Δt-Mamba3D: A Time-Aware Spatio-Temporal State-Space Model for Breast Cancer Risk Prediction

Zhengbo Zhou[1], **Dooman Arefan**[2], **Margarita Louise Zuley**[2], **Shandong Wu**[1,2,3]

[1]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA
[2]Department of Radiology, University of Pittsburgh, Pittsburgh, PA, USA
[3]Department of Biomedical Informatics and Department of Bioengineering, University of Pittsburgh, Pittsburgh, PA, USA
wus3@upmc.edu

Abstract

Longitudinal analysis of sequential radiological images is hampered by a fundamental data challenge: how to effectively model a sequence of high-resolution images captured at irregular time intervals. This data structure contains indispensable spatial and temporal cues that current methods fail to fully exploit. Models often compromise by either collapsing spatial information into vectors or applying spatio-temporal models that are computationally inefficient and incompatible with non-uniform time steps. We address this challenge with Time-Aware Δt-Mamba3D, a novel state-space architecture adapted for longitudinal medical imaging. Our model simultaneously encodes irregular inter-visit intervals and rich spatio-temporal context while remaining computationally efficient. Its core innovation is a continuous-time selective scanning mechanism that explicitly integrates the true time difference between exams into its state transitions. This is complemented by a multi-scale 3D neighborhood fusion module that robustly captures spatio-temporal relationships. In a comprehensive breast cancer risk prediction benchmark using sequential screening mammogram exams, our model shows superior performance, improving the validation C-index by 2–5 percentage points and achieving higher 1–5 year AUC scores compared to established variants of recurrent, transformer, and state-space models. Thanks to its linear complexity, the model can efficiently process long and complex patient screening histories of mammograms, forming a new framework for longitudinal image analysis.

Code — https://github.com/zhouzhengbo2022/dtMamba3D

Introduction

Screening mammography for breast cancer detection is inherently longitudinal. Women return every a few years, image acquisition protocols evolve over time, breasts change with aging, and subtle preclinical lesions may emerge gradually across exams. Radiologists routinely examine longitudinal and cross-view information: they compare current and prior exams and craniocaudal (CC) and mediolateral oblique (MLO) views, assess side-to-side asymmetries, and evaluate interval change when estimating malignancy risk or assigning BI-RADS diagnostic categories (Scutt, Lancaster, and Manning 2006). Yet most deep learning systems for breast imaging still operate on a single imaging exam, ignoring the temporal context that drives clinical decision making (Yala et al. 2021). When multiple exams are available, most methods first collapse each exam into a single per-visit feature vector and then apply a temporal model (e.g., RNNs or GRUs), thereby sacrificing fine-grained lesion morphology and growth patterns (Dadsetan et al. 2022). In addition, the irregular time gaps between exams, an important predictor of breast cancer risk, are usually left unencoded by existing deep-learning approaches (Zhou et al. 2025).

Irregularly timed data frequently occurs in clinical settings, reflecting varying degrees of disease severity since patients with severe conditions tend to have more frequent hospital visits. Despite this, many existing methods, such as standard Recurrent Neural Networks (RNNs) and Transformers, treat patient visits as tokens placed on an evenly spaced temporal grid (Karaman et al. 2024), which discard valuable interval information. Over the past decade, several specialized models have been proposed to address irregular sampling in clinical time series. Time-discretized models, such as GRU-D (Che et al. 2018) and Time-aware LSTM (Nguyen et al. 2020), incorporate elapsed time or its exponential decay directly into hidden state updates. Although these methods capture interval magnitude efficiently, they inherently assume piecewise constant dynamics, limiting their ability to model evolving risk between observations. Continuous-time approaches, including Neural ODEs and Neural CDEs (Rubanova, Chen, and Duvenaud 2019; Kidger et al. 2020), and recent advancements like ContiFormer (Chen et al. 2023), naturally handle irregular intervals and offer continuous-time predictions. However, they have primarily been evaluated on minute-level ECG or sensor data or low-dimensional EHR records. In contrast, screening intervals in medical imaging often span from 0.5 to 3 years, involving extremely high-dimensional features. Such imaging visits are inherently sparse, with each patient encounter represented as a discrete event at a specific integer timestamp, accompanied by a zero-valued signal in between, a scenario inadequately modeled by previous methods. As highlighted by recent advances such as Mamba (Gu and Dao 2023), state-space models (SSMs) with adaptive, context-aware parameters offer enhanced capability for capturing long-range dependencies in dynamic systems. Although Mamba has significantly surpassed conventional re-

current models in domains such as language modeling, it has not explicitly encoded irregular time intervals. Consequently, adapting models like Mamba to effectively handle irregularly timed, high-dimensional imaging data remains largely unexplored and presents an urgent need in the field.

Capturing joint spatial–temporal patterns in longitudinal medical imaging data poses significant methodological and computational challenges. Early 3D CNNs, such as C3D (Tran et al. 2015), I3D (Carreira and Zisserman 2017), apply cubic convolutions to densely sampled frame stacks; their computation scales cubically with spatial resolution and their receptive field remains inherently limited. Video vision transformers, exemplified by TimeSformer (Bertasius, Wang, and Torresani 2021), ViViT (Arnab et al. 2021), and Video Swin (Liu et al. 2022), simultaneously encode spatial structure and densely, uniformly sampled temporal dynamics, yielding joint spatio-temporal representations. Full self-attention scales quadratically with the number of spatio-temporal tokens and becomes impractical when modeling longer longitudinal visits, a limitation that is magnified by the typically small sample sizes in clinical datasets. Furthermore, standard video vision transformers assume equal temporal spacing and do not reflect real-world inter-visit intervals. More recently, structured state-space sequence models (SSMs), S4 and Mamba (Gu, Goel, and Ré 2021; Gu and Dao 2023) and vision variants that devise 2D/3D scan orders or local fusion windows (Zhu et al. 2024; Liu et al. 2024; Xiao et al. 2024), achieve linear-time scanning, but they likewise assume uniformly spaced tokens and must stack multiple costly passes to absorb full 3D context.

