# LUMIN: A Longitudinal Multi-modal Knowledge Decomposition Network for Predicting Breast Cancer Recurrence

**Chunyao Lu[1,2*], Tianyu Zhang[1,2*], Xinglong Liang[1,2], Yuan Gao[1,3], Luyi Han[1,2], Xin Wang[1,3], Nika Rasoolzadeh[1,2], Tao Tan[4†], Ritse Mann[1,2]**

[1]The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX, Amsterdam, The Netherlands
[2]Radboud University Medical Centre, Geert Grooteplein 10, 6525 GA, Nijmegen, The Netherlands
[3]Maastricht University Medical Centre, P. Debyelaan 25, 6202 AZ, Maastricht, The Netherlands
[4]Faculty of Applied Sciences, Macao Polytechnic University, 999078, Macao, China
c.lu@nki.nl, t.zhang@nki.nl, x.liang@nki.nl, y.gao@nki.nl, hanluyi4869@gmail.com, x.wang@nki.nl, nika.rasoolzadeh@radboudumc.nl, taotanjs@gmail.com, Ritse.Mann@radboudumc.nl

## Abstract

Accurate prediction of breast cancer recurrence after treatment is essential for improving long-term outcomes. However, existing models are limited by three key challenges: (1) they typically rely on single-modal data, missing cross-modal interactions; (2) they analyze static snapshots, failing to capture disease progression over time; and (3) they often perform coarse feature fusion, lacking semantic disentanglement and interpretability. To address these issues, we propose LUMIN (Longitudinal Multi-modal Knowledge Decomposition Network), a novel framework that integrates longitudinal mammograms and electronic health records (EHRs) for recurrence prediction. LUMIN leverages a vision-language contrastive pretraining backbone to align multi-modal representations and introduces two knowledge extraction modules: (1) a Cross-Modal Disentangled Knowledge Extractor (CM-DKE) that separates shared, complementary, and modality-specific information across imaging and text; and (2) a Temporal Evolution Disentangled Knowledge Extractor (TE-DKE) that captures time-invariant, time-varying, and time-specific features to model disease dynamics. Experiments on a large-scale dataset of 3,924 patients and 19,684 exams show that LUMIN significantly outperforms state-of-the-art baselines, demonstrating its effectiveness in capturing both multi-modal semantics and temporal heterogeneity for recurrence prediction.

**Code** — https://github.com/Robin54223/LUMIN

## Introduction

Breast cancer is the most commonly diagnosed cancer in women and a leading cause of cancer-related mortality worldwide (Zardavas et al. 2015; Wang et al. 2019). Despite advances in early detection and treatment improving overall survival rates, recurrence and metastasis remain the primary contributors to breast cancer-related deaths (Colleoni et al. 2016; Duffy et al. 2021). Current follow-up strategies predominantly rely on standardized imaging surveillance, including mammography or digital breast tomosynthesis (DBT) (Chlebowski, Aragaki, and Pan 2021; Sprague et al. 2023). However, under this conventional approach, many recurrences are still detected as interval cancers—cases that develop between scheduled screenings—underscoring the critical need for more personalized and effective surveillance strategies.

Fig. 1 illustrates the typical longitudinal monitoring process, where patients undergo repeated mammograms and clinical evaluations over time. While early follow-up exams may appear normal, recurrence can emerge at later time points, often outside routine intervals. This motivates the development of predictive models capable of leveraging both temporal and multi-modal information to anticipate recurrence before it becomes clinically apparent.

Several approaches have been explored to predict breast cancer recurrence using traditional statistical and machine-learning models (El Haji et al. 2023; Lou et al. 2020). However, their performance has been limited by constraints in available data, the need for well-annotated clinical information, and the lack of temporal information, which hinders their ability to fully capture the heterogeneity and progression of cancer. Deep learning-based multi-modal fusion models have demonstrated improved predictive performance (Zhang et al. 2023a; Steyaert et al. 2023). For example, recent research (Yang et al. 2022; Howard et al. 2023) has demonstrated that combining histopathological images with clinical data enhances the accuracy of breast cancer risk prediction. However, these models rely on images captured at a single time point, failing to reflect tumor progression over time and limiting their ability to capture its dynamic characteristics. Recently, deep learning models incorporating longitudinal screening images have demonstrated superior accuracy in predicting primary breast cancer risk. Evidence suggests that leveraging longitudinal data improves predictive performance compared to static imaging (Dadsetan et al. 2022; Wang et al. 2023; Yeoh et al. 2023; Damiani et al. 2023; Wang et al. 2024; Karaman et al. 2024). Despite these advancements, no studies have yet explored the po-

*These authors contributed equally.
†Corresponding author:taotanjs@gmail.com

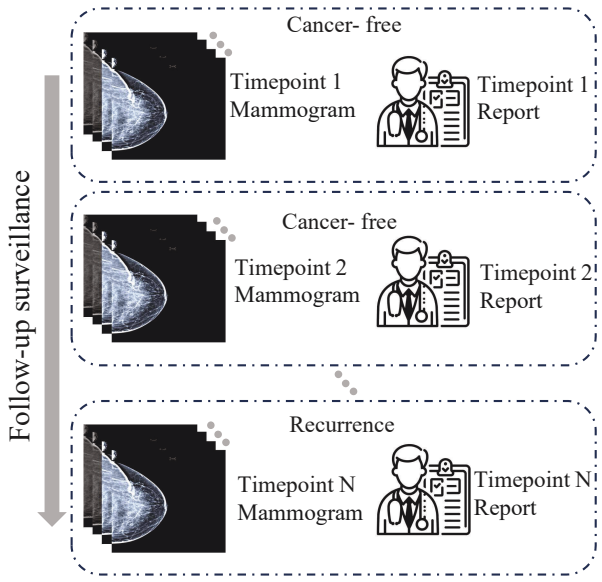


Figure 1: Illustration of longitudinal breast cancer surveillance. Patients undergo repeated mammograms and clinical assessments over time. While early timepoints may appear cancer-free, recurrence can emerge during later follow-up. This motivates the need for predictive models that leverage temporal and multi-modal information for early recurrence detection.

tential of screening data for predicting breast cancer recurrence. This research gap is largely attributed to the challenge of capturing relevant tissue or lesion-related information, as the original tumor site often resolves following treatment. Thus, developing a screening-based recurrence prediction model remains challenging, particularly due to the lack of well-annotated datasets.

