PCa-Mamba: Spatiotemporal State Space Models for Prostate Cancer Detection in Multi-Parametric MRI

Kai Zhao[1,2], Alex Ling Yu Hung[1], Kaifeng Pang[1] Parsa Hajipour[1], Holden Wu[1], Steve Raman[1], Kyunghyun Sung[1],

[1]University of California, Los Angeles [2]Shanghai University

kz@kaizhao.net, alexhung@ucla.edu
{kaifengpang, parsahajipour, holdenwu, sraman, ksung}@mednet.ucla.edu

Abstract

Multiparametric Magnetic Resonance Imaging (mpMRI), including T2-weighted imaging (T2), diffusion-weighted imaging (DWI), and dynamic contrast-enhanced (DCE) imaging is an important technique for the diagnosis of clinically significant prostate cancer (csPCa). Unlike T2 and DWI, which capture spatial contrast, DCE-MRI captures the temporal dynamics before, during, and after the administration of the contrast agent. The temporal dynamics of DCE-MRI characterize the microvascular properties of target tissue, offering valuable information for csPCa diagnosis. However, due to the high data dimensionality and spatial-temporal data incompatibility, DCE-MRI has been largely overlooked or underutilized in prevalent deep learning-based csPCa detection methods. In this paper, we propose PCa-Mamba, a state-space model (SSM)-based framework that, for the first time, fully incorporates the temporal dynamics of DCE-MRI alongside the spatial contrast of T2 and DWI for csPCa detection in mpMRI. SSMs are efficient for long-range sequence modeling, making them well-suited for high-dimensional, heterogeneous mpMRI data. PCa-Mamba comprises two SSM modules: (i) a temporal SSM that captures the temporal dynamics, and (ii) a spatial SSM that extracts spatial contrast. Pharmacokinetic (PK) regularization is introduced to constrain the temporal SSM with the Tofts model, allowing the model to learn the PK properties of the target tissue while reducing overfitting. Permuted sequentialization is introduced into the spatial SSM to rearrange spatial patches along different scan directions. This enhances spatial context while preserving local adjacency. The spatial and temporal features are jointly optimized by interactive fusion for integrated spatiotemporal representations, effectively capturing the complementary information for csPCa detection. We further introduce a dropout mechanism that randomly excludes DCE-MRI during training, allowing the model to handle practical scenarios where DCE-MRI is unavailable. Extensive experiments on our in-house dataset and the PI-CAI dataset demonstrate PCa-Mamba's superiority over existing models for csPCa detection, as well as the advantages of mpMRI over bpMRI in lesion-wise csPCa diagnosis, especially for detecting small lesions and those located in the peripheral zone.

Keywords: mpMRI, DCE-MRI, Prostate MRI, Prostate Cancer, Deep Learning, State Space Models, Mamba

1 Introduction

Prostate cancer (PCa) is the most common cancer among men in the United States [1]. Multi-parametric magnetic resonance imaging (mpMRI) of the prostate, which typically includes T2-weighted imaging (T2), diffusion-weighted imaging (DWI), and dynamic contrast-enhanced MRI (DCE-MRI), has become a central imaging modality for the diagnosis of clinically significant prostate cancer (csPCa). Each imaging component provides complementary diagnostic information for lesion localization, classification, and grading [2]. In contrast to biparametric MRI (bpMRI), which consists of T2 and DWI and primarily provides spatial contrast information, DCE-MRI captures temporal signal variations by acquiring T1-weighted images (T1) before, during, and after the administration of a contrast agent (CA). The resulting time–concentration curves (TCCs) from DCE-MRI encode quantitative information about tissue vascularization, which is closely associated with the occurrence and progression of PCa [3].

Several recent studies have reported the added value of DCE-MRI in csPCa diagnosis. Huebner et al. [4] demonstrated that incorporating DCE-MRI improves the visibility of csPCa lesions. Zhang et al. [5] reported that DCE-MRI enhances diagnostic sensitivity for transition zone (TZ) lesions and is particularly beneficial for the detection of small csPCa lesions. Similar observations were reported by Taghipour et al. [6]. Pellicer et al. [7] showed that incorporating the volume transfer constant (K^{trans}, measured in min^{-1}), derived from quantitative DCE-MRI analysis, improves the sensitivity of deep learning-based csPCa detection. The parameter K^{trans} characterizes the transfer rate of CA from blood plasma into the extracellular extravascular space (EES), thereby reflecting vessel permeability and tissue vascularization. Matsuoka et al. [8] incorporated early- and late-phase DCE-MRI images into deep learning models and also observed perfor-

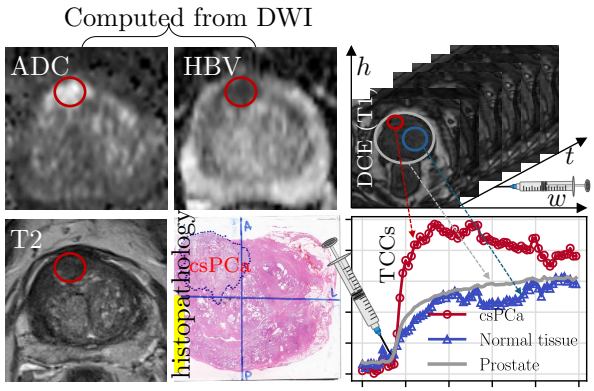

Figure 1: Prostate mpMRI images and the corresponding histological image. The csPCa lesion is located in the anterior-right region of the prostate (top-left in the image) and marked with a red circle, as indicated in the histological image. The TCCs display the average TCCs of the csPCa lesion (red), normal prostate tissue (blue), and the entire prostate (grey). TCCs increase after the injection of CA, with those in the csPCa lesion rising significantly faster and reaching higher levels than in normal prostate tissue.

mance improvements. Collectively, these studies indicate the clinical and computational value of including DCE-MRI in csPCa lesion diagnosis.

With advances in deep learning, numerous models have been proposed for automated csPCa detection [9, 7, 10] and grading [9, 11]. However, most existing approaches rely primarily on bpMRI [9, 12, 10] and exclude or underutilize DCE-MRI due to data heterogeneity and the high dimensionality of temporal information [9, 10]. Although notable performance gains have been achieved through novel loss functions [9, 10] or advanced network architectures [10, 13], there remains substantial potential to further improve sensitivity and specificity by systematically incorporating DCE-MRI information.

Recent studies have explored the integration of mpMRI for csPCa detection by incorporating either early and late DCE-MRI images [8] or parametric maps such as K^{trans} derived from DCE-MRI [7], resulting in improved detection sensitivity. However, parametric maps or sparsely sampled T1 images capture only partial vascularization information. Moreover, DCE-MRI TCCs are often noisy, which can lead to unstable estimation of pharmacokinetic (PK) parameters. In contrast, the original DCE-MRI TCCs contain more comprehensive temporal information related to tissue vascularization. Directly learning from these original TCCs therefore has the potential to further enhance csPCa detection. Nevertheless, the high dimensionality of TCCs and their incompatibility with spatial bpMRI data pose significant challenges for effective integration. For multi-slice prostate MRI, which is already high-dimensional, the inclusion of TCCs further increases the data dimensionality along the temporal axis. Capturing long-range dependencies in such high-dimensional spatiotemporal mpMRI data requires efficient deep learning models capable of modeling both spatial and temporal context.

Motivated by the recent success of state space models (SSM) [14] in long-range sequence modeling, this work proposes an SSM-based architecture for mpMRI-based csPCa detection. The proposed method, termed PCa-Mamba, comprises two core components: (i) a temporal SSM with **pharmacokinetic (PK) regularization** to extract diagnostic and physiologically meaningful features from DCE-MRI TCCs, and (ii) a spatial SSM with **permuted sequentialization** to model spatial features from T2 and DWI images. The spatial SSM follows a UNet [15]-like architecture constructed from SSM blocks, whereas the temporal SSM is a one-dimensional model designed to learn temporal dynamics from TCCs. Features extracted by the two SSM branches are adaptively combined through **interactive feature fusion**.

The **PK regularization** constrains the temporal SSM using the Tofts PK model [16], thereby enabling direct estimation of PK parameters, such as K^{trans}, from input TCCs. This design provides two main advantages. First, PK regularization encourages the extraction of features that are closely related to PK parameters, which have been shown to be strong indicators of csPCa [17, 7, 18, 19]. The incorporation of such features improves csPCa detection performance. Second, training deep models on high-dimensional and noisy data, such as DCE-MRI TCCs, with limited sample sizes is prone to overfitting. PK regularization mitigates this risk by constraining the temporal feature space. The **permuted sequentialization** strategy rearranges spatial feature sequences to enhance contextual modeling in the spatial SSM. The **interactive fusion** mechanism dynamically integrates spatial and temporal representations, facilitating mutual learning between the two SSM branches. To address the potential absence of DCE-MRI TCCs in routine clinical practice, a modality dropout mechanism is incorporated to improve robustness to missing temporal information.

With these components, the aim of this study is to develop a unified deep learning framework for clinically significant prostate cancer detection that effectively integrates spatial information from bpMRI with temporal enhancement dynamics from DCE-MRI. In contrast to prior approaches that underutilize DCE-MRI, depend on precomputed parametric maps, or model temporal information independently of spatial context, the proposed method directly learns from original DCE-MRI temporal dynamics within a coherent spatiotemporal modeling framework. The contributions of this study are summarized as follows:

- A holistic SSM-based framework, PCa-Mamba, is proposed to exploit spatiotemporal information in mpMRI for csPCa detection.
- PK regularization is introduced to integrate pharmacokinetic analysis into the temporal SSM, and an interactive fusion strategy is proposed to combine spatial and temporal features, enabling bidirectional mutual learning between the two SSM branches.
- The contribution of DCE-MRI is systematically evaluated across different csPCa lesion characteristics, demonstrating particular benefits for peripheral zone lesions, small lesions, and MRI-negative lesions.

The remainder of this paper is organized as follows. Sec. 2 reviews related work. Sec. 3 describes the proposed method. Sec. 4 presents the experimental setup and results. Sec. 5 provides a discussion of the findings. Sec. 6 concludes the paper.