To bridge these gaps, we propose Time-Aware Δt-Mamba3D, a spatio-temporal state-space block with two defining features: i) Δt-aware transitions: We find that driving the SSM solely with raw inputs as control signals is ill-suited to irregularly sampled clinical data, leaving the model largely insensitive to the true time gaps. Instead, we modulate every selective-scan update with the true inter-visit interval Δt, enabling continuous-time memory decay or accumulation under irregular sampling while preserving the original content-aware step size; ii) Multi-scale depth-wise 3D fusion: neighborhood-adaptive convolutions jointly encode spatial and temporal context at low cost. Our pipeline first converts each exam into a token sequence via a unidirectional sweep (Fig. 1). The token states then evolve through the closed-form SSM transition whose step size is modulated by the true inter-visit gap, thereby embedding irregular temporal information. These states are then refined by structure-aware 3D convolutions that re-weight neighbouring voxels across multiple receptive fields, and an observation head maps the fused state to output variables. This novel design preserves lesion morphology, accommodates irregular timing, and scales linearly in memory with sequence length. We embed Structure-Aware Δt-Mamba 3D in a prediction pipeline that ingests up to eight prior screening mammogram exams (four views including CC and MLO views of left and right breasts per exam) and outputs year-specific hazards. On two mammogram datasets with irregular inter-exam intervals, our model outperforms time-aware models and spatio-temporal models, improving C-index and 1–5-year AUCs while maintaining linear memory growth. Our contributions are as follows:

1. We extend Mamba with a selective scanning mechanism whose state transition explicitly incorporates the true inter-visit interval Δt at the image level. By utilizing irregular time spans, combined with inputs as control signals for SSM to achieve superior model performance.
2. We embed a multiscale, depth-wise 3D convolution block within the Mamba module to efficiently capture joint spatio-temporal context.
3. On two longitudinal mammography datasets with varying temporal patterns and different class distributions, our model surpasses recurrent, transformer, and visual-SSM baselines in accuracy while maintaining linear memory and computing scaling.

Related Work

Time-Aware Models

Early models incorporated elapsed time between observations by explicitly modulating recurrent network updates. T-LSTM (Nguyen et al. 2020) and GRU-D (Che et al. 2018) adopted data-driven exponential decay mechanisms to inputs and hidden states, effectively reducing the influence of outdated measurements. Attention-based methods have also integrated temporal information through various strategies. Approaches like time2vec (Kazemi et al. 2019) introduced learnable temporal embeddings, while continuous-time attention models explicitly factor in the elapsed time as positional biases or embedding components (Shukla and Marlin 2021). ContiFormer (Chen et al. 2023), a recent advancement, further enhances this line by leveraging continuous-time self-attention mechanisms specifically designed to handle irregularly sampled time series. Beyond these parametric decay and attention-based strategies, continuous-time latent dynamics have been extensively modeled by methods such as Latent ODEs (Rubanova, Chen, and Duvenaud 2019) and Neural Controlled Differential Equations (Neural CDEs) (Kidger et al. 2020). These approaches explicitly integrate hidden state trajectories between irregular event occurrences, effectively capturing complex continuous-time dependencies within data.

Vision State Space Models

Recent visual SSMs tackle spatial coherence by crafting tailored scan patterns (Liu et al. 2024; Zhu et al. 2024). Spatial-Mamba (Xiao et al. 2024) tiles 2D patches and applies a local fusion window to mitigate scan-order bias. For 3D volumes and long-horizon video, Seg-Mamba (Xing et al. 2024) and LongMamba (Zhou et al. 2025) tokenize slices or frames into patch sequences with bespoke spatial–temporal scan orders, ensuring that both intra-slice structure and inter-frame dynamics are captured.

Methods

Preliminary

SSMs are commonly used for analyzing sequential data and modeling continuous linear time-invariant systems

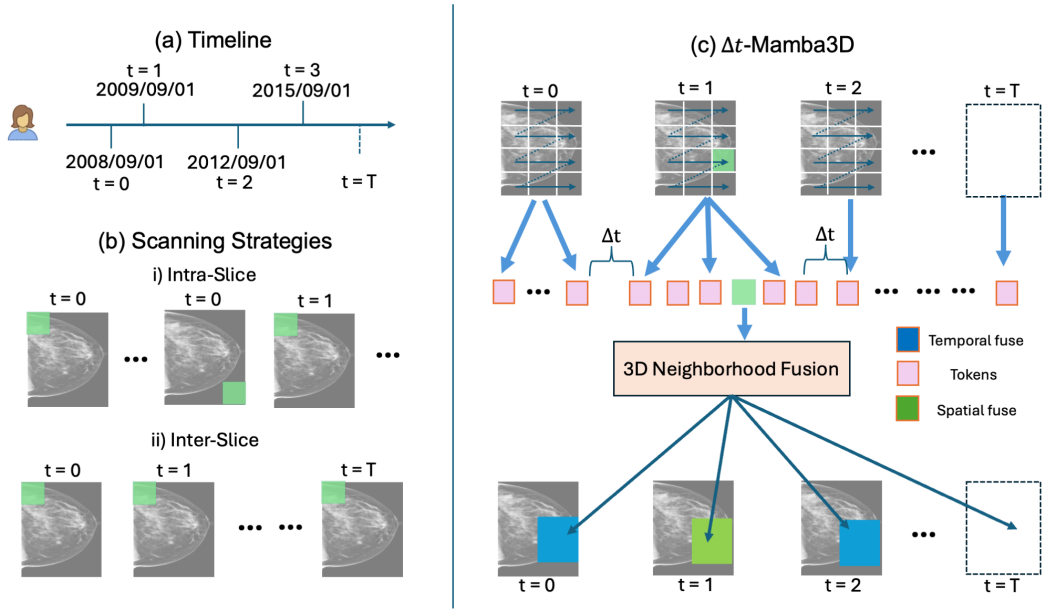

Figure 1: (a) Illustration of a patient's sequential imaging data acquired with irregular inter-visit gaps Δt (e.g., 2008 $\rightarrow$ 2012 $\rightarrow$ 2015). (b) Different scanning strategies for spatio-temporal feature volumes. (c) The scanning mechanism in the proposed method Δt-Mamba3D: time-aware scan modulated by inter-visit gaps Δt with learnable multi-scale 3D neighborhood fusion.

