., 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*. doi: <https://doi.org/10.48550/arXiv.1607.06450>.

- Bhushan, A., Gonsalves, A., and Menon, J. U., 2021. Current state of breast cancer diagnosis, treatment, and theranostics. *Pharmaceutics*, 13(5), 723. doi: <https://doi.org/10.3390/pharmaceutics13050723>.
- Bodewes, F.T.H., van Asselt, A.A., Dorrius, M.D., Greuter, M.J.W., and de Bock, G.H., 2022. Mammographic breast density and the risk of breast cancer: A systematic review and meta-analysis. *The Breast*, 66, 1–13. doi: <https://doi.org/10.1016/j.breast.2022.09.007>.
- Bond, M., Pavey, T., Welch, K., Cooper, C., Garside, R., Dean, S., Hyde, C.J., 2013. Psychological consequences of false-positive screening mammograms in the UK. *BMJ Evidence-Based Medicine* 18(2), pp. 54–61.
- Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., and Jemal, A., 2024. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3). doi: <https://doi.org/10.3322/caac.21834>.
- Chen, H., Martel, A.L., 2025. Breast cancer detection from multi-view screening mammograms with visual prompt tuning. In: *Artificial Intelligence and Imaging for Diagnostic and Treatment Challenges in Breast Care (Deep-Breath)*, pp. 91-100. doi: https://doi.org/10.1007/978-3-032-05559-0_10.
- Clift, A.K., Dodwell, D., Lord, S., Petrou, S., Brady, S.M., Collins, G.S., Hippisley-Cox, J., 2022. The current status of risk-stratified breast screening. *Br. J. Cancer* 126(4), pp. 533–550. doi: <https://doi.org/10.1038/s41416-021-01550-3>.
- Dadsetan, S., Arefan, D., Berg, W.A., Zuley, M.L., Sumkin, J.H., Wu, S., 2022. Deep learning of longitudinal mammogram examinations for breast cancer risk prediction. *Pattern Recognit.* 132. doi: <https://doi.org/10.1016/j.patcog.2022.108919>.
- Dembrower, K., Lindholm, P., Strand, F., 2020. A multi-million mammography image dataset and population-based screening cohort for the training and evaluation of deep neural networks—the cohort of screen-aged women (CSAW). *J. Digit. Imaging* 33, pp. 408-413. doi: <https://doi.org/10.1007/s10278-019-00278-0>.
- Donnelly, J., Moffett, L., Barnett, A.J., Trivedi, H., Schwartz, F., Lo, J., Rudin, C., 2024. Asym-Mirai: Interpretable mammography-based deep learning model for 1–5-year breast cancer risk prediction. *Radiology* 310(3), e232780. doi: <https://doi.org/10.1148/radiol.232780>.
- Habib, A.R., Grady, D., Redberg, R.F., 2021. Recommendations from breast cancer centers for frequent screening mammography in younger women may do more harm than good. *JAMA Internal Medicine* 181(5), pp. 588–589.
- Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, pp. 29–36.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778. doi: <https://doi.org/10.1109/CVPR.2016.90>.

- Huang, G., Sun, Y., Liu, Z., Sedra, D., and Weinberger, K.Q., 2016. Deep networks with stochastic depth. In: Proceedings of the European Conference on Computer Vision (ECCV).
- Hussain, S., Ali, M., Naseem, U., Nezhadmoghadam, F., Jatoi, M.A., Gulliver, T.A., and Tamez-Peña, J.G., 2024. Breast cancer risk prediction using machine learning: A systematic review. *Frontiers in Oncology*, 14. doi: <https://doi.org/10.3389/fonc.2024.1343627>.
- Jeong, J., Vey, B.L., Bhimireddy, A., Kim, T., Santos, T., Correa, R., Dutt, R., Mosunjac, M., Oprea-Ilie, G., Smith, G., Woo, M., McAdams, C.R., Newell, M.S., Banerjee, I., Gichoya, J., Trivedi, H., 2023. The EMory BrEast imaging Dataset (EMBED): a racially diverse, granular dataset of 3.4 million screening and diagnostic mammographic images. *Radiol. Artif. Intell.* 5, 1. doi: <https://doi.org/10.1148/ryai.220047>.