2 Related Work

2.1 Deep Learning-based MRI csPCa diagnosis

A line of deep-learning models has been proposed for MRI csPCa diagnosis in the past decade. These methods extract pixel or region-level features from MRI images and predict csPCa locations or probability maps. Song et al. [20] used region-based Convolutional Neural Networks (CNNs) to classify biopsy-proven prostate cancer lesions and non-lesion tissue. Kiraly et al. [21] used an encoder-decoder architecture with two output branches for csPCa detection and classification. Seetharaman [22] proposed to distinguish indolent PCa and csPCa with a T2 and Apparent Diffusion Coefficient (ADC) images. Cao [9] proposed the ordinal encoding to distinguish csPCa of different severity for multi-class csPCa detection. A mutual finding loss was proposed to simulate radiological interpretation of csPCa from T2 and ADC. Li [23] used dilated convolution networks to simultaneously segment prostate and PCa. A line of work has introduced the anatomical zonal information into the csPCa detection framework. Hoss et al. [24] incorporated zonal segmentation with mpMRI as the model input. Vente et al. [25] integrated zonal information with ordinal encoding to detect PCa of different aggressiveness. Duran et al. [26] confirmed the importance of zonal masks in csPCa detection by comparing the performance with prostate and zonal segmentations. A number of network architectures have also been introduced to improve csPCa detection performance. Yu et al. [27] proposed a multi-scale patch-wise network enhanced with a squeeze-and-excitation (SE) block. This architecture is designed to reduce false positive (FP) predictions by letting the model automatically learn the FP patterns from the context information provided by the multi-scale patches. Saha et al. [28] introduced an auxiliary network to classify if a given image patch contains csPCa lesions and then used the classification results to depress FP predictions. Recently, Zheng et al. [10] took into account the anatomical asymmetry of csPCa lesions and proposed an anatomy-aware network architecture, along with a loss function, to reduce FP detections with symmetric patterns. Hung et al. [29] utilized cross-slice information to perform evidential deep learning for csPCa lesion detection.

2.2 DCE-MRI in csPCa Diagnosis

DCE-MRI measures changes in CA concentration over time and helps identify areas with abnormal vascularization, often associated with tumors. The analysis of DCE-MRI involves the quantification and interpretation of DCE-MRI TCCs. In general, there are qualitative and quantitative methods for prostate DCE-MRI analysis. Qualitative methods involve visually assessing the TCCs in the target area, as the csPCa demonstrates more rapid and more intense enhancement [30, 31]. Fig. 1 illustrates the average TCCs observed in normal tissue compared to csPCa. Quantitative DCE-MRI leverages pharmacokinetic (PK) models to quantify contrast agent (CA) exchange between the intravascular space and the EES. The widely used Tofts model [16] estimates PK parameters by fitting them to TCCs. The PK parameters are relevant to physiological properties of the target tissue and are indicative of csPCa [17, 7, 18, 19]. In addition to the Tofts model, MR dispersion imaging (MRDI) [32] was proposed to characterize microvascular architecture by quantifying intravascular dispersion and permeability without requiring explicit arterial input function (AIF) measurement or assumptions, and has demonstrated improved accuracy for clinically significant prostate cancer (csPCa) detection [33, 17].

Although limited benefits have been reported in patient-wise csPCa diagnostics, DCE-MRI has shown added value in lesion-level csPCa detection [4, 7, 8]. The visibility of lesions was significantly better with mpMRI when compared to mpMRI in visual grading characteristic analysis [4].

Recently, deep learning-based methods [18, 34] have been proposed to directly estimate PK parameters from TCCs. Another line of work utilizes DCE-MRI or its derived parameters as input features for deep learning models. Pellicer-Valero et al. [7] incorporated K^{trans} parametric maps derived from TCCs as model input, achieving notable improvements in lesion-wise csPCa detection but minimal gains in patient-wise diagnostics. Matsuoka et al. [8] included DCE images of early and late phases

as input. However, neither method accounts for the full temporal dynamics of DCE-MRI, resulting in limited utilization of DCE-MRI.

2.3 State Space Models for Long Sequence Modeling

Gu et al. [35] proposed the Structured State-Space Sequence (S4) model, an alternative to Transformers for sequential modeling. Due to their linear complexity with respect to sequence length, SSMs have attracted a number of further explorations. Smith et al. [36] simplified S4 by replacing multiple single-input, single-output SSMs, with one multi-input, multi-output SSM. [37] proposed an SSM layer that nearly fills the performance gap between SSMs and Transformer attention in language modeling.

The original SSMs were designed for one-dimensional sequences, while vision-related tasks require processing multi-dimensional inputs like images. Inspired by the visual transformer the vision mamba block [38] was proposed for multi-dimensional image data. SegMamba [39] extended SSMs into volumetric medical image segmentation.

3 Methodology

In this section, we present PCa-Mamba for csPCa detection. We begin with a brief overview of SSMs and the SSM block in Sec. 3.1, followed by details on spatial SSMs in Sec. 3.2 and temporal SSMs in Sec. 3.3. Finally, we discuss spatial-temporal interactions and the end-to-end training of PCa-Mamba in Sec. 3.4.

3.1 State Space Models

The state space models (SSMs) are inspired by the continuous system that maps a 1-D sequence $x_t \in \mathbb{R} \mapsto y_t \in \mathbb{R}$ through a hidden state variable $h_t \in \mathbb{R}^N$. This state h_t captures the internal status of the system at time step t and summarizes all necessary past information to determine future behavior. This system uses $\mathbf{A} \in \mathbf{R}^{N\times N}$ as the evolution parameter and $\mathbf{B} \in \mathbb{R}^{N\times 1}, \mathbf{C} \in \mathbb{R}^{1\times N}$ as the projection parameters. The discrete version of the system can be formulated as:

$$\begin{aligned} h_t &= \mathbf{A}h_{t-1} + \mathbf{B}x_t \\ y_t &= \mathbf{C}h_t, \end{aligned} \tag{1}$$

where x_t and y_t are the input and output of the t-th node, and h_{t-1} is the hidden state of the previous node. Eq. (1) can be reformulated as a parallel computation:

$$\begin{aligned} K &= [CB, CAB, CAAB, ..., CA^{T-1}B] \\ \mathbf{y} &= x \circledast K \end{aligned} \tag{2}$$

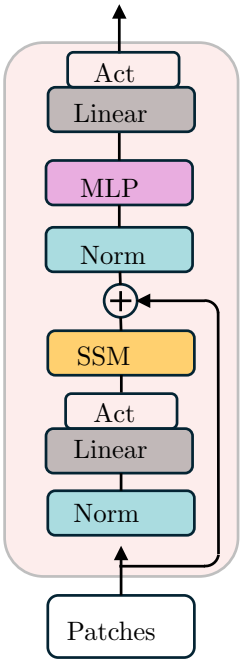

Figure 2: The SSM block. The input sequence is initially normalized using LayerNorm and then projected through a linear layer followed by the SiLU activation function (Act). The SSM captures the global context, and a residual connection is added to the SSM's output. The features undergo a second normalization and are then processed by an MLP, followed by another linear layer and activation.

where $\mathbf{x} = [x_1, x_2, ..., x_L], \mathbf{y} = [y_1, y_2, ..., y_L]$ are the input and output, $\circledast$ is the convolution operator, L is the length of the sequence, and K is the convolution kernel. Eq. (2) enable efficient parallelizable computation and makes SSMs much more efficient than recurrent models. The hidden state h_t in Eq. (1) contains information on all previous nodes in the sequence. The representation capacity of h_t is limited, and it struggles to fully encapsulate context of very distant nodes.

Eq. (1) and (2) enables partial global context, allowing the output y_t to access all inputs prior to t. However, it does not permit access to inputs occurring after t. To enable the bidirectional information access and learning global context, we reverse the input sequence and add the outputs of forward and reverse SSM at each time step:

$$\begin{aligned} [\overrightarrow{y}_1, \overrightarrow{y}_2, ..., \overrightarrow{y}_L] &= [x_1, x_2, ..., x_L] \circledast K \\ [\overleftarrow{y}_L, \overleftarrow{y}_2, ..., \overleftarrow{y}_1] &= [x_L, x_{L-1}, ..., x_1] \circledast K \\ \mathbf{y} = [y_1, y_2, ..., y_L] &= [\overrightarrow{y}_1 + \overleftarrow{y}_1, \overrightarrow{y}_2 + \overleftarrow{y}_2, ..., \overrightarrow{y}_L + \overleftarrow{y}_L] \end{aligned} \tag{3}$$

In Eq. (3), $\overrightarrow{y}_t$ and $\overleftarrow{y}_t$ are the outputs of the forward and reverse SSMs at time step t, which have access to information prior to and after t, respectively. $y_t = \overrightarrow{y}_t + \overleftarrow{y}_t$ is the final output at time step t, which capture the global context from both directions. From this point forward, we

will use $\mathbf{y} = \text{SSM}(\mathbf{x})$ to represent the bidirectional SSM defined in Eq. (3).

We follow the design of vision transformer and replace the self-attention modules with SSMs defined in Eq. (3) to form SSM-based networks that are more efficient in modeling long-term dependencies.

An SSM block comprises several operations. The input sequence $\mathbf{x} = [x_1, x_2, ..., x_L]$ is first normalized, then passed through the SSM operation, and finally processed by a multi-layer perceptron (MLP) to produce the output $\mathbf{y} = [y_1, y_2, ..., y_L]$. The computation flow of a single SSM block is illustrated in Fig. 2.

3.2 Spatial SSM with Permuted Sequentialization

The spatial SSM extracts features from spatial images, including T2, ADC, HBV, and zonal masks. These images are stacked as a multi-channel multi-slice volume. The input images are first transformed into the sequence of patches, and then processed by multi-layer U-shaped SSM encoder to learn hierarchical representations. Zonal masks were used for both the in-house and PI-CAI datasets. They were generated using a zonal segmentation model [40] pretrained on the in-house dataset, which contains human-annotated zonal ground truths. To avoid any potential data leakage, the zonal segmentation model was trained on exactly the same set of cases as PCa-Mamba. The resulting pretrained model was then applied to generate zonal masks for the PI-CAI dataset.

3.2.1 Multi-slice MRI to a Patch Sequence

The SSMs are designed for the sequence-to-sequence problems. To adapt SSMs to our task, we first transform the multi-slice images into flattened square patches. The flattened patches are regarded as a sequence where each patch denotes a node in the sequence.

Let $I \in \mathbb{R}^{H \times W \times D}$ be the spatial data of a multi-slice mpMRI scan including T2, ADC, HBV, and zonal masks, where $H \times W$ is the in-plane resolution of each slice, D is the number of slices, and $C = 4$ is the number of channels. In our experiments, the four images are stacked along the channel dimension, resulting in $C = 4$.