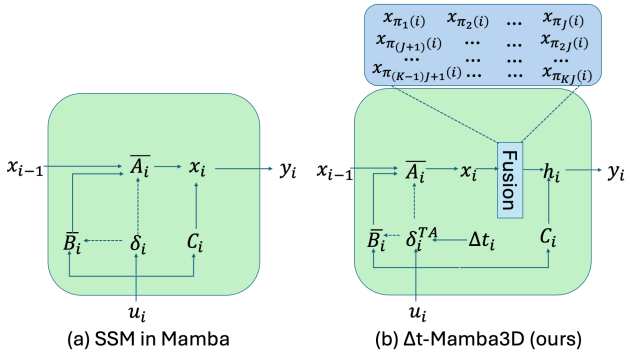

Figure 2: State-space modules. (a) Standard Mamba SSM: input u_i produces parameters $(\bar{A}_i, \bar{B}_i, \delta_i, C_i)$ that update state x_i and emit y_i. (b) Time-Aware Δt-Mamba3D (ours) generalizes to irregularly sampled spatio–temporal grids: each token's step size is modulated by the inter-visit gap Δt_i, and a multi-scale 3D neighborhood fusion aggregates local structure across visits before producing h_i and y_i, where J is the number of spatial neighbors and K the number of temporal neighbors.

(Williams, Lawrence et al. 2007). This dynamic system can be described by the linear state transition and observation equations (Kalman 1960). A standard linear continuous-time state-space model (SSM) can be expressed as:

$$\mathbf{h}'(i) = \mathbf{A}\mathbf{h}(i) + \mathbf{B}\mathbf{x}(i), \ \mathbf{y}(i) = \mathbf{C}\mathbf{h}(i) + \mathbf{D}\mathbf{x}(i). \quad (1)$$

where $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are the weighting trainable parameters, and $\mathbf{D}$ always equals to 0. To effectively integrate continuous-time SSMs into the deep learning framework, it is essential to discretize the continuous-time models. Sampling this continuous-time SSM at intervals of size δ (assuming zero-order hold) yields the discrete counterpart:

$$\mathbf{h_i} = \bar{\mathbf{A}}\mathbf{h}_{i-1} + \bar{\mathbf{B}}\mathbf{x}_i, \ \mathbf{y}_i = \mathbf{C}\mathbf{h}_i + \mathbf{D}\mathbf{x}_i, \quad (2)$$

with

$$\bar{\mathbf{A}} = e^{\delta \mathbf{A}}, \ \bar{\mathbf{B}} = \left(e^{\delta \mathbf{A}} - \mathbf{I}\right) \mathbf{A}^{-1}\mathbf{B}.$$

By transforming the parameters from $(\delta, \mathbf{A}, \mathbf{B})$ to $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$, the SSM model becomes a sequence-to-sequence mapping framework from discrete input to output.

Real-world dynamics are seldom linear time-invariant (LTI); their behaviour shifts with context, input, and time. As shown in Mamba (Gu and Dao 2023), making the state-space model content-varying allows the network to focus on relevant signals and better capture nonstationary processes. Mamba achieves this by modulating the SSM parameters with selective, data-dependent gates, yielding a context-aware adaptive transition. This is achieved by modifying the parameters as functions of the input sequence. For each token $u_i \in \mathbb{R}^d$ at step i with d-dimensional, we compute adaptive parameters:

$$[\hat{\delta}_i, B_i, C_i] = W_{proj} u_i + \mathbf{b}_{proj}, \ \delta_i = \text{softplus}(\hat{\delta}_i). \quad (3)$$

Here, a single linear projection with weights $W_{proj}$and bias b_{proj} of the input token $u_i \in \mathbb{R}^d$ produces three vectors: (1) $\hat{\delta}_i$: a step–size logit. After softplus, $\delta_i > 0$ is used as the content-dependent step in the discretized SSM update; (2) B_i: an input gate that scales the driving term of the state update; (3) C_i: an output/skip gate that scales the direct contribution of the input to the output.

Formulation of Time-Aware Δt-Mamba3D

Time-Aware Δt-Mamba3D aims to capture the clinically important inter-visit time gap and the spatial–temporal dependencies among neighbouring latent states. Unlike earlier visual Mamba variants that rely on multiple scan directions and content-only adaptive steps, we introduce two key modifications: (i) the true time gap Δt is injected directly into the selective scan, and (ii) a 3D neighborhood fusion term is added to the original Mamba equations. The detailed workflow is illustrated in Fig. 2

Time-aware Modulation Mamba (Gu and Dao 2023) is designed for language modeling, implicitly assuming uniform steps between tokens (i.e., $\Delta t = 1$). To mimic variable dependence, it uses the current input u_i as a control signal and predicts an effective step via $\delta_i = \mathrm{softplus}(\mathrm{Linear}(u_i))$, selectively copying past inputs. However, this content-gated surrogate does not encode actual, irregular clock times—ubiquitous in clinical longitudinal data. We therefore augment the model with explicit time information, using real Δt (exam gaps) alongside u_i as control signals. Let Δt_i denote the real calendar interval (in months) between the current time t_i and its preceding time t_{i-1}, with a minimum interval $\tau_{\min} = 12$ months for normalizing time gap. The time-aware step size Δ_i is defined by:

$$\delta_i^{\mathrm{TA}} = \delta_i \left(1 + \gamma \frac{\Delta t(i)}{\tau_{\min}}\right), \quad 0 < \gamma \leq 1. \tag{4}$$

Thus, tokens within the same visit retain the original microscopic step δ_i, whereas tokens at visit boundaries scale proportionally to the actual elapsed interval, which decide how much past information is carried forward. This ensures numerical stability and meaningful time-awareness, where γ is a mixing coefficient that controls how strongly real time stretches the step size. To theoretically demonstrate that the proposed time-aware encoding can control the importance between historical memory and current input, we give the following Theorem proven by (Li et al. 2024).