In this paper, to address the aforementioned challenges, we introduce LUMIN (Longitudinal Multi-modal Knowledge Decomposition Network), a novel predictive framework for breast cancer recurrence that effectively integrates longitudinal mammograms and EHRs. Our key contributions are as follows: (1) We leverage a pre-trained contrastive learning framework to ensure robust feature alignment between longitudinal mammograms and corresponding textual EHR data. (2) We propose two novel knowledge extraction modules: CM-DKE, which extracts shared, complementary, text-specific, and image-specific knowledge to enhance cross-modal interaction; and TE-DKE, which extracts time-invariant, time-varying, Time1-specific, and Time2-specific knowledge to effectively model disease progression. (3) We evaluate LUMIN on a dataset of 3,924 patients and 19,684 mammograms, demonstrating significant improvements in predictive performance. These findings underscore LUMIN's potential to advance personalized breast cancer surveillance and enhance early recurrence detection.

# Materials and Methods

## Model development

**Pre-Training Stage: Dual Direction Alignment** Fig. 2 presents an overview of our model framework, comprising two main stages: pre-training and downstream prediction. In scenarios with limited annotated clinical data, vision-language pre-training models—such as CLIP (Gao et al. 2025), BLIP (Li et al. 2022), Flamingo (Alayrac et al. 2022), and LLaMA (Touvron et al. 2023)—have demonstrated strong generalization capabilities by leveraging large-scale image–text pairs to learn transferable representations. These models exploit semantic correspondences between modalities, thereby reducing the dependency on labor-intensive manual labeling in domain-specific applications like medical imaging.

Motivated by these advances, we propose a dual-modal alignment (DMA) framework that learns a shared latent space where visual and textual features are semantically aligned. To this end, we employ a symmetric contrastive loss that enforces bidirectional alignment between image and text embeddings. Specifically, for a mini-batch of $N$ matched image–text pairs $\{(I_i, T_i)\}_{i=1}^{N}$, we first encode the inputs using modality-specific encoders $f_I(\cdot)$ and $f_T(\cdot)$, producing normalized embeddings:

$$\hat{v}_i = \frac{f_I(I_i)}{\|f_I(I_i)\|}, \quad \hat{t}_j = \frac{f_T(T_j)}{\|f_T(T_j)\|}. \tag{1}$$

The cosine similarity between an image $I_i$ and a text $T_j$ is defined as:

$$s_{i,j} = \hat{v}_i^\top \hat{t}_j = \frac{f_I(I_i) \cdot f_T(T_j)}{\|f_I(I_i)\| \|f_T(T_j)\|}. \tag{2}$$

To optimize cross-modal alignment, we adopt a dual-directional InfoNCE-style contrastive loss. The overall DMA loss is formulated as:

$$\mathcal{L}_{DMA} = -\frac{1}{2N} \sum_{i=1}^{N} \left[ \log \frac{\exp(s_{i,i}/\tau)}{\sum_{j=1}^{N} \exp(s_{i,j}/\tau)} + \log \frac{\exp(s_{i,i}/\tau)}{\sum_{j=1}^{N} \exp(s_{j,i}/\tau)} \right] \tag{3}$$

where $\tau > 0$ is a learnable temperature parameter that controls the concentration level of the distribution. The first term promotes alignment between image $I_i$ and its paired text $T_i$ over all other texts in the batch (image-to-text), while the second term promotes alignment from text to image (text-to-image).

Minimizing $\mathcal{L}_{DMA}$ encourages the model to bring matched image–text pairs closer and push mismatched pairs apart in the shared embedding space, enabling robust and generalizable representation learning for downstream medical tasks.

**Prediction Stage: Knowledge Decomposition Guided Prediction** We propose **LUMIN**, a longitudinal multi-modal framework that predicts breast cancer recurrence by integrating mammographic and clinical information across time. This model leverages pre-trained encoders to extract representations from mammograms and EHRs, and disentangles modality- and time-specific factors via a hierarchical

knowledge decomposition process. Inspired by recent advances in interpretable representation learning (Zhou, Zhou, and Chen 2024), we introduce two complementary modules: the **Cross-Modal Disentangled Knowledge Extractor (CM-DKE)** and the **Temporal Evolution Disentangled Knowledge Extractor (TE-DKE)**, which collaboratively decompose, align, and integrate knowledge across modalities and time points.

**Cross-Modal Disentangled Knowledge Extractor (CM-DKE)** The CM-DKE module disentangles multi-modal knowledge at each time point $t$ into three disentangled components:

**(1) Modality-Specific Knowledge.** For imaging $I^{(t)}$ and textual $T^{(t)}$ inputs, we apply independent transformation networks to extract intra-modal features:

$$K_{spec}^{(t)} = \left[\phi_I(W_I I^{(t)} + b_I),\ \phi_T(W_T T^{(t)} + b_T)\right], \quad (4)$$

where $\phi_I$ and $\phi_T$ are non-linear projection layers specific to each modality.