We split the stacked images into square, non-overlapping patches in the $H \times W$ plane. Let $\mathbf{P}_l \in \mathbb{R}^{p^2 \times C}$ be the l-th patch where p is the patch size, and $L = H/p \times W/p \times D$ is the number of patches (or length of the sequence). Following the vision transformer, we employ 16×16 patch size to get a sequential representation of non-overlapping patches. Next, we linearly project each patch into d-dimensional embeddings and add position embeddings $\mathbf{E}_{pos} \in \mathbb{R}^{L \times d}$. The input sequence to the spatial SSM is

$$\mathbf{x} = [\mathbf{P}_1 W, \mathbf{P}_2 W, ..., \mathbf{P}_L W] + \mathbf{E}_{pos} \quad (4)$$

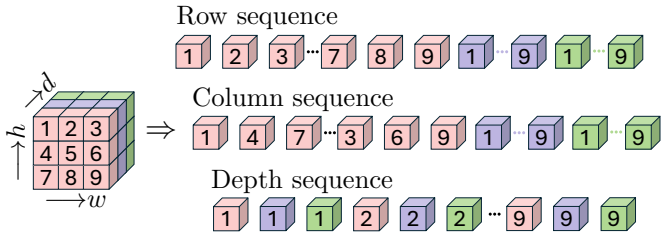

Figure 3: Illustration of the permuted sequentialization of the 3D spatial bpMRI data. The 3-D image is sequentialized along different axes, i.e. *w*idth, *h*eight, and *d*epth, ensuring that spatially neighboring voxels have shortcut access to the context of each other in at least one of the sequences.

where $W \in \mathbb{R}^{(p^2 \times C) \times d}$ is the learnable projection matrix. $\mathbf{x} \in \mathbb{R}^{L \times d}$ is the projected embeddings sequence.

With the aforementioned sequentialization, the sequence length would be $L = H/p \times W/p \times D$. For example, with patch size $p = 16$, our in-house mpMRI data with dimension $128 \times 128 \times 20$, the sequence length will be $L = 128/16 \times 128/16 \times 20 = 1280$.

3.2.2 Permuted Sequentialization

The SSMs compress contextual information into the hidden state h_t and therefore allow efficient linear-time reasoning. However, the context of distant nodes gradually diminishes over the recurrence, potentially leaving a node with insufficient context for distant nodes. Let $\mathbf{P}_{i,j,k}$ be a patch in a multi-slice MRI, where i, j are the in-plane indices of the patch and k is the slice index. $\mathbf{P}_{i,j,k}$ and $\mathbf{P}_{i,j,k\pm1}$ are spatially adjacent patches in the same in-plane position but with adjacent slices. However, after sequentialization, the two patches become distant from each other, resulting in insufficient mutual context. These spatially adjacent patches should share rich context with each other to learn consistent features.

To ensure that spatially adjacent patches have sufficient context of each other, we propose the permuted sequentialization that rearranges the sequence with different permutations of the patches. As illustrated in Fig. 3, the original 3D spatial bpMRI is sequentialized along the row, column, and depth axes, resulting in three distinct permutations. The permuted sequentialization ensures that spatially neighboring patches, such as two patches in the same row but in adjacent columns within the same slice, are close to each other in at least one of the sequentializations. To enhance contextual modeling without increasing computational cost, we apply a single permutation within each spatial SSM block and alternate the permutation scheme across successive blocks. For instance, row-, column-, and depth-wise sequentialization are used in the first, second, and third spatial SSM blocks, respectively, and this alternation is repeated in subsequent blocks.

3.2.3 Spatial SSM with U-shaped architecture

Using the SSM block and permuted sequentialization, we construct the spatial SSM network, which resembles the UNet architecture by progressively downsampling and then upsampling features to learn hierarchical representations at various scales. This architecture is similar to recent SSM-based models such as Vision Mamba [38] and SegMamba [39]. Fig. 4 illustrates the architecture of the spatial SSM. The spatial SSM has three downsampling and three upsampling stages. Each stage consists of three SSM blocks and a patch merging (or expanding) module. The three SSM blocks utilize row, column, and depth sequentialization to learn the spatial correspondences. As shown in Fig. 4, the patch merging module combines $2 \times 2 = 4$ neighboring patches by concatenating their embeddings and then applying a linear projection to convert the embeddings back to the original dimension. The patch expanding module first increases the embedding dimension by a factor of four using a linear projection, and then splits the embeddings into four separate patches. It is important to note that both patch merging and patch expanding are performed within the in-plane $H \times W$ plane, while the depth axis (number of slices) remains unchanged.

3.3 Temporal SSM with Pharmacokinetic Regularization

3.3.1 Temporal SSM

The temporal SSM extracts temporal dynamics from DCE-MRI time-concentration curves (TCCs). Let $C_t \in \mathbb{R}^{H\times W\times D\times T}$ be the TCCs of DCE-MRI, computed from the acquired DCE-MRI T1 images using the method described in [17]. $H\times W\times D$ is the spatial dimension, $T = 75$ is the total temporal steps of the DCE-MRI data.

The temporal SSM operates on the time–concentration curve (TCC) of each spatial voxel independently. Let $c_t \in \mathbb{R}^T$ be the TCC of a specific voxel, the temporal SSM extracts features from each c_t. In this setting, the sequence length of the temporal SSM is T. The extracted temporal features are combined with the features of spatial SSM for csPCa prediction. The input TCC is first projected into high-dimensional features with a linear layer and then processed by four subsequent SSM blocks.

The temporal SSM extracts temporal features F_t from input TCCs C_t, and F_t is used for both csPCa estimation and PK parameters estimation. Fig. 5 illustrates the computation flow of the temporal SSM. The PK parameter estimation is performed only during training and is essentially a regularization to the temporal SSM. We will elaborate this in the following context.

3.3.2 Tofts Pharmacokinetic Model

We used the Tofts model for PK regularization and PK parameter estimation. The Tofts model formulates the TCC (c_t) as the convolution between the plasma CA concentration (c_p), also known as the arterial input function (AIF), and the tissue impulse response:

$$c_t(t) = \int_0^t \underbrace{c_p(\tau - t_0)}_{\text{AIF}} \cdot \underbrace{K^{trans} e^{-k_{ep}\cdot(t-\tau)}}_{\text{impulse}} d\tau. \quad (5)$$

In Eq. (5), K^{trans} denotes the volume transfer constant and k_{ep} represents the rate constant, both measured in min^{-1}. The parameter t_0 indicates the bolus arrival time. All these parameters are free variables to be estimated, and represent the vascularization of the target tissue. c_p is the AIF, and the Parker population-based AIF [41] is used in our model and thus c_p is a constant. Given acquired $c_t(t)$, we can estimate the three PK parameters using the Tofts model in Eq. (5). Conversely, provided the three PK parameters K^{trans}, k_{ep} and t_0, we can calculate corresponding c_t by inverting the equation.

3.3.3 Pharmacokinetic Regularization

We constrain the temporal SSM with the Tofts model, which enables estimation of PK parameters, i.e. K^{trans}, k_{ep} and t_0, from input TCCs c_t. As shown in Fig. 5, the temporal features F_t is not only used for csPCa detection, but also used to estimate the PK parameters. The estimation of the PK parameters is enforced by a reconstruction loss. Let $\hat{K}_{trans}, \hat{k}_{ep}$, and t_0 be the estimated PK parameters, the regularization loss term is the L1 distance between the original and the fitting TCCs

$$\mathcal{L}_{pk} = L_1\big(C_t, \hat{C}_t\big) = L_1\big(C_t - \text{Tofts}(\hat{K}_{trans}, \hat{k}_{ep}, t_0)\big), \quad (6)$$

where $\hat{C}_t = \text{Tofts}(\hat{K}_{trans}, \hat{k}_{ep}, t_0)$ is the fitting, and $L_1(\cdot,\cdot)$ denotes the L1 distance. As illustrated in Fig. 5, the PK parameter estimation branch is utilized only during training and discarded during testing. The temporal SSM is constrained to emulate the Tofts model, and the learned features encapsulate both csPCa diagnostic information and vascularization characteristics of the target tissue.

3.4 Interactive Spatiotemporal Fusion

We have introduced spatial and temporal SSMs designed to learn independently from spatial and temporal MRI data, respectively. To integrate spatial and temporal features for csPCa detection and facilitate mutual learning between the SSMs, we propose an interactive fusion approach that dynamically combines spatiotemporal features.

As demonstrated in Fig. 6 (middle), after each SSM block, we alternately add features from the temporal (or

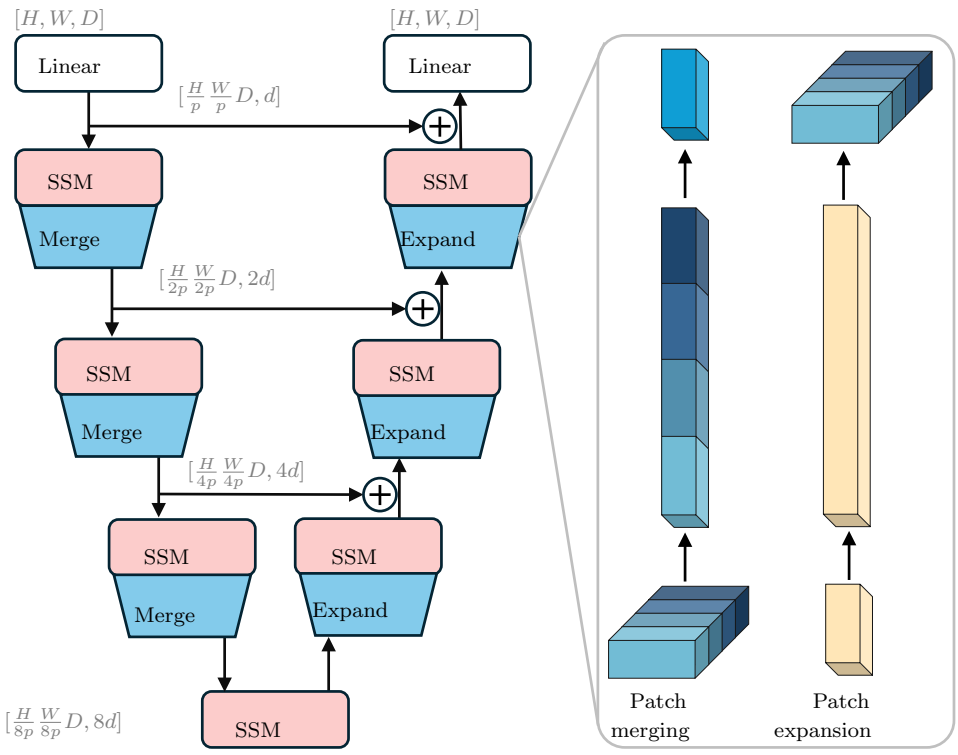

Figure 4: The architecture of spatial SSM. Patch merging and expanding are used to downsample and upsample the features to learn hierarchical representations. The feature dimensions at different levels are annotated in the figure.