Theorem 1 *Let $A = V\Lambda V^{-1}$ with eigenvalues $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ and $\mathrm{Re}(\lambda_i) \leq 0$.*

$$\bar{A}_k = \mathrm{diag}\left(e^{\lambda_1 \delta_i^{TA}}, \ldots, e^{\lambda_N \delta_i^{TA}}\right), \tag{5}$$

$$\bar{B}_k = \left(\lambda_1^{-1}\left(e^{\lambda_1 \delta_i^{TA}} - 1\right), \ldots, \lambda_N^{-1}\left(e^{\lambda_N \delta_i^{TA}} - 1\right)\right). \tag{6}$$

The k-th coordinate-wise hidden of h_i state update is:

$$h_{k,i} = e^{\lambda_k \delta_i^{\mathrm{TA}}} h_{k,i-1} + \lambda_k^{-1}(e^{\lambda_k \delta_i^{\mathrm{TA}}} - 1) u_{k,i}. \tag{7}$$

According to the Theorem, if δ_i^{TA} is small enough, we observe $h_{k,i} \approx h_{k,i-1}$, demonstrating that a short δ_i^{TA} arising from a small visit gap or uninformative content persists in the historical state that ignores the current input. And a larger δ_i^{TA} drives the $\lambda_k \delta_i^{\mathrm{TA}} \ll 0$ and thus $h_{k,i} \approx -\lambda_k^{-1} u_{k,i}$ meaning the contribution from the previous state is forgotten and the update is dominated by the current input.

3D Neighborhood Fusion After selectively scanning all tokens by injecting time aware visit gap, if dependencies are still modeled along a single 1D order, it will leave residual spatial-temporal interactions underexplored. We therefore apply a 3D neighborhood fusion using depth-wise 3D convolutions aiming to capture the spatial and temporal dependencies of neighboring features in the latent state space:

$$\begin{aligned} x_i &= \bar{A}\, x_{i-1} + \bar{B}\, u_i, \\ h_i &= \sum_{k \in \Omega} \alpha_k\, x_{\pi_k(i)}, \\ y_i &= C_i\, h_i + D\, u_i, \end{aligned} \tag{8}$$

where x_i is the original state variable, h_i is the spatio-temporal aware state variable, Ω is the neighbor set, a_k is a learning weight, and π_k indexes the 3D coordinates of the k-th neighbor. The original state variable x_i is directly influenced by its previous state with newly added Δt_i while the spatio-temporal aware variable h_i incorporates additional neighboring state variables through a fusing mechanism. For each state x_i, we linearly weight its neighboring states $\pi_k(i)$ in Ω with coefficients α_k to integrate spatiotemporal context into a new state h_i. By considering both the global long-range and the local spatial and temporal information, the fused state variable gains a richer context.

To capture residual spatio-temporal dependencies, we linearly re-weight adjacent states with depth-wise 3-D convolutions. Given the maximum number of visits T, we employ two asymmetric kernels: $(1, 3, 3)$ to capture purely spatial context and $(\min(3, T), 3, 3)$ to capture joint spatio–temporal information.

$$\mathcal{K} = \Big\{(1, 3, 3),\ (\min(3, T),\ 3,\ 3)\Big\}.$$

and fuse their responses as

$$\mathbf{h} = \sum_{s \in \mathcal{K}} \beta_s\ \text{DW-Conv3D}^{(s)}(\hat{\mathbf{x}}),\ \beta_s = \frac{e^{\alpha_s}}{\sum_{s \in \mathcal{K}} e^{\alpha_s}}. \tag{9}$$

Let $\hat{\mathbf{x}} \in \mathbb{R}^{d \times T \times H \times W}$ be the state tensor reshaped to its spatio-temporal grid. We apply a depth-wise 3D convolution $\text{DW-Conv3D}^{(s)}(\cdot)$ with kernel size s, padding to the same size, and groups $= d$, so that each of the d channels is filtered independently. This preserves the spatial–temporal receptive field of a standard Conv3D while reducing parameters and FLOPs by a factor of d. The α_s are learnable logits; β_s are their softmax-normalized weights, ensuring $\beta_s \geq 0$ and $\sum_s \beta_s = 1$.

Model Architecture

As shown in Fig. 3, Each patient contributes a longitudinal series of imaging visits at irregular times $0 < t_1 < \cdots < t_T$. At visit t, a standard mammography study provides four projections/views: left/right craniocaudal (L/R-CC) and left/right mediolateral oblique (L/R-MLO). Let $\mathbf{I}_{t,v} \in \mathbb{R}^{3 \times H \times W}$ denote the v-th view ($v \in \{1, \ldots, 4\}$). We process each image with a Swin-V2 backbone (Liu et al. 2021) that produces a low-resolution feature map