**(2) Shared Cross-Modal Knowledge.** To model joint semantics, we concatenate $I^{(t)}$ and $T^{(t)}$ and map them into a common latent space:

$$K_{shared}^{(t)} = \phi_{shared}\left(W_{shared}[I^{(t)}; T^{(t)}] + b_{shared}\right). \quad (5)$$

**(3) Complementary Cross-Modal Knowledge.** We compute cross-attentive interactions between modalities via bilinear alignment. Specifically, we first obtain aligned representations:

$$A^{(t)} = I_{align}^{(t)} \otimes T_{align}^{(t)}, \quad (6)$$

where $\otimes$ denotes the outer product. The complementary knowledge is then computed as:

$$K_{comp}^{(t)} = I_{align}^{(t)} \odot \sigma(W_I^* A^{(t)}) + T_{align}^{(t)} \odot \sigma(W_T^* A^{(t)\top}), \quad (7)$$

where $\sigma$ is the sigmoid activation and $\odot$ denotes element-wise product. This operation enables each modality to selectively attend to complementary features from the other.

The full cross-modal representation at time $t$ is the concatenation:

$$K_{mod}^{(t)} = [K_{spec}^{(t)}; K_{shared}^{(t)}; K_{comp}^{(t)}]. \quad (8)$$

**Temporal Evolution Disentangled Knowledge Extractor (TE-DKE)** While CM-DKE captures within-timepoint multimodal relationships, TE-DKE focuses on disentangling knowledge across time. Given representations at two timepoints $t_1$ and $t_2$, we model stable and evolving components as:

$$K_{temp} = \phi_{temp}\left(W_{temp}\left[K_{mod}^{(t_1)}; K_{mod}^{(t_2)}; K_{mod}^{(t_1)} - K_{mod}^{(t_2)}\right] + b_{temp}\right) \quad (9)$$

Here, the additive and subtractive operations allow the model to explicitly encode invariant and varying aspects of the patient's condition over time. This separation enhances temporal reasoning and helps mitigate information dilution from simple fusion strategies.

**Aggregated Knowledge Integration and Prediction** We concatenate the knowledge components from both CM-DKE and TE-DKE into a unified sequence representation:

$$K_{agg} = [K_{mod}^{(t_1)}; K_{mod}^{(t_2)}; K_{temp}], \quad (10)$$

This representation is then fed into a Transformer-based module to model high-order dependencies and generate the recurrence prediction. The Transformer enables global reasoning across both modalities and timepoints, making the final decision interpretable and temporally aware.

# Experiment Setting

## Data Collection

In this study, we curated a temporally-aligned multi-modal dataset from the Netherlands Cancer Institute (IRBd21-059), including 3,924 breast cancer patients, 19,684 longitudinal mammograms, and corresponding electronic health records. To our knowledge, this is the first large-scale dataset aligning multi-timepoint mammograms with clinical text. All patients had confirmed breast cancer and received standard-of-care treatments based on tumor subtype and stage.

All patients had at least one year of post-treatment follow-up, with a consistent schedule of screening mammography. Each follow-up exam included the standard four views—bilateral craniocaudal (CC) and mediolateral oblique (MLO)—ensuring uniform spatial coverage across time. In total, 11.7% of patients developed breast cancer recurrence during follow-up, categorized as local, regional, or distant (metastatic) relapse, while the remaining 88.3% remained recurrence-free throughout the observation period.

To ensure consistent temporal modeling across patients, we selected the two most recent follow-up mammographic exams prior to recurrence (or censoring, for non-recurrent cases). This design enables the model to capture the latest signs of disease progression while maintaining uniform temporal input. In cases with more than two follow-up exams, only the last two satisfying the interval criterion were retained. Importantly, we required a minimum time gap of six months between the selected exams to ensure clinically meaningful temporal variation. This approach balances modeling consistency and the ability to learn disease evolution from real-world longitudinal data.

## Experiment Setting

We conducted our experiments in a two-stage framework comprising a cross-modal pre-training stage and a recurrence prediction stage. The dataset was split at the patient level to prevent information leakage, with 70% allocated to the pre-training phase and the remaining 30% reserved exclusively for downstream prediction and evaluation.

For both stages, we adopted ResNet-50 (Koonce and Koonce 2021) as the visual backbone and RadioBERT (Zhang et al. 2023b) as the textual encoder, given their proven effectiveness in capturing high-level semantic features in medical imaging and clinical narratives, respectively. During pre-training, we utilized all available mammographic exams and corresponding EHR entries prior to recurrence events (or end of follow-up for non-recurrent cases)