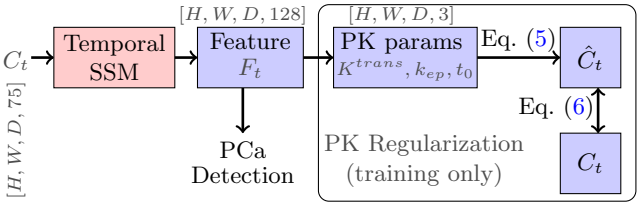

Figure 5: Computation flow of the temporal SSM. The temporal features F_t extracted from input TCCs C_t are used for (1) prostate cancer detection and (2) predicting PK parameters, i.e. K^{trans}, k_{ep} and t_0. The PK parameter estimation occurs exclusively during training and serves as a regularization mechanism for the temporal SSM.

spatial) SSM to the corresponding spatial (or temporal) SSM block. A linear layer aligns the feature dimensions, while patch merging (or expansion) ensures consistency in spatial size. To improve flexibility and enable inference without DCE-MRI TCCs, we randomly drop the temporal TCC input and disable interaction during training with a probability of 0.2. This trains the model to handle both spatial-only and spatiotemporal inputs at inference time.

4 Experiments and Results

We validate the performance of PCa-Mamba and the effect of the proposed components through extensive experiments. In particular, we investigate the following questions:

1. Can PCa-Mamba improve the performance of csPCa detection over other models? (Fig. 9 and 10 and Tab. 2)
2. Are there benefits to including DCE-MRI along with bpMRI for csPCa lesion detection? (Fig. 9, 10 and 13)
3. Do the proposed components, i.e. permuted sequentialization, PK regularization, and interactive fusion, improve the performance of PCa-Mamba? (Tab. 4)

We first introduce the MRI data in Sec. 4.1 and experimental details Sec. 4.2, and then report the experimental results in the rest of the section.

4.1 Datasets and Data Preprocessing

Our experiments are conducted on two datasets: our in-house dataset and the public PI-CAI dataset [42].

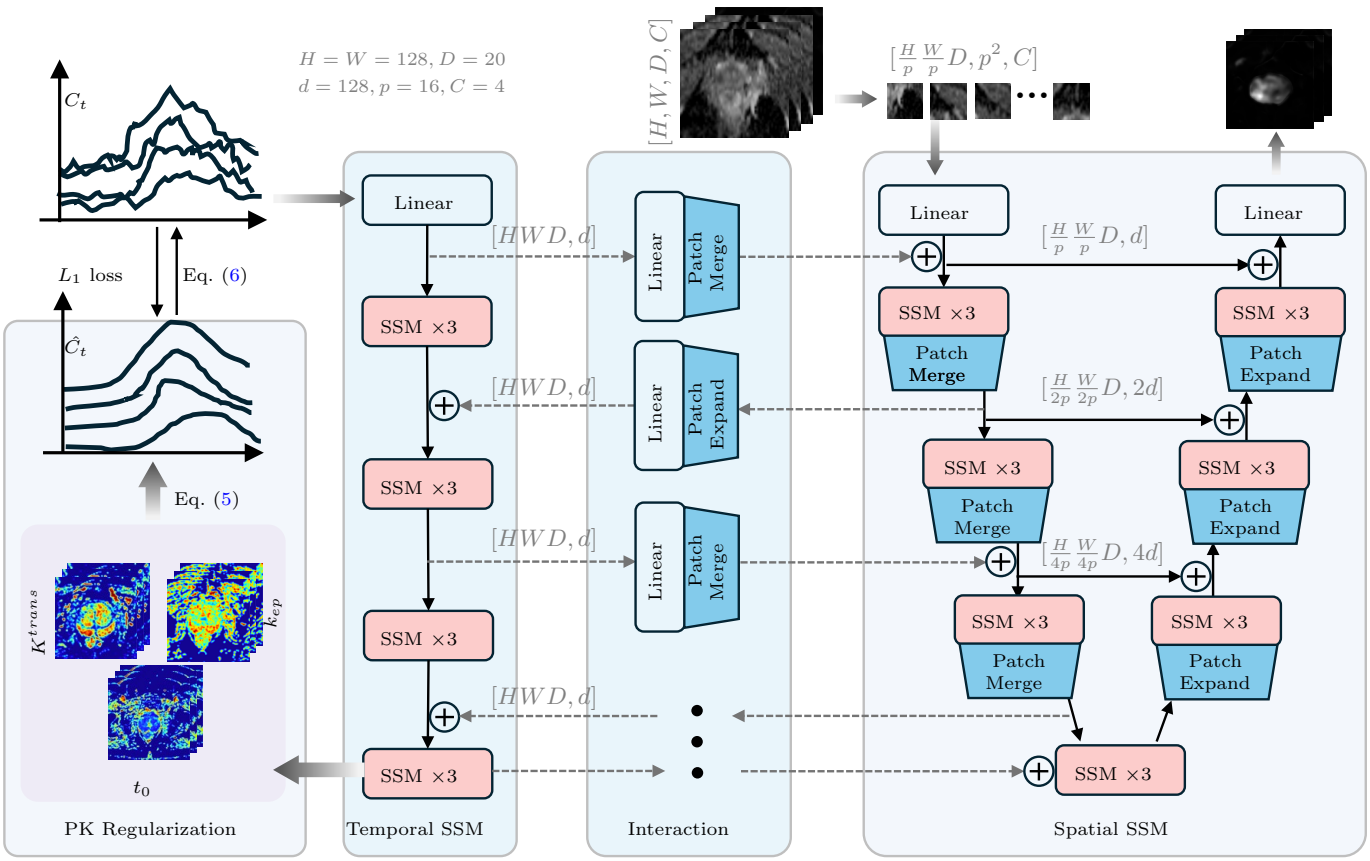

Figure 6: The architecture of spatiotemporal SSMs. The spatial SSM takes multi-slice spatial MRI images as input and extracts hierarchical features using a U-shaped encoder-decoder network. The temporal SSM takes DCE-MRI TCCs as input, and the extracted features are used for both csPCa detection and PK parameter estimation. The features of spatial and temporal SSMs interact to enrich the spatiotemporal context.

Table 1: Overview of the datasets used in our experiments.

Dataset	Scan	Mag	Resolution (mm)	Slice Spacing (mm)	Matrix Size
In-house	T2	3T	0.625	3.6	320 × 320
	ADC		2.0	3.6	128 × 128
	DCE		1.625	3.6	160 × 160
PI-CAI	T2	3T/1.5T	0.3	3.6	640 × 640
	ADC		2.0	3.6	128 × 120

4.1.1 PI-CAI dataset

The public release of the PI-CAI dataset consists of mpMRI scans of 1,500 patients including T2, DWI with b-values 50, 400, and 800 s/mm^2, ADC map (calculated from the b-values). All scans were acquired on Siemens MRI systems operating at 1.5 T or 3 T.

Among these cases, 220 positive cases have voxel-level lesion annotations provided by human experts, while 1,075 cases are negative (i.e., with empty lesion labels). We therefore used a total of 1,295 cases in our experiments. A subset of 204 cases, originating from the ProstateX dataset, additionally includes K^{trans} maps derived from DCE-MRI. Since we randomly drop channels during training, we include cases without K^{trans} maps in the training/testing process. Further details of the PI-CAI dataset are available on the official website[1].

[1] https://pi-cai.grand-challenge.org/DATA/

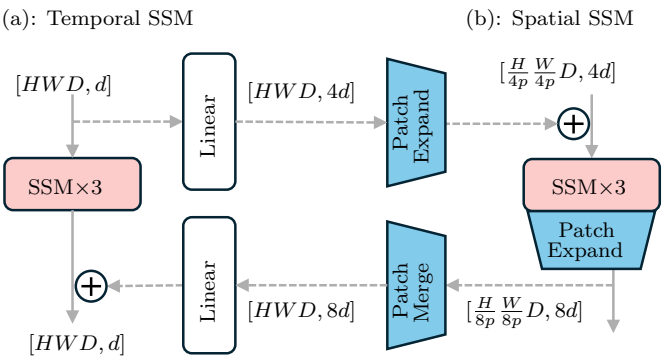

Figure 7: The interactions between spatial and temporal SSMs. The features from the one SSM are first aligned using patch merging (or expansion) and linear projection to match the feature dimensions of the target SSM. The patch merging (or expansion) ensures that the spatial size is maintained, and the linear projection aligns the feature dimensions. The fusion is optional and is applied only when DCE-MRI TCCs are available.

4.1.2 In-house Dataset

Dataset and data acquisition: Our in-house dataset consists of 652 patients with mpMRI, including: T2, DWI with b-values 0/100/400/800 s/mm^2, and DCE-MRI T1 images. All images are performed on Siemens 3T scanners with the standardized clinical prostate mpMRI protocol [31, 2]. The patients are composed of two parts: (1) pre-operative mpMRI images from 220 patients who had totally 246 confirmed PCa lesions with whole-mount histopathology after radical prostatectomy, and (2) mpMRI images from 432 patients who did not have indications of PCa lesions, confirmed by systematic biopsies followed by negative mpMRI (PI-RADS$\leq$2). We include mpMRI images with no indications of PCa lesions to balance the data distribution on model training and testing, as well as to perform patient-level classification evaluation.

We calculate the apparent diffusion coefficient (ADC) maps using linear least squares curve fitting of four DWIs against their corresponding b values (0/100/400/800 s/mm^2). We also include the DWI images with b=1400 s/mm2, referred to as high-b value DWI (or HBV for short). We compute zonal masks using the T2 with a pretrained zonal segmentation model [40].