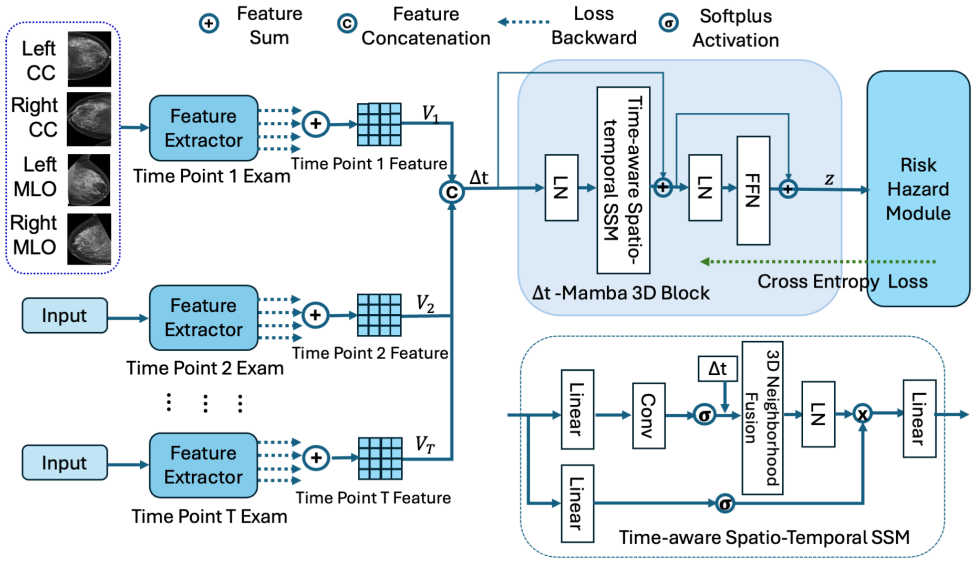

Figure 3: Overall architecture of the proposed Time-Aware Δt-Mamba3D. Top: longitudinal multi-visit, 4-view mammography is encoded per view with Swin-V2, fused by visit, and processed by a hierarchy of Δt-Mamba3D blocks before the Risk Hazard Module. Bottom right: Expanded diagram of a single Δt-Mamba3D block.

$\mathbf{F}_{t,v} = \text{Swin}(\mathbf{I}_{t,v}) \in \mathbb{R}^{d\times H_0\times W_0}$ ($H_0, W_0 \ll H, W$; e.g., H_0=W_0=8, d=768). Because radiologists integrate information across symmetric left/right and CC/MLO views, we fuse per-visit features by summation, yielding a fused spatial tensor $\mathbf{V}_t \in \mathbb{R}^{d\times H_0\times W_0}$. Stacking across valid visits gives $\mathbf{V} \in \mathbb{R}^{d\times T\times H_0\times W_0}$.

Time-Aware Δt-Mamba3D block. We apply the proposed spatio-temporal state-space block to $\mathbf{V}$ together with the inter-visit gaps $\Delta t = t_i - t_{i-1}$ ($\Delta t_1 = 0$). The block: (i) flattens the $T_0\times H_0\times W_0$ extracted feature to a token sequence, (ii) runs a Mamba selective scan whose state update is modulated by the true Δt for all tokens belonging to visit t, and (iii) reshapes back to 3D format and applies a learnable mixture of different depth-wise 3D kernels for 3D neighborhood fusion. The output of the block is $\mathbf{Z} \in \mathbb{R}^{d\times T\times H\times W}$.

Patient embedding and Risk Hazard Module. After encoding the irregular time gaps and the spatio-temporal context, we aggregate $\mathbf{V}$ across space (mean pool over H_0, W_0) and time (masked mean over the T valid visits) to obtain a patient embedding $\mathbf{z} \in \mathbb{R}^d$. This embedding feeds the Risk Hazard Module. In our breast cancer risk prediction setting we utilize an additive hazard (Yala et al. 2021; Karaman et al. 2024) and integrate the entire history embedding z, to estimate the future risk of developing breast cancer. The cumulative risk over k years (where $k \in \{1, 2, \ldots, 5\}$, as we target to predict 1- to 5-year risk) is computed by summing the baseline risk $B_r(z)$ with the annual hazard term $H_i(z)$

$$P(t_{\text{cancer}} = k \mid z) = \sigma\left(B_r(z) + \sum_{i=1}^{k} H_i(z)\right). \quad (10)$$

Model	Params (M)	FLOPs (G)	Peak tokens/s ($\times 10^6$)
I3D (3D CNN)	15.9	8.2	5.9
Transformer	7.1	4.0	11.9
GRU-Δt	3.5	0.06	79.0
Neural ODE (6 steps)	1.2	0.36	56.5
ContiFormer	7.5	4.1	11.6
TimeSformer	7.1	2.4	19.6
SegMamba	5.3	0.90	49.7
LongMamba	5.4	0.90	49.7
Δt-Mamba3D (ours)	**1.8**	**0.32**	**59.3**

Table 1: Single-layer efficiency on a 512-token input ($T = 8$, $H = W = 8$) with hidden width $d = 768$. FLOPs are fused multiply–adds (billions). "Peak tokens/s" is an estimate for an FP16 A100 (312 TFLOP/s) at 30 % utilization.

Handling variable-length series. Because patients have different numbers of exams, we left-pad every sequence to a fixed maximum length T. A binary mask suppresses padded tokens in the temporal pooling step and supplies the correct Δt values (zero at padded positions).

Experiments

Study Cohorts and Datasets

Our experiments used two independent patient cohorts and imaging datasets. The first is the Karolinska Case-Control **(CSAW-CC) Dataset** (Strand 2022), which is a part of the Cohort of Screen-Aged Women (CSAW). The CSAW-CC dataset was specifically curated for developing breast imaging-based AI tools. It includes women aged 40–74 years old who underwent mammographic screening between