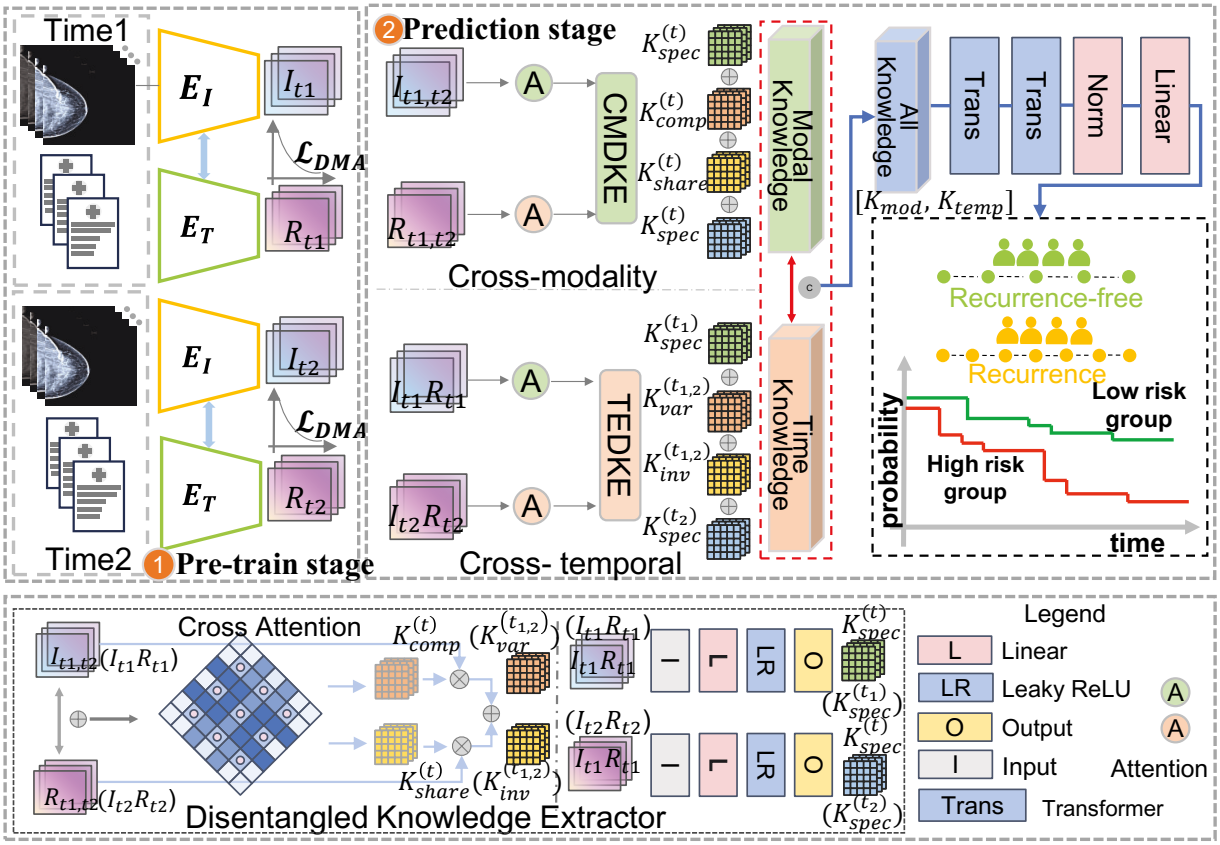


Figure 2: Overview of LUMIN, a predictive framework integrating longitudinal mammograms and EHRs for breast cancer recurrence prediction via modality- and time-aware knowledge extraction.

to learn robust cross-modal representations. Input mammograms were rescaled to $512 \times 512$ pixels, and all text inputs were truncated or padded to a maximum of 256 tokens. The model was trained using Adam optimizer for 20 epochs with a batch size of 24. The learning rates for the image and text encoders were set to $1 \times 10^{-4}$ and $1 \times 10^{-6}$, respectively, following standard warm-up and cosine decay schedules.

For the downstream recurrence prediction stage, the pre-training set was further split into 80% training and 20% validation. The held-out 30% from the initial dataset split was used as an independent test set. During fine-tuning, learning rates were reduced to 10% of their pre-training values to retain previously learned representations while adapting to the target task.

## Evaluation Metrics

To evaluate the effectiveness of our model from both predictive and clinical perspectives, we employ a comprehensive set of metrics. For classification performance, we report accuracy (ACC), area under the receiver operating characteristic curve (AUC), sensitivity, and specificity. All classification metrics are computed on the held-out test set, and 95% confidence intervals are estimated using 1,000 bootstrap samples for statistical robustness. To assess statistical significance of AUC differences between models, we performed pairwise DeLong's test with $p < 0.05$ considered significant.

Beyond binary classification, we assess the model's clinical utility via risk stratification analysis. Specifically, predicted recurrence scores are used in a Cox proportional hazards model to compute hazard ratios (HRs) between high- and low-risk groups. We report HR values alongside their 95% confidence intervals and log-rank test p-values to evaluate statistical significance. This survival-based analysis validates whether the model's risk scores correspond to meaningful differences in recurrence-free survival, which is critical for personalized follow-up planning.

To further support model interpretability, we apply Grad-CAM-based visualization on the image encoder and attention weight analysis on the EHR encoder. These methods highlight salient regions and clinical terms that contribute most to the recurrence prediction, offering insights into model decision-making. This interpretability analysis enables qualitative validation and promotes clinical trust in the proposed framework.

## Results and Discussion

**Performance Comparison Across Modalities.** We first establish unimodal baselines using either mammographic images or EHR data. As shown in Table 1, the image-only

| Model | Module | Modality | ACC | AUC | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| Resnet-50 | *NA* | Image | 0.699 (0.673-0.726) | 0.572 (0.519-0.629) | 0.477 (0.400-0.557) | 0.703 (0.691-0.757) |
| ST-I | *NA* | Image | 0.611 (0.585-0.640) | 0.605 (0.551-0.656) | 0.558 (0.472-0.644) | 0.619 (0.590-0.650) |
| ST-R | *NA* | EHR | 0.621 (0.593-0.653) | 0.686 (0.649-0.739) | 0.693 (0.612-0.767) | 0.606 (0.576-0.637) |
| ST-B | *NA* | I&R | 0.700 (0.675-0.728) | 0.693 (0.642-0.740) | 0.623 (0.539-0.703) | 0.711 (0.683-0.741) |
| L-B | *NA* | I&R | 0.644 (0.617-0.672) | 0.676 (0.630-0.723) | 0.686 (0.609-0.761) | 0.638 (0.608-0.668) |
| LUMIN-M | CM-DKE | I&R | 0.672 (0.644-0.700) | 0.721 (0.680-0.764) | **0.702*** **(0.630-0.777)** | 0.668 (0.638-0.697) |
| LUMIN-T | TE-DKE | I&R | 0.682 (0.655-0.709) | 0.714 (0.669-0.759) | 0.680 (0.609-0.755) | 0.682 (0.655-0.711) |
| LUMIN | TE-DKE CM-DKE | I&R | **0.718*** **(0.692-0.744)** | **0.729*** **(0.685-0.771)** | 0.623 (0.544-0.700) | **0.732*** **(0.703-0.758)** |

Table 1: Performance comparison of different models for breast cancer recurrence prediction. Note: ST-I: Single Time Image; ST-R: Single Time EHR; ST-B: Single Time Baseline; L-B: Longitudinal Baseline; I: Image; R: EHR