Clinical interpretation and annotations: mpMRI images were reviewed by three genitourinary (GU) radiologists, each with over 10 years of clinical experience in prostate MRI interpretation, as part of routine clinical practice in accordance with established clinical guidelines [31, 2]. Lesions with a PI-RADS score $\geq$ 3 were considered MRI-positive findings, whereas lesions with PI-RADS score $<$ 3 were considered MRI-negative findings in this study.

The ground truth for lesion annotations was established using whole-mount histopathology (WMHP) obtained after radical prostatectomy (RP), spatially matched to the preoperative mpMRI. Blinded to all MRI-related information, the sliced WMHP specimens were independently examined and reported by three GU pathologists with 14, 8, and 5 years of experience in clinical prostate histopathology interpretation, as part of standard clinical workflow. Each PCa lesion was contoured and assigned a corresponding Gleason Score (GS) based on WMHP. Lesions with GS $\geq$ 7 were defined as csPCa and served as the detection targets in this study.

Under the supervision of attending GU radiologists, GU radiology research fellows retrospectively reviewed each mpMRI examination and manually contoured regions of interest (ROIs) for MRI-visible lesions on T2-weighted images, with reference to the WMHP examination reports. MRI-positive findings were categorized as true positives if radiological findings corresponded to histopathological lesions, and as false positives if no corresponding prostate cancer was identified in histopathology. Prospectively missed lesions that were retrospectively identified during the rereview process were defined as false-negative (FN) lesions. Prostate cancer lesions that were not visible on mpMRI and could not be retrospectively identified were excluded from the study, as reliable contouring of these lesions was not feasible.

The retrospective study on the in-house dataset is in compliance with the United States Health Insurance Portability and Accountability Act (HIPAA) of 1996 with approval from the institutional review board (IRB) of the University of California, Los Angeles with a waiver of the requirement for informed consent. All experiments adhere strictly to the relevant guidelines and regulations.

4.1.3 Preprocessing

All the MRI images are in a DICOM format. The DCE-MRI TCCs are computed from T1 images using the same method described in [17]. The dimensions of bpMRI images, including T2, ADC, HBV, and zonal masks, are $H \times W \times D$, where $H \times W$ are height and width, and D is the number of slices. TCCs are with dimension of $H \times W \times D \times T$ where $T = 75$ is the number of time steps.

4.2 Implementation and Evaluation Details

We implement our models with the PyTorch framework. We follow the training recipe of Vision Mamba [38]. Our model is trained for 20 epochs with a single NVIDIA 3090 GPU. Random left-right flipping is performed for data augmentation.

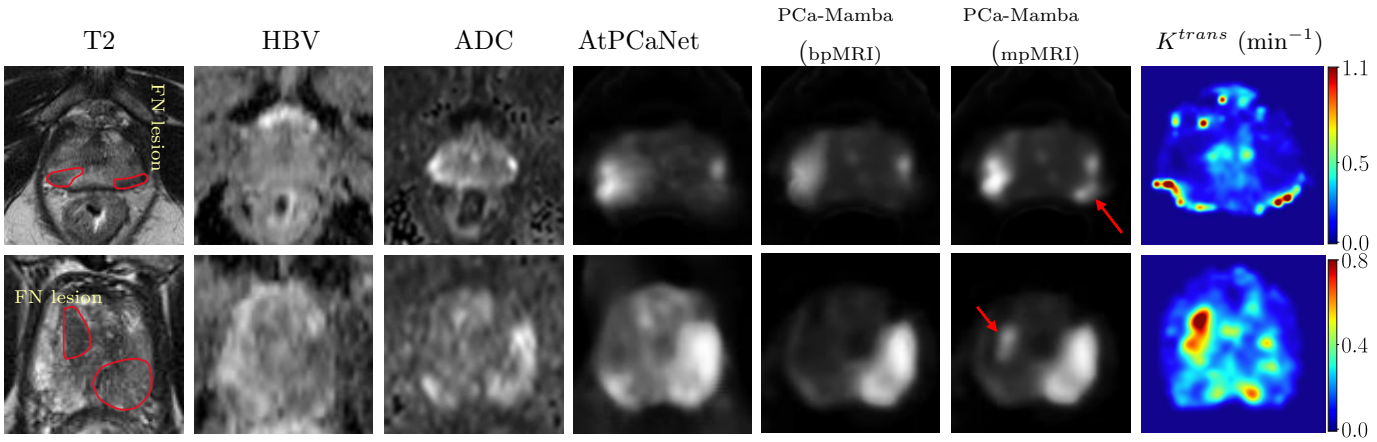

Figure 8: Examples of the csPCa detection results on our in-house dataset. The PCa-Mamba model with bpMRI produces less noisy results than the current state-of-the-art AtPCaNet [10], because of its broader context. Using PCa-Mamba with mpMRI, which includes DCE-MRI, enhances sensitivity by detecting MRI false negative (FN) lesions that are not visible in bpMRI (highlighted by text and red arrows). The K^{trans} maps are shown in the last column as a reference.

4.2.1 Other csPCa Detection Methods

We compare our proposed method with various csPCa detection methods and network architectures, including AtPCaNet [10], Pellicer et al [7], VoxResNet [43], VNet [44], nnUNet [45], UNeTR [46], and SEResUNet [47]. AtPCaNet [10] is a bpMRI PCa detection method with anatomical designs in network architecture and loss function. It takes bpMRI including T2, ADC, HBV, and zonal masks as input. Pellicer et al [7] is a mpMRI csPCa detection method that takes bpMRI and K^{trans} maps as input. When comparing different network architectures, we use the same input images and loss function as AtPCaNet [10] for fair comparisons.

4.2.2 Quantitative Evaluation and Statistical Testing

On each dataset, we perform 5-fold cross-validation for quantitative performance evaluation. In each fold, four-fifths of cases are used for model training with the others for testing. Following the common practice of many previous studies [48, 9, 10, 7], we evaluate lesion-level csPCa lesion detection using free-response receiver operating characteristic (FROC) analysis. The FROC curve characterizes the relationship between model detection sensitivity (recall) and the level of false positive predictions per patient. A detection point is defined as a true positive (TP) when the point is within 5 mm of a ground truth annotations to account for a potential mismatch between histopathology and radiology ROI [48, 9, 10]. The sensitivity results at each number of FP predictions per patient are compared using the Chi-squared Test, in accordance with 95% confidence interval (95% CI), correspondingly. In addition to FROC, we also report precision-recall (PR) curves that characterize the relationship between precision and recall.

We also report per-patient binary classification results using FROC and PR curves. A patient is labeled positive if they have at least one csPCa lesion, and negative otherwise. The patient-level prediction is defined as the maximum probability across all predicted volumes.

4.3 Comparison With the State-of-the-arts

We first compared PCa-Mamba with other state-of-the-art models for both lesion detection and patient-level classification.

We compared our method with various bpMRI csPCa detection methods, including AtPCaNet [10], VoxResNet [43], VNet [44], nnUNet [45], UNeTR [46], and SEResUNet [47]. We also compared with UMamba [49], a recent SSM-based architecture for medical image segmentation. All these methods used bpMRI data, including T2, ADC, HBV, and zonal mask as input. We also compared our method with two mpMRI methods [7, 8] that underutilized mpMRI data. To replicate [7], we used PCa-Mamba with only spatial SSM, accepting bpMRI and K^{trans}, referred to as "PCa-Mamba (bpMRI + K^{trans})." For the replication of [8], we used PCa-Mamba with only spatial SSM, accepting bpMRI, along with early and late T1 images, referred to as "PCa-Mamba (bpMRI + early + late)." We tested the full PCa-Mamba with spatiotemporal SSMs on our in-house dataset.

4.3.1 Lesion-level Detection Results

Fig. 8 showcases example detection results of the state-of-the-art AtPCaNet [10] and PCa-Mamba. The two cases in Fig. 8 are from our in-house dataset. The mean FROC curves for the two datasets are displayed in Fig. 9 and 10,

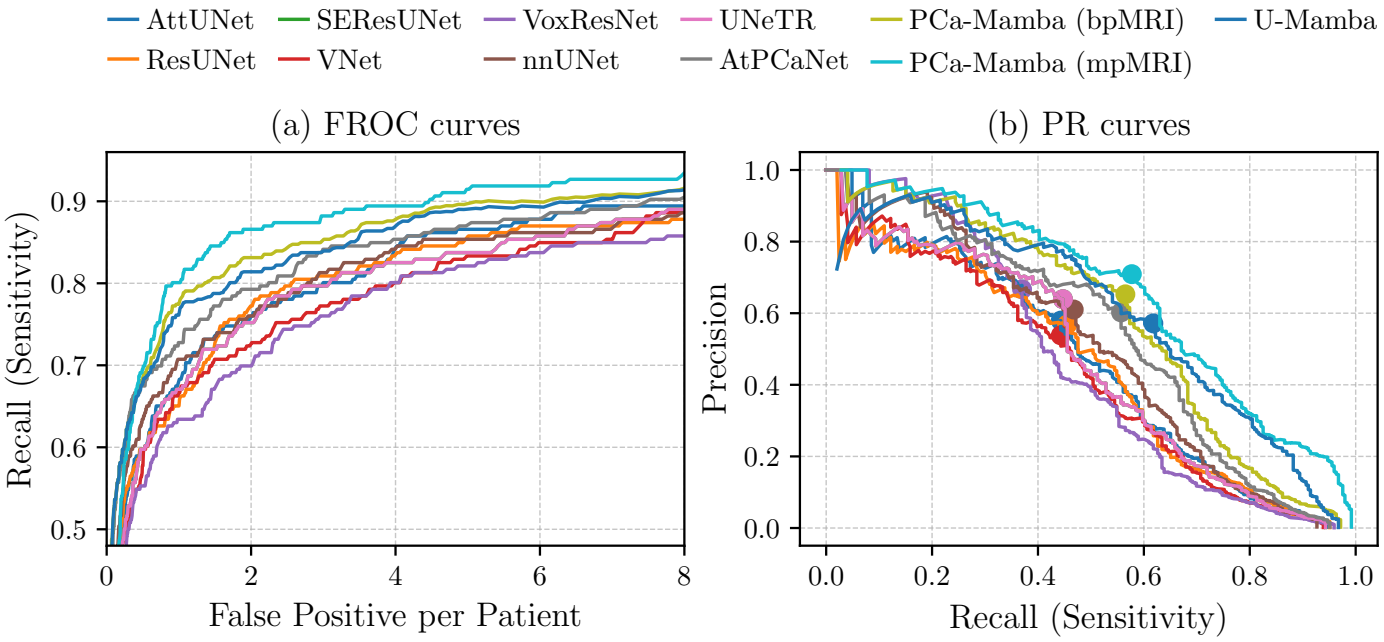

Figure 9: Lesion detection performance on our in-house dataset. The circles on PR curves indicate the point with best F1 score. PCa-Mamba with mpMRI achieves the best performance, demonstrating the effectiveness of PCa-Mamba in incorporating temporal dynamics in DCE-MRI.