Model	CSAW Dataset(max 4 prior exams)					Independent Dataset (max 8 prior exams)					
	C-index	AUC_{2y}	AUC_{3y}	AUC_{4y}	AUC_{5y}	C-index	AUC_{1y}	AUC_{2y}	AUC_{3y}	AUC_{4y}	AUC_{5y}
Time-aware (pooled) baselines											
GRU-Δt	0.661±0.03	0.677±0.01	0.661±0.01	0.647±0.01	0.643±0.01	0.576±0.01	0.620±0.01	0.610±0.01	0.587±0.01	0.573±0.01	0.571±0.02
Transformer	0.635±0.01	0.634±0.02	0.631±0.01	0.625±0.01	0.623±0.01	0.647±0.04	0.659±0.01	0.629±0.02	0.605±0.01	0.584±0.02	0.573±0.02
Neural ODE	0.649±0.01	0.657±0.01	0.656±0.01	0.658±0.01	0.658±0.01	0.642±0.02	0.651±0.03	0.652±0.04	0.641±0.02	0.626±0.01	0.609±0.03
ContiFormer	0.659±0.01	0.661±0.01	0.673±0.01	0.681±0.01	0.680±0.01	0.714±0.02	0.695±0.03	0.709±0.03	0.694±0.04	0.665±0.05	0.643±0.05
Δt-Mamba	0.716±0.01	0.738±0.02	0.714±0.01	0.700±0.01	0.695±0.02	0.717±0.01	0.731±0.01	0.726±0.01	0.708±0.02	0.677±0.03	0.657±0.04
Spatio-temporal baselines (uniform time)											
I3D	0.690±0.01	0.700±0.01	0.688±0.02	0.680±0.02	0.673±0.02	0.712±0.01	0.722±0.01	0.717±0.02	0.683±0.02	0.654±0.02	0.636±0.01
TimeSformer	0.662±0.01	0.679±0.01	0.674±0.01	0.636±0.02	0.633±0.02	0.717±0.01	0.725±0.01	0.709±0.01	0.687±0.02	0.665±0.02	0.643±0.04
SegMamba	0.704±0.01	0.732±0.03	0.692±0.01	0.684±0.01	0.672±0.01	0.714±0.02	0.723±0.02	0.719±0.01	0.709±0.01	0.688±0.01	0.658±0.01
LongMamba	0.711±0.02	0.722±0.03	0.707±0.02	0.698±0.02	0.689±0.03	0.712±0.01	0.695±0.01	0.695±0.02	0.671±0.02	0.656±0.01	0.621±0.01
Mamba3D	0.713±0.02	0.737±0.02	0.707±0.02	0.710±0.02	0.708±0.03	0.716±0.03	0.729±0.03	0.719±0.02	0.715±0.01	0.698±0.04	0.679±0.02
Δt-Mamba3D	**0.742±0.01**	**0.754±0.02**	**0.743±0.02**	**0.730±0.01**	**0.720±0.01**	**0.738±0.01**	**0.749±0.02**	**0.752±0.02**	**0.733±0.03**	**0.719±0.04**	**0.705±0.05**

Table 2: Performance comparisons (mean ± std) on the **CSAW** (max 4 exams) and **Independent** (max 8 exams) datasets. All models share the same Swin-V2 per-visit encoder; model names indicate whether inter-visit gaps (Δt) are used.

Model	CSAW Dataset(max 4 prior exams)					Independent Dataset (max 8 prior exams)					
	C-index	AUC_{2y}	AUC_{3y}	AUC_{4y}	AUC_{5y}	C-index	AUC_{1y}	AUC_{2y}	AUC_{3y}	AUC_{4y}	AUC_{5y}
Baseline	0.701±0.02	0.726±0.02	0.694±0.05	0.680±0.05	0.670±0.06	0.688±0.01	0.699±0.03	0.697±0.01	0.676±0.01	0.660±0.01	0.642±0.02
$k = \{1, 3, 3\}$	0.705±0.01	0.730±0.01	0.709±0.01	0.699±0.01	0.697±0.01	0.710±0.03	0.724 ±0.04	0.702±0.02	0.694±0.02	0.688±0.01	0.649±0.01
$k = \{3, 3, 3\}$	0.702±0.01	0.725±0.01	0.695±0.02	0.692±0.01	0.682±0.00	0.718±0.01	0.718±0.01	0.715±0.01	0.702±0.01	0.681±0.02	0.661±0.02
$k = \{1\&3, 3, 3\}$	0.713±0.02	0.737±0.02	0.707±0.02	0.710±0.02	0.708±0.03	0.716±0.03	0.729±0.03	0.719±0.02	0.715±0.01	0.698±0.04	0.679±0.02
Inter-Slice Δt-Mamba	0.682±0.01	0.691±0.01	0.658±0.01	0.657±0.01	0.662±0.02	0.692±0.01	0.707±0.01	0.706±0.02	0.697±0.05	0.675±0.06	0.656±0.06
Δt-Mamba	0.716±0.01	0.738±0.02	0.714±0.01	0.700±0.01	0.695±0.02	0.717±0.01	0.731±0.01	0.726±0.01	0.708±0.02	0.677±0.03	0.657±0.04
Δt-Mamba3D	**0.742±0.01**	**0.754±0.02**	**0.743±0.02**	**0.730±0.01**	**0.720±0.01**	**0.738±0.01**	**0.749±0.02**	**0.752±0.02**	**0.733±0.03**	**0.719±0.04**	**0.705±0.05**

Table 3: Ablation study on the spatial–temporal module design on the CSAW and Independent datasets. Values are mean ± std.

2008 and 2016 using Hologic imaging systems. To mitigate potential bias in the risk prediction due to early cancer signs or early-detectable cancers, patients diagnosed with breast cancer within six months following the "present" screening exam were excluded. Our analysis included subjects who have at least two sequential screening exams. The final CSAW-CC cohort consisted of 406 breast cancer cases (all biopsy-proven) and 6,053 normal controls, with inter-exam intervals ranging from 12 to 36 months. The second dataset (denoted as **Independent Dataset**) is a retrospectively collected case-control cohort at University of Pittsburgh Medical Center, with individuals who participated in routine breast cancer screening from 2007 to 2014 also using Hologic systems. We have data use agreement for this not-publicly-available dataset. This cohort comprises 293 breast cancer cases (all biopsy-proven) and 297 normal controls (at least 1-year follow-up to ensure normal status). Each subject had at least two sequential screening mammogram exams, with inter-exam intervals ranging from 12 to 24 months.

Implementation Details

Δt-Mamba3D models were trained to predict 1- to 5-year breast cancer risk using sequential screening mammograms. For each mammogram exam of a patient's data, it is treated as a reference time-point (Prior 0) and we then traced backward up to maximum three prior exams in CSAW dataset and up to seven in the Independent Dataset, with irregular intervals of 12–36 months between consecutive exams. Patient outcomes (e.g., cancer vs. normal status) were determined based on the next follow-up exam occurring after k years since Prior 0, where k corresponds to the prediction horizon (1–5 years) (Yala et al. 2021). All dataset splits were rigorously performed at the patient level to prevent data leakage.