- Kang, L., Chen, W., Petrick, N.A., and Gallas, B.D., 2015. Comparing two correlated c indices with right-censored survival outcome: a one-shot nonparametric approach. *Stat. Med.*, 34(4), 685–703. doi: <https://doi.org/10.1002/sim.6375>.
- Karaman, B.K., Dodelzon, K., Akar, G.B., Sabuncu, M.R., 2024. Longitudinal mammogram risk prediction. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 437–446. doi: https://doi.org/10.1007/978-3-031-72086-4_41.
- Keller, B.M., Nathan, D.L., Wang, Y., Zheng, Y., Gee, J.C., Conant, E.F., Kontos, D., 2012. Estimation of breast percent density in raw and processed full field digital mammography images via adaptive fuzzy c-means clustering and support vector machine segmentation. *Med. Phys.* 39, pp. 4903–4917. doi: <https://doi.org/10.1118/1.4736530>.
- Lee, H., Kim, J., Park, E., Kim, M., Kim, T., Kooi, T., 2023. Enhancing breast cancer risk prediction by incorporating prior images. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 389–398. doi: https://doi.org/10.1007/978-3-031-43904-9_38.
- Loshchilov, I. and Hutter, F., 2019. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.
- McCann, J., Stockton, D., Godward, S., 2002. Impact of false-positive mammography on subsequent screening attendance and risk of cancer. *Breast Cancer Research* 4(5), R11. doi: <https://doi.org/10.1186/bcr455>.
- Meng, M., Bi, L., Fulham, M., Feng, D., Kim, J., 2023. Non-iterative coarse-to-fine transformer networks for joint affine and deformable image registration. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 750–760. doi: https://doi.org/10.1007/978-3-031-43999-5_71.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: an imperative style, high-performance deep learning library. In: *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 32, pp. 8024–8035.

- Ren, W., Chen, M., Qiao, Y., Zhao, F., 2022. Global guidelines for breast cancer screening: A systematic review. *The Breast* 64, pp. 85–99. doi: <https://doi.org/10.1016/j.breast.2022.04.003>.
- Rubio, I.T., Drukker, C.A., and Esgueva, A., 2024. Risk-based breast cancer screening: what are the challenges? *Tumori J.* doi: <https://doi.org/10.1177/03008916241306971>.
- Sarker, S., Sarker, P., Bebis, G., Tavakkoli, A., 2024. MV-Swin-T: Mammogram classification with multi-view Swin transformer. *arXiv preprint arXiv:2402.16298*. doi: <https://arxiv.org/abs/2402.16298>.
- Sickles, E.A., ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System, American College of Radiology, 2013.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15, 1929–1958. <http://jmlr.org/papers/v15/srivastava14a.html>.
- Sun, Z., Thrun, S., Kampffmeyer, M., 2025. VMRA-MaR: An Asymmetry-Aware Temporal Framework for Longitudinal Breast Cancer Risk Prediction. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp 660 - 669. doi: https://doi.org/10.1007/978-3-032-05182-0_64.
- Thrun, S., Hansen, S., Sun, Z., Blum, N., Salahuddin, S. A., Wang, X., Wickstrøm, K., Wetzer, E., Jenssen, R., Stille, M., and Kampffmeyer, M., 2025. The impact of longitudinal mammogram alignment on breast cancer risk assessment. *arXiv preprint arXiv:2511.08328*.
- Thrun, S., Hansen, S., Sun, Z., Blum, N., Salahuddin, S.A., Wickstrøm, K., Wetzer, E., Jenssen, R., Stille, M., Kampffmeyer, M., 2025. Reconsidering Explicit Longitudinal Mammography Alignment for Enhanced Breast Cancer Risk Prediction. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp 495 - 505. doi: https://doi.org/10.1007/978-3-032-04937-7_47.
- Thrun, S., Sun, Z., Salahuddin, S.A., Wickstrøm, K., Wetzer, E., Hansen, S., Jenssen, R., Kampffmeyer, M., 2026. Longitudinal Multi-View Breast Cancer Risk Prediction. In: *Medical Image Computing and Computer Assisted Intervention*. pp 530 - 540. doi: https://doi.org/10.1007/978-3-032-38470-6_50.
- Tyrer, J., Duffy, S. W., and Cuzick, J., 2004. A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in Medicine* 23(7), 1111–1130. doi: <https://doi.org/10.1002/sim.1668>.