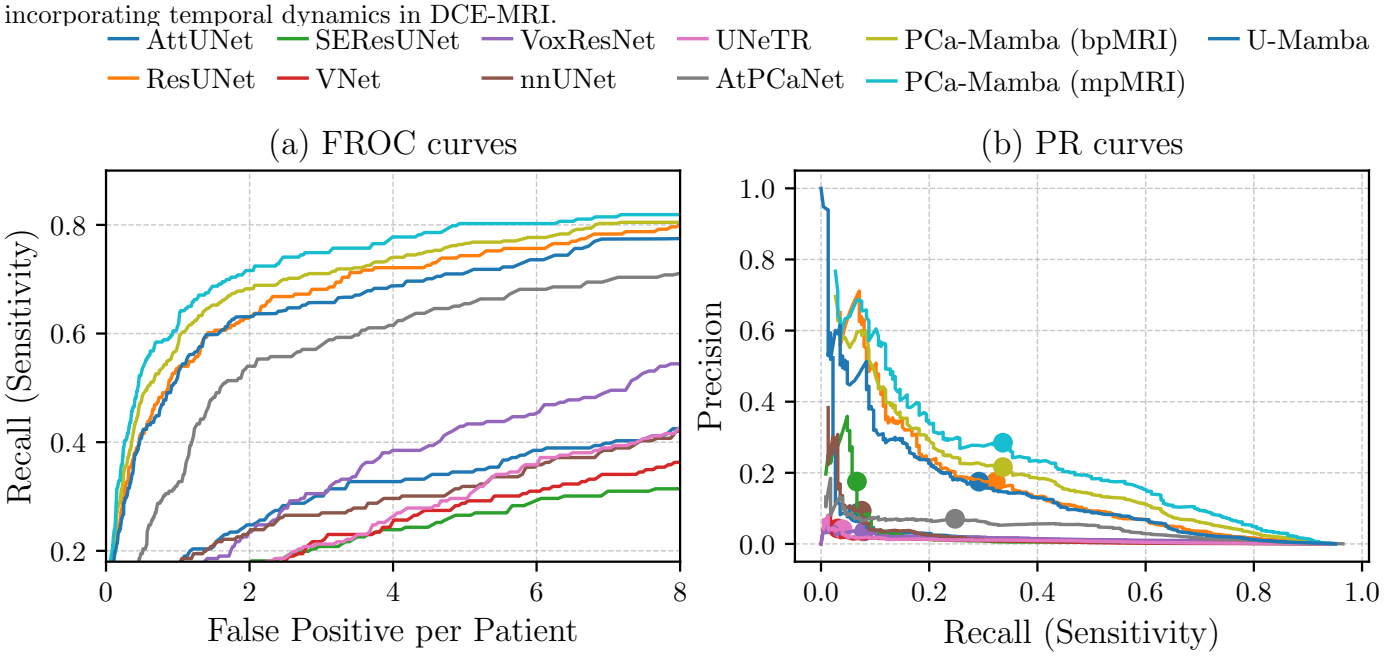

Figure 10: Lesion detection performance on the PI-CAI dataset. The circles on PR curves indicate the point with best F1 score. PCa-Mamba with mpMRI achieves the best performance, demonstrating the effectiveness of PCa-Mamba in incorporating temporal dynamics in DCE-MRI.

and the mean and variance of sensitivity values at various false positive rates are summarized in Tab. 2, where the sensitivity values were analyzed using the Chi-squared test with a 95% confidence interval (95% CI).

The FROC curves in Fig. 9 and 10 clearly show the benefit of including DCE-MRI and the superiority of our method over existing methods. As shown in Fig. 9 and 10, incorporating DCE-MRI leads to substantial improvements over bpMRI. Using DCE-MRI TCCs achieves the best performance, which reveals that DCE-MRI helps identify lesions that are invisible to other bpMRI methods. As illustrated in Fig. 9, including DCE-MRI TCCs

Table 2: Mean and standard deviation of sensitivity at different false positives per patient. The sensitivity results at each number of FP per patient were compared using the Chi-squared Test in accordance with a 95% confidence interval.

Models	Input	Sensitivity@#False positives per patient (95%CI)							
		In-house dataset				PI-CAI dataset			
		1	2	4	8	1	2	4	8
AttUNet	bp	$0.681_{\pm 0.070}$	$0.760_{\pm 0.051}$	$0.843_{\pm 0.057}$	$0.894_{\pm 0.064}$	$0.173_{\pm 0.079}$	$0.248_{\pm 0.080}$	$0.327_{\pm 0.073}$	$0.427_{\pm 0.077}$
ResUNet	bp	$0.650_{\pm 0.071}$	$0.772_{\pm 0.061}$	$0.835_{\pm 0.078}$	$0.878_{\pm 0.051}$	$0.535_{\pm 0.070}$	$0.630_{\pm 0.054}$	$0.721_{\pm 0.055}$	$0.800_{\pm 0.076}$
SEResUNet	bp	$0.671_{\pm 0.072}$	$0.752_{\pm 0.072}$	$0.825_{\pm 0.059}$	$0.890_{\pm 0.067}$	$0.146_{\pm 0.056}$	$0.178_{\pm 0.075}$	$0.239_{\pm 0.074}$	$0.314_{\pm 0.078}$
VNet	bp	$0.667_{\pm 0.080}$	$0.726_{\pm 0.075}$	$0.801_{\pm 0.057}$	$0.886_{\pm 0.064}$	$0.084_{\pm 0.060}$	$0.150_{\pm 0.066}$	$0.254_{\pm 0.064}$	$0.363_{\pm 0.076}$
VoxResNet	bp	$0.634_{\pm 0.056}$	$0.699_{\pm 0.080}$	$0.801_{\pm 0.076}$	$0.858_{\pm 0.062}$	$0.124_{\pm 0.064}$	$0.235_{\pm 0.075}$	$0.385_{\pm 0.070}$	$0.544_{\pm 0.054}$
nnUNet	bp	$0.707_{\pm 0.072}$	$0.760_{\pm 0.067}$	$0.842_{\pm 0.054}$	$0.886_{\pm 0.068}$	$0.177_{\pm 0.059}$	$0.236_{\pm 0.054}$	$0.296_{\pm 0.077}$	$0.420_{\pm 0.071}$
UNeTR	bp	$0.671_{\pm 0.080}$	$0.752_{\pm 0.077}$	$0.825_{\pm 0.075}$	$0.890_{\pm 0.055}$	$0.093_{\pm 0.063}$	$0.146_{\pm 0.060}$	$0.263_{\pm 0.070}$	$0.427_{\pm 0.064}$
AtPCaNet	bp	$0.728_{\pm 0.074}$	$0.793_{\pm 0.059}$	$0.854_{\pm 0.062}$	$0.905_{\pm 0.070}$	$0.323_{\pm 0.079}$	$0.540_{\pm 0.066}$	$0.615_{\pm 0.053}$	$0.711_{\pm 0.059}$
PCa-Mamba	bp	$0.775_{\pm 0.051}$	$0.831_{\pm 0.075}$	$0.878_{\pm 0.053}$	$0.916_{\pm 0.076}$	$0.578_{\pm 0.054}$	$0.682_{\pm 0.053}$	$0.740_{\pm 0.071}$	$0.805_{\pm 0.079}$
PCa-Mamba	mp	bpMRI + TCCs				bpMRI + K^{trans}			
		$0.806_{\pm 0.056}$	$0.867_{\pm 0.051}$	$0.896_{\pm 0.075}$	$0.937_{\pm 0.065}$	$0.614_{\pm 0.058}$	$0.716_{\pm 0.057}$	$0.778_{\pm 0.078}$	$0.819_{\pm 0.069}$

(PCa-Mamba with mpMRI) significantly outperforms the use of K^{trans} maps or early and late-phase T1 images. The results in Fig. 10 also demonstrated the superiority of SSMs over other architectures, such as Transformers and CNNs.

The FROC protocol defines a TP as a prediction within 5 mm of any ground truth annotation, but it does not quantify the actual distance to the ground truth. To further analyze localization errors, we compute the distance between a prediction and corresponding ground-truth. For each method, we apply the threshold that maximizes the F1 score to binarize the predictions, then compute the distance between each prediction and its nearest ground truth. The results are summarized in Tab. 3, which shows the mean and median distances, as well as the percentage of predictions within 3 mm and 5 mm of the ground truth. A prediction is considered a TP if it lies within 5 mm of a ground-truth lesion; otherwise, it is treated as a false positive. Under this criterion, "% $\leq$ 5 mm" corresponds to precision.

The results in Tab. 3 show that nnUNet outperforms conventional CNN-based methods such as AttUNet. Furthermore, SSM-based methods like PCa-Mamba and U-Mamba achieve superior performance compared to both nnUNet and AttUNet. Among all methods, PCa-Mamba with mpMRI achieves the lowest localization error and the highest percentage of predictions within 3 mm and 5 mm of the ground truth on both datasets.

4.3.2 Patient-level Classification Results

We also evaluate the patient-level binary classification performance. The prediction score of each patient is the maximum prediction score of the patient volume. Fig. 11 and 12 show the specificity-sensitivity and precision-recall curves for patient-level classification. The results demonstrate that incorporating DCE-MRI improves patient-level diagnostic performance, possibly because DCE-MRI helps identify lesions that are not visible in bpMRI. The results in Fig. 11 and 12 also demonstrate the superiority of SSM-based methods, e.g. UMamba [49] and PCa-Mamba, over CNN- or Transformer-based methods, e.g. AtPCaNet [10] and UNeTR [46].

Both lesion-level and patient-level results consistently demonstrate better performance on the in-house dataset than on the PI-CAI dataset, which is likely due to the higher imaging quality and more consistent annotation in our in-house dataset.