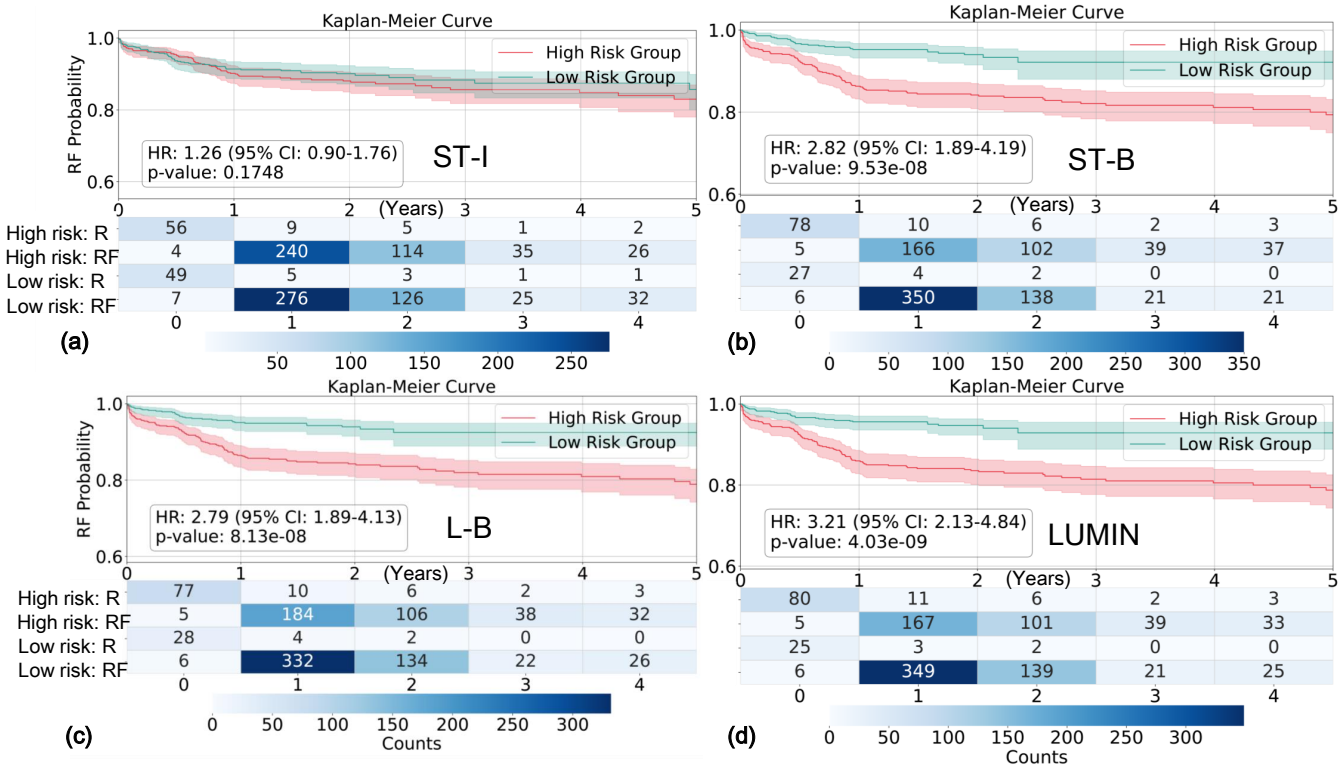


Figure 3: Kaplan-Meier curves showing recurrence-free probability over time and heatmaps showing the distribution of recurrence and recurrence-free cases across risk groups for different models (a) ST-I (b) ST-B (c) L-B (d) LUMIN

.

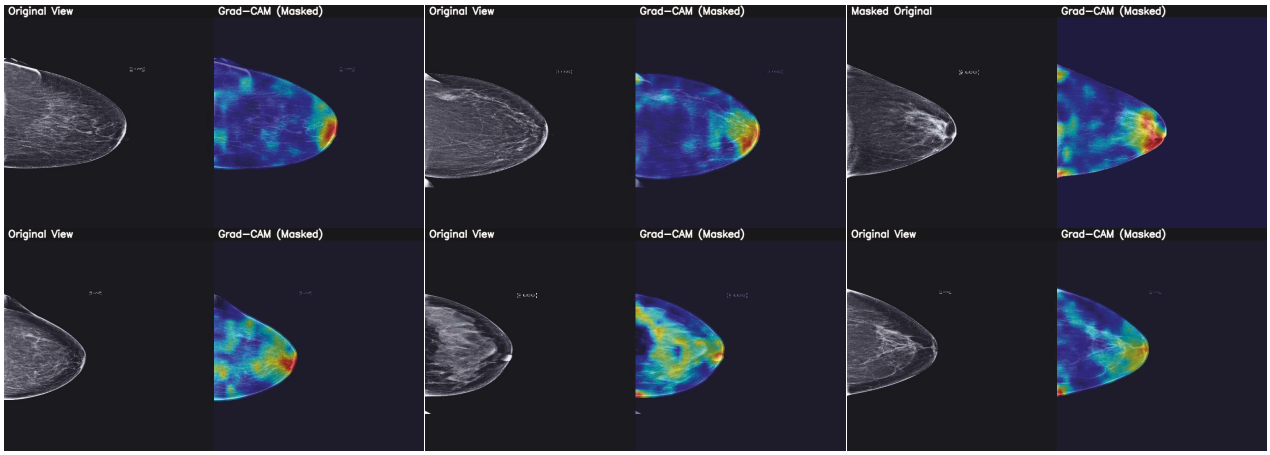


Figure 4: Interpretability visualization using Grad-CAM on left craniocaudal (LCC) mammographic views. Each case presents the original mammogram and the corresponding Grad-CAM heatmap highlighting the model's attention regions. The visualizations are generated using the LUMIN model and demonstrate its spatial focus when predicting future recurrence risk.All examples are taken from LCC views across different patients and timepoints.