We employed patient-wise 5-fold cross-validation to evaluate the performance of the proposed Δt-Mamba3D model on both datasets. In each fold, data is split into training and testing set in an 80%-20% ratio. To focus the model learning on breast tissue, we first used LIBRA (Keller et al. 2012, 2015) to segment the breasts and discard the background, producing images of size 350×400 pixels. To mitigate class imbalance in the CSAW dataset, we adopt the reweighted cross-entropy loss function. The model was trained for 30 epochs with a batch size of 8, and the best checkpoint was selected via a grid search over learning rate of 5e-5 and 1e-5. All experiments were conducted on an NVIDIA TESLA A100 GPU, courtesy of our institution's computing resources. Model performance was evaluated using C-index and the Area Under the ROC Curve (AUC), with the mean AUC and standard deviations computed over 5-fold cross-validation for predicting the 1- to 5-year risk.

We compared our method with state-of-the-art methods including: (i) vector-based time-series methods—GRU-Δt (Che et al. 2018), vanilla Transformer (Vaswani et al. 2017), Neural ODE (Rubanova, Chen, and Duvenaud 2019), and ContiFormer (Chen et al. 2023); For benchmarking, each exam is associated with a single inter-visit gap Δt, which we feed to GRU-Δt, Neural ODE, and ContiFormer while leaving all other model components unaltered. We grid-searched hidden sizes $\{256, 512\}$ and depths $\{1–3\}$ to give each model its best setting. (ii) spatio-temporal approaches-TimeSformer (Bertasius, Wang, and Torresani 2021), SegMamba (Xing et al. 2024), and LongMamba (Zhou et al. 2025). Two variants of our proposed framework were also included for comparison: Mamba3D, which uses only 3D neighborhood fusion, and Δt-Mamba, a time-aware state-space model without 3D neighborhood fusion. In addition, we also compared performance of our model to several representative longitudinal breast cancer risk models: Mirai (Yala et al. 2021), LRP-NET (Dadsetan et al. 2022), Prime+ (Lee et al. 2023), and LoMaR (Karaman et al. 2024). Finally, we compared the computational complexity of different methods.

Results

Table 2 reports validation performance of Δt-Mamba3D along with compared methods under two settings: using up to four prior exams on CSAW Dataset and up to eight exams on Independent Dataset for breast cancer risk prediction . All models utilized the same Swin-V2 per-visit encoder, differing primarily in their handling of inter-visit intervals (Δt) and in capturing spatio-temporal information.

For the CSAW dataset, time-aware pooled baselines such as GRU-Δt, Transformer, Neural ODE, ContiFormer, and Δt-Mamba achieved C-indices ranging from 0.635 to 0.716. Among these, Δt-Mamba exhibited the highest performance with a C-index of 0.716 and notable AUC scores across all evaluated intervals. Spatio-temporal baselines, including I3D, TimeSformer, SegMamba, LongMamba, and Mamba3D, showed improved performance, with Mamba3D achieving a C-index of 0.713. Our proposed Δt-Mamba3D significantly outperformed all baselines, achieving the highest C-index of 0.742 and superior AUC performance consistently across 2-year (0.754), 3-year (0.743), 4-year (0.730), and 5-year (0.720) risk prediction.

For the Independent Dataset, similar trends were observed. The time-aware pooled baselines achieved moderate performance, with Δt-Mamba reaching the highest C-index of 0.714. Spatio-temporal baselines generally showed improvements, with Mamba3D obtaining a notable C-index of 0.716. Our proposed Δt-Mamba3D model again delivered the highest performance, yielding a C-index of 0.738 and showing consistently superior AUC scores across all intervals evaluated, including 1-year (0.749), 2-year (0.752), 3-year (0.733), 4-year (0.719), and 5-year (0.705). Overall, these results underscore the effectiveness of our proposed method by incorporating true inter-visit intervals (Δt) and spatio-temporal information into the Δt-Mamba3D architecture, highlighting significant improvements over existing state-of-the-art methods, particularly on long-sequence data.

Computational Complexity

As shown in Table 1, **Δt-Mamba3D** delivers an unrivalled efficiency trade-off. With only 1.8 M parameters and 0.32 GFLOPs per 512-token layer, it achieves a throughput of 59.3 M tokens/s, much lighter and faster than a vanilla Transformer and I3D. It also surpasses other attention-based TimeSformer and ContiFormer by 3–5$\times$ in speed while using far less compute. Compared with its own family variants, SegMamba and LongMamba, it trims about two-thirds of their parameters and FLOPs yet is still at least 20% faster. Although the recurrent GRU-Δt baseline records the highest raw speed (79.0 M tokens/s), it requires twice as many parameters and lacks 3D spatial modelling depth, while the smaller Neural ODE matches throughput only at a slightly higher FLOP budget.

Ablation Study

In this section, we ablate several key components of our proposed method to evaluate effects by using the CSAW and Independent datasets. We construct the baseline model using the original mamba. 1) **Neighbor set**: We introduce two depth-wise 3-D branches: a purely spatial filter $(1 \times 3 \times 3)$ and a spatio-temporal filter $(3 \times 3 \times 3)$. Either branch alone improves upon the baseline, and using them together yields an even larger overall performance gain. **Time-aware module**: Inter-Slice Δt-Mamba denotes alternating patches across visits (inter-slice scan) while injecting the true time gap at every hop. This variant performs poorly because it (i) breaks spatial coherence, and (ii) repeatedly triggers Δt gates, adding noise and destabilizing learning. Adding Δt alone raises the model performance, and the full Δt-Mamba3D achieves the top scores across all horizons. Overall observations: (i) coupling the spatial-only branch with the spatio-temporal branch delivers the best results, because each captures complementary cues. In our setting, scanning an image intra-slice before tokenization preserves these advantages and further boosts performance; (ii) coupling time-aware scanning with 3D neighborhood fusion yields the most effective spatio-temporal integration.