These results suggest improved csPCa detection performance on our in-house dataset compared with PI-CAI. We hypothesize that this may be due to differences in imaging acquisition and annotation. Specifically, our in-house data were acquired exclusively at 3 T, whereas PI-CAI includes both 1.5 T and 3 T scans, introducing lower image quality and greater variability. In addition, lesion annotations in our in-house dataset were guided by whole-mount histopathology, while PI-CAI annotations were based on MRI-guided biopsy, which may affect the accuracy of lesion delineation.

4.4 The Added Value of DCE-MRI

In this experiment, we verify the added value of the DCE-MRI in different lesions on our in-house dataset. In particular, we test PCa-Mamba with and without DCE-MRI as input. When testing with DCE-MRI, we employ the full PCa-Mamba architecture with spatiotemporal SSMs. When testing without DCE-MRI, we use only the spatial SSM which takes T2, ADC, HBV, and zonal masks as input. We explore the added value of DCE-MRI in different csPCa lesions. In particular, we evaluate the performance in the peripheral zone (PZ) and transition zone (TZ) le-

Table 3: Mean, median, and percentage of distances between predicted lesions and ground-truth annotations.

Data	Method	Mean (mm)	Median (mm)	% ≤ 3 mm	% ≤ 5 mm
in-house	AttUNet	4.77	4.38	41.82	57.94
	nnUNet	4.19	4.06	46.28	61.03
	U-Mamba	3.76	3.75	53.59	57.16
	PCaMb (bp)	3.56	3.13	58.79	65.27
	PCaMb (mp)	3.35	3.01	63.20	70.91
PI-CAI	AttUNet	8.01	7.92	6.97	8.62
	nnUNet	7.85	7.71	6.16	9.33
	U-Mamba	6.19	6.11	14.86	17.51
	PCaMb (bp)	5.66	5.51	17.15	21.60
	PCaMb (mp)	5.53	5.42	20.21	28.46

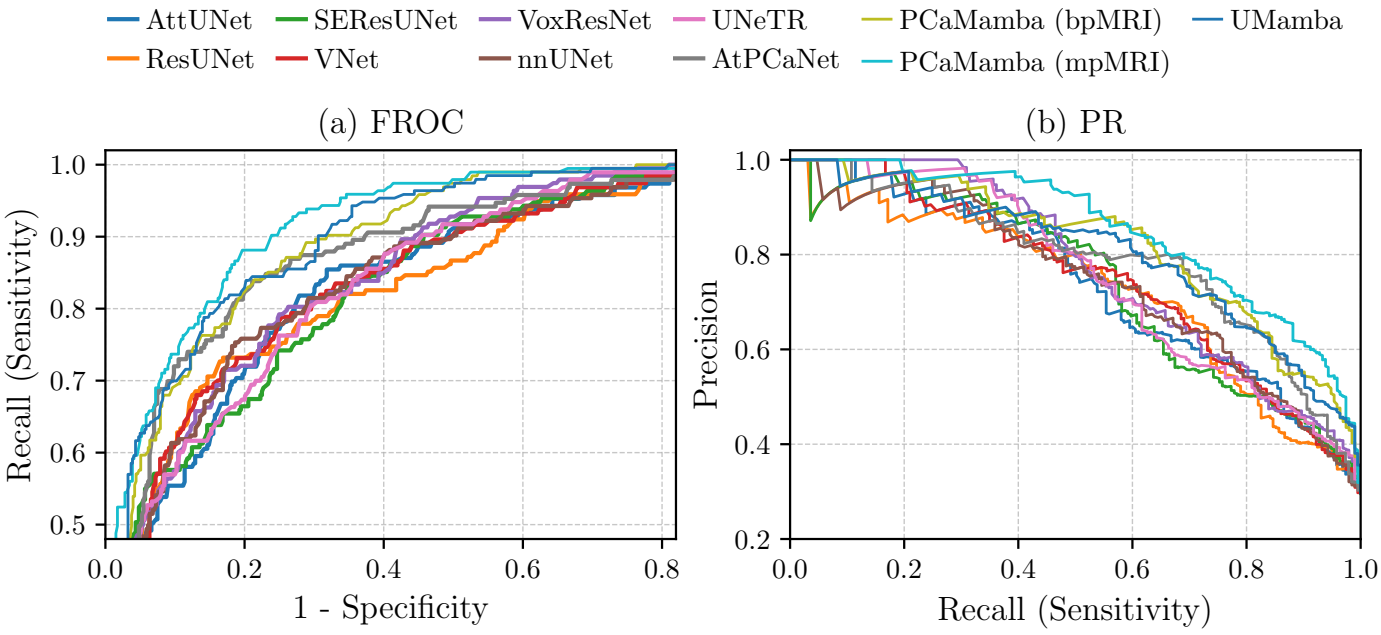

Figure 11: Per-patient classification results on our in-house dataset.

sions, lesions that are smaller and larger than 15 mm, as well as false negative (FN) and TP MRI lesions.

A lesion is assigned to the PZ or TZ based on spatial overlap with the corresponding zonal mask. Lesions whose segmentation masks overlap both zonal regions are categorized as belonging to both PZ and TZ. Lesions that were not detected during the initial prospective reading but were identified retrospectively during expert review are defined as false-negative (FN) lesions, whereas lesions detected prospectively are considered true-positive (TP). Our in-house dataset comprises 246 lesions in total, including 182 PZ lesions and 83 TZ lesions, as well as 120 small lesions and 126 large lesions. Among these, 57 lesions are classified as false negatives and 332 as true positives.

4.4.1 PCa lesions in PZ and TZ

First, we compare the performance of PCa-Mamba using bpMRI and mpMRI data on PZ and TZ lesions. As shown in Fig. 13 (a, b), the inclusion of DCE-MRI enhance detection of lesions in both the TZ and PZ, though the improvement is more significant in the PZ. This aligns with clinical observations that mpMRI is more successful in peripheral zone [50] and the impact of DCE-MRI on TZ lesions is less substantial [5].

4.4.2 Performance by lesion size

In this experiment, we examine the effect of DCE-MRI on lesions of varying sizes. Lesion size per slice is defined by the diameter of the largest enclosing circle, and overall

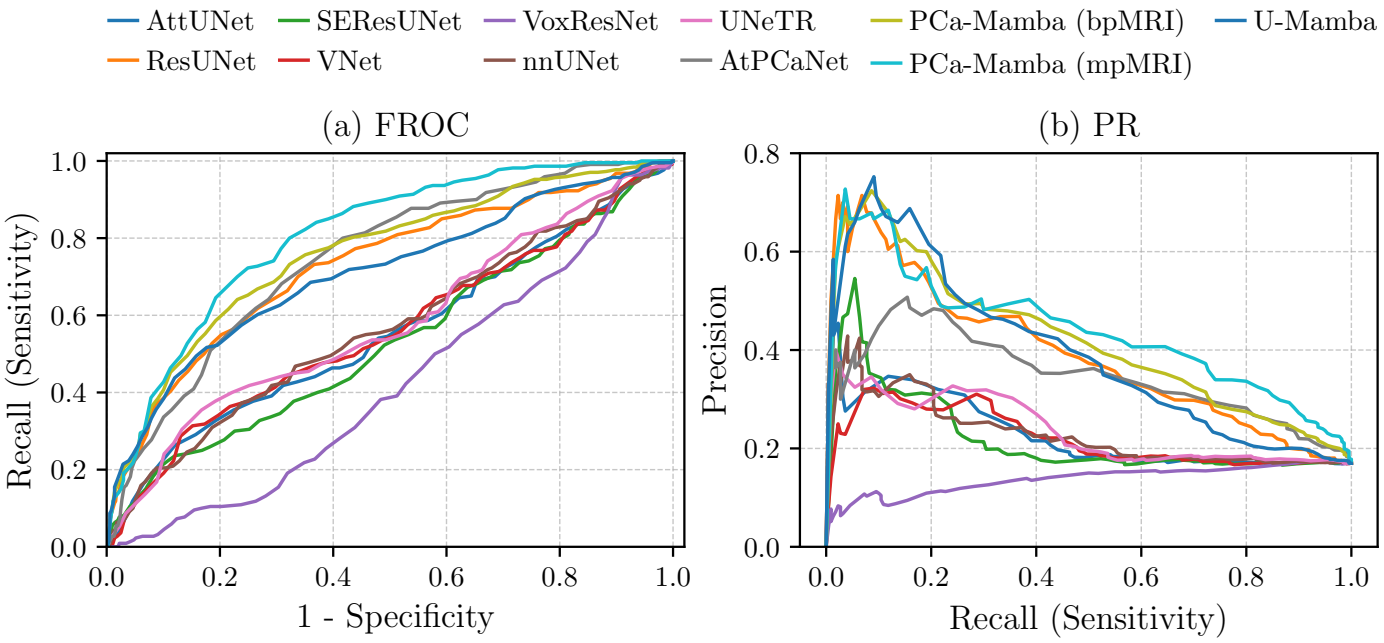

Figure 12: Per-patient classification results on PI-CAI dataset.

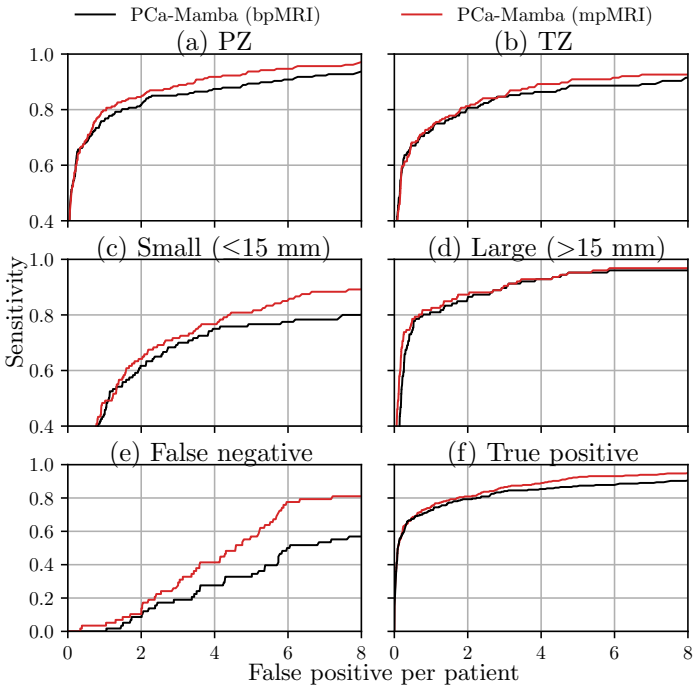

Figure 13: FROC curves of PCa-Mamba using bpMRI (T2+ADC+HBV+zonal mask) and mpMRI (bpMRI + TCCs) across peripheral (a) and transition zones (b), small (c) and large (d) lesions, as well as MRI false-negative (e) and true-positive (f) lesions.

lesion size by the maximum diameter across slices. Lesions are classified as small (≤15 mm) or large (>15 mm).