ResNet-50 model achieves an AUC of 0.572, while its temporal variant (ST-I) slightly improves to 0.605. In contrast, the EHR-only model (ST-R) yields a significantly higher AUC of 0.686 ($p < 0.001$, DeLong's test), suggesting that structured clinical data contain stronger recurrence-related signals than imaging alone.

**Impact of Temporal and Cross-modal Information.** We then assess models that incorporate both multimodal and longitudinal information. Naïve fusion of image and text from a single timepoint (ST-B) achieves moderate performance (AUC = 0.693), while its longitudinal extension (L-B), which directly concatenates information from two exams without temporal modeling, performs slightly worse (AUC = 0.676). These results highlight the limitations of unstructured fusion in capturing disease progression.

In contrast, our proposed LUMIN framework introduces explicit knowledge decomposition across modalities and time, achieving the highest performance with an AUC of 0.729 and accuracy of 0.718. This demonstrates the benefit of disentangled representation learning in modeling longitudinal multimodal data. The AUC gains of LUMIN over ST-I ($p < 0.001$), ST-R ($p = 0.018$), ST-B ($p = 0.004$), and L-B ($p = 0.007$) are all statistically significant, confirming its superior predictive performance.

**Risk Stratification and Survival Analysis.** We further evaluate the model's ability to stratify patients by recurrence risk. Using the median predicted risk score from the validation set, we divide the test cohort into high- and low-risk groups and compute hazard ratios (HRs), as shown in Table 2. Text-only (ST-R) achieves an HR of 2.82, compared to 1.26 for image-only (ST-I). Naïve multimodal models (ST-B and L-B) yield moderate HRs, while LUMIN-M (HR = 3.06) and LUMIN-T (HR = 2.93) offer substantial improvements. LUMIN attains the highest HR (3.21), confirming its superior risk stratification capabilities.

Fig. 3(a–d) presents Kaplan-Meier curves and recurrence heatmaps for different models. LUMIN (Fig. 2(d)) achieves the clearest separation between high- and low-risk groups. Within five years post-treatment, the recurrence rate in the high-risk group is 3.4 times that of the low-risk group (102 vs. 30), while the low-risk group contains 1.5 times more recurrence-free cases (540 vs. 345). These findings demonstrate LUMIN's ability to prioritize high-risk patients while minimizing false alarms, making it a promising tool for personalized post-treatment surveillance.

| Model | HR (95% CI) | P-value |
|---|---|---|
| ResNet-50 | 1.07 (0.67–1.33) | 0.81 |
| ST-I | 1.26 (0.90–1.76) | 0.17 |
| ST-R | 2.82 (1.89–4.19) | $1.24 \times 10^{-5}$ |
| ST-B | 2.92 (1.95–4.29) | $9.53 \times 10^{-8}$ |
| L-B | 2.79 (1.89–4.13) | $8.13 \times 10^{-8}$ |
| LUMIN-M | 3.06 (2.04–4.60) | $1.29 \times 10^{-8}$ |
| LUMIN-T | 2.93 (1.96–4.37) | $3.96 \times 10^{-8}$ |
| **LUMIN** | **3.21 (2.13–4.84)** | $\mathbf{4.03 \times 10^{-9}}$ |

Table 2: Risk stratification performance of different models based on HR.

**Ablation and Component Analysis** To evaluate the effectiveness of each component in the proposed LUMIN framework, we performed a structured ablation study, including both high-level module removal and fine-grained submodule analysis. The full LUMIN model, equipped with contrastive pretraining, CM-DKE, and TE-DKE, achieves the highest AUC of 0.729 (Table 3).

We first conducted module-level ablations by independently removing each component from the full model. Disabling pretraining (*w/o Pretraining*) leads to a marked AUC

| Model Variant | Pretrain | CM-DKE | TE-DKE | AUC | Baseline |
|---|---|---|---|---|---|
| Full LUMIN | ✓ | ✓ | ✓ | **0.729** | ✓ |
| w/o Pretraining | ✗ | ✓ | ✓ | 0.703 | ✗ |
| w/o CM-DKE | ✓ | ✗ | ✓ | 0.714 | ✗ |
| w/o TE-DKE | ✓ | ✓ | ✗ | 0.721 | ✗ |

Table 3: Ablation study of LUMIN backbone and major modules. ✓: enabled, ✗: removed. Bold: highest AUC

| Model Variant | Pretrain | CM-DKE | TE-DKE | AUC | Baseline |
|---|---|---|---|---|---|
| *Sub-module ablations (within CM-DKE)* | | | | | |
| CM-DKE only | ✓ | ✓ | ✗ | **0.721** | ✓ |
| w/o $K_{comp}$ | ✓ | * | ✗ | 0.711 | ✗ |
| w/o $K_{img}$ | ✓ | * | ✗ | 0.706 | ✗ |
| w/o $K_{text}$ | ✓ | * | ✗ | 0.691 | ✗ |
| w/o $K_{shared}$ | ✓ | * | ✗ | 0.708 | ✗ |
| *Sub-module ablations (within TE-DKE)* | | | | | |
| TE-DKE only | ✓ | ✗ | ✓ | **0.714** | ✓ |
| w/o $K_{var}$ | ✓ | ✗ | * | 0.701 | ✗ |
| w/o $K_{inv}$ | ✓ | ✗ | * | 0.704 | ✗ |
| w/o $K_{T1}$ | ✓ | ✗ | * | 0.695 | ✗ |
| w/o $K_{T2}$ | ✓ | ✗ | * | 0.699 | ✗ |

Table 4: Ablation study of sub-components within CM-DKE and TE-DKE. ✓: enabled, ✗: removed, *: partially removed.

decrease to 0.703, highlighting the role of semantic initialization for cross-modal fusion. Removing CM-DKE (*w/o CM-DKE*) and TE-DKE (*w/o TE-DKE*) also results in performance drops (AUC = 0.721 and 0.714, respectively), validating the importance of disentangling modality-specific and temporal knowledge.