Conclusion

We introduced Time-Aware Δt-Mamba3D, a spatial–temporal state–space block for modeling longitudinal breast imaging. This method modulates each state update by the true inter-visit interval Δt, and applies a learnable multiscale, depth-wise 3D neighborhood fusion module that jointly models spatio-temporal structure and irregular temporal patterns. Integrated into a breast cancer risk prediction framework (up to eight sequential imaging exams/visits; four views per exam), our model consistently outperforms recurrent, transformer, and prior visual SSM baselines in C-index and long-horizon AUCs (1–5y) under two different datasets. Our proposed method scales linearly in sequence length, enabling computationally efficient modeling of decade-long patient histories where quadratic attention methods like transformer become impractical. Future work will focus on extending the proposed method to additional medical imaging modalities and related clinical tasks.

Acknowledgments

This work was supported in part by a National Institutes of Health (NIH) Other Transaction research contract #1OT2OD037972-01; a National Science Foundation (NSF) grant (CICI: SIVD; #2115082); grant 1R01EB032896 as part of the NSF/NIH Smart Health and Biomedical Research in the Era of Artificial Intelligence and Advanced Data Science Program; an Amazon Machine Learning Research Award; and the University of Pittsburgh Momentum Funds (Scaling Grant) for the Pittsburgh Center for AI Innovation in Medical Imaging. This work used Bridges-2 at the Pittsburgh Supercomputing Center through allocation [MED200006] from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by NSF grants #2138259, #2138286, #2138307, #2137603, and #2138296. This research was also supported in computing resources by the University of Pittsburgh Center for Research Computing and Data (RRID: SCR_022735) through the resources provided by the H2P cluster, which is supported by NSF award OAC-2117681. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing official policies, either expressed or implied, of the NIH or NSF.

References

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6836–6846.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is space-time attention all you need for video understanding? In *Icml*, volume 2, 4.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Che, Z.; Purushotham, S.; Cho, K.; Sontag, D.; and Liu, Y. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1): 6085.

Chen, Y.; Ren, K.; Wang, Y.; Fang, Y.; Sun, W.; and Li, D. 2023. Contiformer: Continuous-time transformer for irregular time series modeling. *Advances in Neural Information Processing Systems*, 36: 47143–47175.

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.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A.; Goel, K.; and Ré, C. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Kalman, R. E. 1960. A new approach to linear filtering and prediction problems.

Karaman, B. K.; Dodelzon, K.; Akar, G. B.; and Sabuncu, M. R. 2024. Longitudinal Mammogram Risk Prediction. *arXiv preprint arXiv:2404.19083*.

Kazemi, S. M.; Goel, R.; Eghbali, S.; Ramanan, J.; Sahota, J.; Thakur, S.; Wu, S.; Smyth, C.; Poupart, P.; and Brubaker, M. 2019. Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321*.

Keller, B. M.; Chen, J.; Daye, D.; Conant, E. F.; and Kontos, D. 2015. Preliminary evaluation of the publicly available Laboratory for Breast Radiodensity Assessment (LIBRA) software tool: comparison of fully automated area and volumetric density measures in a case–control study with digital mammography. *Breast cancer research*, 17: 1–17.

Keller, B. M.; Nathan, D. L.; Wang, Y.; Zheng, Y.; Gee, J. C.; Conant, E. F.; and 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. *Medical physics*, 39(8): 4903–4917.

Kidger, P.; Morrill, J.; Foster, J.; and Lyons, T. 2020. Neural controlled differential equations for irregular time series. *Advances in neural information processing systems*, 33: 6696–6707.

Lee, H.; Kim, J.; Park, E.; Kim, M.; Kim, T.; and Kooi, T. 2023. Enhancing Breast Cancer Risk Prediction by Incorporating Prior Images. *arXiv preprint arXiv:2303.15699*.

Li, D.; Tan, S.; Zhang, Y.; Jin, M.; Pan, S.; Okumura, M.; and Jiang, R. 2024. Dyg-mamba: Continuous state space modeling on dynamic graphs. *arXiv preprint arXiv:2408.06966*.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; and Liu, Y. 2024. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37: 103031–103063.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3202–3211.

Nguyen, A.; Chatterjee, S.; Weinzierl, S.; Schwinn, L.; Matzner, M.; and Eskofier, B. 2020. Time matters: Time-aware lstms for predictive business process monitoring. In *International Conference on Process Mining*, 112–123. Springer.

Rubanova, Y.; Chen, R. T.; and Duvenaud, D. K. 2019. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32.

Scutt, D.; Lancaster, G. A.; and Manning, J. T. 2006. Breast asymmetry and predisposition to breast cancer. *Breast cancer research*, 8: 1–7.

Shukla, S. N.; and Marlin, B. M. 2021. Multi-time attention networks for irregularly sampled time series. *arXiv preprint arXiv:2101.10318*.

Strand, F. 2022. CSAW-CC (mammography) – a dataset for AI research to improve screening, diagnostics and prognostics of breast cancer. *Dataset*.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Williams, R. L.; Lawrence, D. A.; et al. 2007. *Linear state-space control systems*. John Wiley & Sons.

Xiao, C.; Li, M.; Zhang, Z.; Meng, D.; and Zhang, L. 2024. Spatial-mamba: Effective visual state space models via structure-aware state fusion. *arXiv preprint arXiv:2410.15091*.

Xing, Z.; Ye, T.; Yang, Y.; Liu, G.; and Zhu, L. 2024. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, 578–588. Springer.

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.

Zhou, Z.; Arefan, D.; Zuley, M.; Sumkin, J.; and Wu, S. 2025. Longmambattn: A Novel Architecture for Enhanced Breast Cancer Risk Prediction Using Variable-Length Longitudinal Mammograms. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, 1–5. IEEE.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*.