- Uno, H., Cai, T., Pencina, M.J., D’Agostino, R.B., Wei, L.J., 2011. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat. Med.* 30(10). doi: <https://doi.org/10.1002/sim.4154>.
- van Tulder, G., Tong, Y., and Marchiori, E., 2021. Multi-View Analysis of Unregistered Medical Images Using Cross-View Transformers. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 104–113. doi: https://doi.org/10.1007/978-3-030-87199-4_10.

- van Winkel, S.L., Samperna, R., Loehrer, E.A., Kroes, J., Rodriguez-Ruiz, A., Mann, R.M., 2025. Using AI to Select Women with Intermediate Breast Cancer Risk for Breast Screening with MRI. *Radiology* 314(2), e233067. doi: <https://doi.org/10.1148/radiol.233067>.
- Varela, C., Karssemeijer, N., Hendriks, J.H.C.L., Holland, R., 2005. Use of prior mammograms in the classification of benign and malignant masses. *European Journal of Radiology* 56(2).
- Wang, A.T., Vachon, C.M., Brandt, K.R., and Ghosh, K., 2014. Breast density and breast cancer risk: A practical review. *Mayo Clin. Proc.*, 89(4), 548–557. doi: <https://doi.org/10.1016/j.mayocp.2013.12.014>.
- Wang, X., Tan, T., Gao, Y., Su, R., Zhang, T., Han, L., Teuwen, J., D'Angelo, A., Drukker, C.A., Schmidt, M.K., Beets-Tan, R., Karssemeijer, N., Mann, R., 2023. Predicting up to 10 year breast cancer risk using longitudinal mammographic screening history. medRxiv.
- Wang, X., Tan, T., Gao, Y., Marcus, E., Han, L., Portaluri, A., Zhang, T., Lu, C., Liang, X., Beets-Tan, R., Teuwen, J., and Mann, R., 2024. Ordinal learning: longitudinal attention alignment model for predicting time to future breast cancer events from mammograms. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 155–165. doi: https://doi.org/10.1007/978-3-031-72378-0_15.
- Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzębski, S., Févry, T., Katsnelson, J., Kim, E., Wolfson, S., Parikh, U., Gaddam, S., Lin, L.Y., Ho, K., Weinstein, J.D., Reig, B., Gao, Y., Toth, H., Pysarenko, K., Lewin, A., Lee, J., Airola, K., Mema, E., Chung, S., Hwang, E., Samreen, N., Kim, S.G., Heacock, L., and Moy, L., 2020. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE Transactions on Medical Imaging* 39(4), 1184–1194. doi: <https://doi.org/10.1109/TMI.2019.2945514>.
- Yala, A., Mikhael, P. G., Strand, F., Lin, G., Smith, K., Wan, Y.-L., Lamb, L., Hughes, K., Lehman, C., and Barzilay, R., 2021. Toward robust mammography-based models for breast cancer risk. *Science Translational Medicine* 13(578), eaba4373. doi: <https://doi.org/10.1126/scitranslmed.aba4373>.
- Yeoh, H. H., Liew, A., Phan, R., Strand, F., Rahmat, K., Nguyen, T. L., and Tan, M., 2025. RAD-IFUSION: a multi-radiomics deep learning based breast cancer risk prediction model using sequential mammographic images with image attention and bilateral asymmetry refinement. *Physics in Medicine & Biology* 70(23). doi: <https://doi.org/10.1088/1361-6560/ae1ee4>.
- Yeoh, H.H., Strand, F., Phan, R., Rahmat, K., Tan, M., 2026. A new time-decay radiomics integrated network (TRINet) for breast cancer risk prediction. *Med. Image Anal.* 107, 103829. doi: <https://doi.org/10.1016/j.media.2025.103829>.
- Zielonke, N., Gini, A., Jansen, E.E.L., Anttila, A., Segnan, N., Ponti, A., Veerus, P., de Koning, H.J., van Ravesteyn, N.T., Heijnsdijk, E.A.M., et al., 2020. Evidence for reducing cancer-specific mortality due to screening for breast cancer in Europe: A systematic review. *Eur. J. Cancer*, 127. doi: <https://doi.org/10.1016/j.ejca.2019.12.010>.