Next, we examined the internal mechanisms of CM-DKE and TE-DKE in isolation by retaining only one module at a time as baseline, then progressively removing its subcomponents (Table 4). Using CM-DKE alone (with pretraining, but without TE-DKE) yields AUC = 0.721. Within this setup, removing text-specific knowledge ($K_{text}$) causes the largest drop (AUC = 0.691), followed by image-specific ($K_{img}$), complementary ($K_{comp}$), and shared knowledge ($K_{shared}$), indicating that text-derived features play a dominant role in cross-modal fusion.

Similarly, TE-DKE alone (without CM-DKE) produces AUC = 0.714. Ablating invariant ($K_{inv}$), timepoint-specific ($K_{T1}$, $K_{T2}$), or variation-aware ($K_{var}$) knowledge all degrade performance (to 0.699–0.704), suggesting that temporal progression and inter-timepoint contrasts are crucial for longitudinal modeling.

Overall, these results demonstrate that each module and its associated knowledge branches contribute incrementally to performance, and the removal of any component consistently results in performance degradation, confirming the necessity of the proposed disentangled architecture.

**Model Interpretability via Grad-CAM.** To gain spatial insights into model decisions, we visualize Grad-CAM heatmaps generated from the image encoder on left craniocaudal (LCC) mammograms (Fig. 4). For each case, we present the original image, the masked image (if applicable), and the corresponding Grad-CAM heatmap. The highlighted regions consistently focus on fibroglandular tissue with architectural distortion, asymmetry, or suspicious density patterns—areas of clinical relevance. Importantly, attention remains consistent across follow-up exams, suggesting that the model not only localizes disease-related features but also aligns them temporally. These visualizations serve as an external validation of the model's spatial reasoning and enhance transparency for clinical deployment.

**Representation Analysis via t-SNE.** To assess the learned representations, we visualize the embedding spaces from CM-DKE and TE-DKE using t-SNE (Fig. 5). CM-DKE shows distinct clusters of image-specific, text-specific, shared, and complementary features, reflecting clear modality disentanglement. TE-DKE embeddings exhibit smooth transitions among time-invariant, time-varying, and time-specific components, indicating effective modeling of temporal progression. The distinct separability of learned features supports the interpretability and structural fidelity of the LUMIN framework.

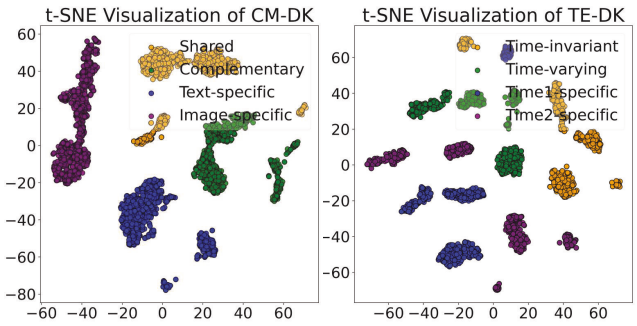


Figure 5: t-SNE visualization of disentangled feature representations. (a) CM-DKE: The feature space is decomposed into shared (orange), complementary (purple), text-specific (green), and image-specific (blue) components. (b) TE-DKE: Temporal features are disentangled into time-invariant (orange), time-varying (green), and time-specific components for each timepoint.

## Conclusion

In this study, we propose LUMIN, a Longitudinal Multimodal Knowledge Decomposition Network for breast cancer recurrence prediction, which integrates longitudinal mammograms and EHRs. LUMIN captures cross-modal interactions and models disease progression through two modules: CM-DKE, which disentangles modality-specific, shared, and complementary knowledge, and TE-DKE, which distinguishes stable, evolving, and time-specific information. Built upon a contrastive pretraining framework, LUMIN effectively aligns multi-modal representations and leverages longitudinal context to improve predictive performance. Extensive experiments on a large-scale clinical dataset demonstrate that LUMIN consistently outperforms unimodal and non-disentangled baselines in both AUC and risk stratification. These results underscore the importance of structured knowledge decomposition for modeling disease progression and highlight the potential of LUMIN to support personalized post-treatment surveillance in breast cancer care.

## Acknowledgments

Tao Tan acknowledges support from Shenzhen Medical Research Fund (D2501013). Chunyao Lu holds a Chinese Scholarship Council Studentship (No.202206830036). Tianyu Zhang acknowledges support from the Guangzhou Elite Program (TZ-JY201948). Tianyu Zhang and Ritse Mann also acknowledge funding from the ODELIA project (European Union's Horizon Europe research and innovation programme, grant agreement No. 101057091) and the SAFE-MRI project (Horizon Europe, grant agreement No. 101087701).

## References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Chlebowski, R. T.; Aragaki, A. K.; and Pan, K. 2021. Breast cancer prevention: time for change. *JCO oncology practice*, 17(12): 709–716.

Colleoni, M.; Sun, Z.; Price, K. N.; Karlsson, P.; Forbes, J. F.; Thürlimann, B.; Gianni, L.; Castiglione, M.; Gelber, R. D.; Coates, A. S.; et al. 2016. Annual hazard rates of recurrence for breast cancer during 24 years of follow-up: results from the international breast cancer study group trials I to V. *Journal of clinical oncology*, 34(9): 927–935.

Dadsetan, S.; Arefan, D.; Berg, W. A.; Zuley, M. L.; Sumkin, J. H.; and Wu, S. 2022. Deep learning of longitudinal mammogram examinations for breast cancer risk prediction. *Pattern recognition*, 132: 108919.

Damiani, C.; Kalliatakis, G.; Sreenivas, M.; Al-Attar, M.; Rose, J.; Pudney, C.; Lane, E. F.; Cuzick, J.; Montana, G.; and Brentnall, A. R. 2023. Evaluation of an AI model to assess future breast cancer risk. *Radiology*, 307(5): e222679.

Duffy, S. W.; Tabár, L.; Yen, A. M.-F.; Dean, P. B.; Smith, R. A.; Jonsson, H.; Törnberg, S.; Chiu, S. Y.-H.; Chen, S. L.-S.; Jen, G. H.-H.; et al. 2021. Beneficial effect of consecutive screening mammography examinations on mortality from breast cancer: a prospective study. *Radiology*, 299(3): 541–547.

El Haji, H.; Souadka, A.; Patel, B. N.; Sbihi, N.; Ramasamy, G.; Patel, B. K.; Ghogho, M.; and Banerjee, I. 2023. Evolution of breast cancer recurrence risk prediction: a systematic review of statistical and machine learning–based models. *JCO clinical cancer informatics*, 7: e2300049.

Gao, Y.; Tan, T.; Wang, X.; Beets-Tan, R.; Zhang, T.; Han, L.; Portaluri, A.; Lu, C.; Liang, X.; Teuwen, J.; et al. 2025. Multi-modal Longitudinal Representation Learning for Predicting Neoadjuvant Therapy Response in Breast Cancer Treatment. *IEEE Journal of Biomedical and Health Informatics*.

Howard, F. M.; Dolezal, J.; Kochanny, S.; Khramtsova, G.; Vickery, J.; Srisuwananukorn, A.; Woodard, A.; Chen, N.; Nanda, R.; Perou, C. M.; et al. 2023. Integration of clinical features and deep learning on pathology for the prediction of breast cancer recurrence assays and risk of recurrence. *NPJ Breast Cancer*, 9(1): 25.

Karaman, B. K.; Dodelzon, K.; Akar, G. B.; and Sabuncu, M. R. 2024. Longitudinal Mammogram Risk Prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 437–446. Springer.

Koonce, B.; and Koonce, B. 2021. ResNet 50. *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, 63–72.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Lou, S.-J.; Hou, M.-F.; Chang, H.-T.; Chiu, C.-C.; Lee, H.-H.; Yeh, S.-C. J.; and Shi, H.-Y. 2020. Machine learning algorithms to predict recurrence within 10 years after breast cancer surgery: a prospective cohort study. *Cancers*, 12(12): 3817.

Sprague, B. L.; Coley, R. Y.; Lowry, K. P.; Kerlikowske, K.; Henderson, L. M.; Su, Y.-R.; Lee, C. I.; Onega, T.; Bowles, E. J.; Herschorn, S. D.; et al. 2023. Digital breast tomosynthesis versus digital mammography screening performance on successive screening rounds from the Breast Cancer Surveillance Consortium. *Radiology*, 307(5): e223142.

Steyaert, S.; Pizurica, M.; Nagaraj, D.; Khandelwal, P.; Hernandez-Boussard, T.; Gentles, A. J.; and Gevaert, O. 2023. Multimodal data fusion for cancer biomarker discovery with deep learning. *Nature machine intelligence*, 5(4): 351–362.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, F.; Shu, X.; Meszoely, I.; Pal, T.; Mayer, I. A.; Yu, Z.; Zheng, W.; Bailey, C. E.; and Shu, X.-O. 2019. Overall mortality after diagnosis of breast cancer in men vs women. *JAMA oncology*, 5(11): 1589–1596.

Wang, X.; Tan, T.; Gao, Y.; Marcus, E.; Han, L.; Portaluri, A.; Zhang, T.; Lu, C.; Liang, X.; Beets-Tan, R.; et al. 2024. Ordinal Learning: Longitudinal Attention Alignment Model for Predicting Time to Future Breast Cancer Events from Mammograms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 155–165. Springer.

Wang, X.; Tan, T.; Gao, Y.; Su, R.; Zhang, T.; Han, L.; Teuwen, J.; D'Angelo, A.; Drukker, C. A.; Schmidt, M. K.; et al. 2023. Predicting up to 10 year breast cancer risk using longitudinal mammographic screening history. *medRxiv*, 2023–06.

Yang, J.; Ju, J.; Guo, L.; Ji, B.; Shi, S.; Yang, Z.; Gao, S.; Yuan, X.; Tian, G.; Liang, Y.; et al. 2022. Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. *Computational and structural biotechnology journal*, 20: 333–342.

Yeoh, H. H.; Liew, A.; Phan, R.; Strand, F.; Rahmat, K.; Nguyen, T. L.; Hopper, J. L.; and Tan, M. 2023. RADIFUSION: A multi-radiomics deep learning based breast cancer risk prediction model using sequential mammographic images with image attention and bilateral asymmetry refinement. *arXiv preprint arXiv:2304.00257*.

Zardavas, D.; Irrthum, A.; Swanton, C.; and Piccart, M. 2015. Clinical management of breast cancer heterogeneity. *Nature reviews Clinical oncology*, 12(7): 381–394.

Zhang, T.; Tan, T.; Samperna, R.; Li, Z.; Gao, Y.; Wang, X.; Han, L.; Yu, Q.; Beets-Tan, R. G.; and Mann, R. M. 2023a. Radiomics and artificial intelligence in breast imaging: a survey. *Artificial Intelligence Review*, 56(Suppl 1): 857–892.

Zhang, T.; Tan, T.; Wang, X.; Gao, Y.; Han, L.; Balkenende, L.; D'Angelo, A.; Bao, L.; Horlings, H. M.; Teuwen, J.; et al. 2023b. RadioLOGIC, a healthcare model for processing electronic health records and decision-making in breast disease. *Cell Reports Medicine*, 4(8).

Zhou, H.; Zhou, F.; and Chen, H. 2024. Cohort-individual cooperative learning for multimodal cancer survival analysis. *IEEE Transactions on Medical Imaging*.