The results in Fig. 13 (b, c) show that, overall, detection accuracy is higher for larger lesions compared to smaller ones. The benefits of incorporating DCE-MRI are more evident in small lesions, primarily due to the saturated performance observed in large lesions, and the small lesions are sometimes invisible in T2 and DWI images. Similar results were also reported by another research [5].

4.4.3 MRI false negative lesions

As demonstrated in Fig. 9 and 10, one advantage of incorporating DCE-MRI is its enhanced sensitivity, particularly with a high rate of false positives. This indicates that mpMRI is capable of detecting lesions that are invisible in bpMRI. To validate this further, we evaluate the performance on MRI FN and TP lesions.

The results in Fig. 13 (e, f) indicate that including DCE-MRI enhances performance for both TP and FN lesions, with a more pronounced margin observed in FN lesions. Fig. 8 showcases two examples that DCE-MRI can help detect lesions that are invisible in bpMRI. This suggests that DCE-MRI could complement bpMRI, improving the visibility of csPCa lesions.

4.5 Ablation Study

We conduct ablation studies on the in-house dataset to investigate the effectiveness of key designs in PCa-Mamba.

4.5.1 Interactive Fusion

We first verify the effectiveness of interactive fusion by comparing different strategies to combine temporal DCE-MRI with bpMRI. In particular, we test the following four different approaches.

1. Data combination that directly integrates DCE TCCs as additional images alongside T2, ADC, HBV, and zonal masks. This results in input images with 79 channels (four spatial images and 75 TCC timesteps).
2. Parametric map combination, which is similar to [7],
3. Feature combination that merges temporal features with spatial. This method is equivalent to PCa-Mamba without interactive feature fusion.
4. The proposed interactive combination, as depicted in Sec. 3.4.

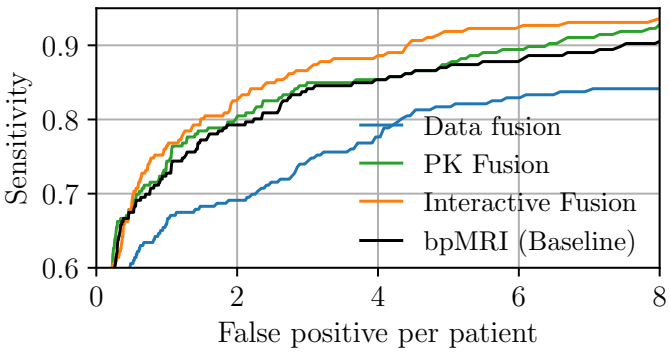

Figure 14: Comparison of various spatiotemporal feature fusion approaches.

Fig. 14 demonstrates the performance of various fusion methods, along with a baseline that excludes DCE-MRI for reference. Data fusion not only does not improve, but worsens the performance. This could be because concatenating high-dimensional DCE-MRI with other spatial images might lead to model overfitting. PK parametric map fusion results in a slight performance improvement, as PK parameters are indicative of csPCa. Our proposed interactive fusion outperforms feature fusion and achieves the best overall performance, because the interactions allow PCa-Mamba to learn more powerful features via dynamical feature fusion and mutual learning.

4.5.2 Regularization and Sequentialization

In addition, we ablated PK regularization and permuted sequentialization, and the results are summarized in Tab. 4 and 5. PCa-Mamba with mpMRI data, excluding PK regularization (PK Reg in Tab. 4), performs even worse than PCa-Mamba with bpMRI data. This is because, without PK regularization, the temporal SSM can be prone to overfitting, which leads to poor performance. PK regularization not only constrains the feature space of the spatial SSM but also enables its features to encode the tissue's physiological information. The model's performance also slightly decreases without permuted sequentialization. Overall, the results in Tab. 4 demonstrate the positive impact of the proposed components in PCa-Mamba.

Table 4: Sensitivity of PCa-Mamba when excluding one of the proposed components.

Models	Sensitivity@#FP per patient (95%CI)			
	1	2	4	8
PCa-Mamba (mpMRI)	0.806	0.867	0.896	0.937
w/o Per Seq	0.787	0.849	0.871	0.921
w/o PK Reg	0.742	0.799	0.825	0.885
PCa-Mamba (bpMRI)	0.775	0.831	0.878	0.916

Table 5: Performance of PCa-Mamba (bpMRI) with different sequentialization directions.

Direction	Sensitivity@#FP per patient (95%CI)			
	1	2	4	8
Vertical	0.763	0.819	0.864	0.913
Horizontal	0.759	0.818	0.860	0.911
Depth	0.725	0.799	0.848	0.896
Ver+Hor+Dep	0.775	0.831	0.878	0.916

The ablation study in Tab. 5 evaluates PCa-Mamba with different sequentialization directions. We test PCa-Mamba with bpMRI data the sequentialization is only performed on bpMRI data. Performance varies significantly by direction: in-plane horizontal or vertical sequentialization performs best among single directions, while sequentialization along the depth axis performs worst. Combining all three directions yields the highest overall performance. The results suggest that the inplane images provide more informative features for csPCa detection than the depth images. The results in Tab. 5 demonstrate that in-plane images provide richer and more discriminative features for csPCa detection than depth images. This advantage stems from their higher spatial resolution, lower noise level, and clearer depiction of anatomical structures—consistent with clinical practice, where radiologists primarily rely on in-plane views. Furthermore, incorporating through-plane images introduces complementary spatial information, leading to improved overall model performance when combined with in-plane data.

4.5.3 Efficiency Comparisons with Other Sequence Models

We select state space models (SSMs) as the backbone for PCa-Mamba due to their efficiency, particularly in handling the high spatial and temporal dimensionality of mpMRI data. To evaluate PCa-Mamba's efficiency in modeling long-range dependencies, we compare its performance, inference time, and GPU memory usage against other sequence models, including the Vision Transformer (ViT) [51], Gated Recurrent Unit (GRU) [52]. GPU memory and inference time were measured based on the usage and latency for processing a single patient volume.

We conduct comparisons on both bpMRI and mpMRI data using the spatial and spatial-temporal variants of PCa-Mamba. The spatial-temporal evaluations are performed on our in-house dataset, while the spatial evaluations were conducted on the public PI-CAI dataset.

For the spatial PCa-Mamba, we only use the spatial SSM, replace the bidirectional SSM in spatial SSM with alternative sequence models, including GRU and Transformer. For the spatial-temporal PCa-Mamba, we replace the bidirectional SSM with GRU and Transformer. For all the experiments in this section, we only replace the SSM (highlighted in pink in Fig. 6) with the alternative sequence models, e.g. GRU and Transformer, for fair comparisons.

It is worth noting that training mpMRI models with Transformers is very memory-consuming. Therefore all the experiments in this section are conducted on an H100 GPU with 80GB of memory.

Results in Tab. 6 show that linear-complexity models like SSMs and GRUs are significantly faster and more memory-efficient than quadratic-complexity models such as Transformers. Among them, SSMs outperform GRUs in modeling long-range dependencies in volumetric MRI, offering a better balance between performance and efficiency. This makes SSMs well-suited for csPCa detection in spatiotemporal mpMRI, supporting their use in PCa-Mamba.

4.5.4 The effectiveness of DCE-MRI dropping

Our framework randomly drops DCE-MRI during training to improve robustness to missing sequences at inference. This strategy enables the model to operate with or without DCE-MRI and even boosts performance when it is available. Treated as a form of data augmentation, the random dropping mitigates overfitting and enhances generalization.

The results in Tab. 7 demonstrate that training with DCE-MRI dropping improves performance when DCE-MRI is available at inference, and barely affects performance when DCE-MRI is absent.

4.5.5 Different Pharmacokinetic Models

The Tofts model used in Eq. (6) to regularize the temporal SSM assumes a well-mixed compartment and neglects the intravascular contrast contribution. To verify the generalizability of our framework to other PK models, we tested our method using different PK models for regularization.

We investigated three PK models: the original MRDI model [32], and two its variants [17, 18]. We tested the csPCa detection performance as well as the average L1 loss on the test dataset using the Tofts mode and the two variants for PK regularization.

The results in Tab. 8 show that the three alternative pharmacokinetic (PK) models yield lower L1 fitting errors than the standard Tofts model, indicating improved capability in modeling DCE-MRI time–concentration curves (TCCs). However, with respect to clinically significant prostate cancer (csPCa) detection performance, the model in [18] only marginally outperforms the Tofts model, while the approaches in [17] and [32] exhibit slightly inferior results. Paired t-tests conducted across four operating points (false alarm rates ranging from 10^{-6} to 10^{-9}) revealed no statistically significant differences among these three models (all $p > 0.05$).

This degraded performance of MRDI-based PK models can be attributed to the role of the PK model in our framework. Specifically, the PK model is employed solely as a regularization term during training, rather than being directly used for csPCa classification. While more expressive PK models can better fit DCE-MRI TCCs, their increased complexity and non-convexity introduce additional optimization challenges during training. As a result, these models may hinder stable convergence and limit the effectiveness of PK-based regularization, ultimately leading to slightly degraded csPCa detection performance.

4.5.6 Interpretability Study Using Grad-CAM

To understand how the model makes decisions for csPCa detection, we employed Grad-CAM [53] to reveal the specific regions that PCa-Mamba attends to during inference. Grad-CAM computes the contribution of each feature to a scalar output. Since PCa-Mamba produces voxel-wise output, we first averaged the outputs over the entire volume to obtain a single scalar on which Grad-CAM can be computed. The averaged output can be interpreted as the probability of csPCa for the patient. Grad-CAM maps were then generated for both the spatial and temporal branches, using the feature maps immediately before fusion. The resulting Grad-CAM maps were interpolated to match the resolution of the original images. We use $\mathcal{G}_s\mathbb{R}^{H\times W\times D}$ to denote the Grad-CAM from the spatial branch, and $\mathcal{G}_t \in \mathbb{R}^{H\times W\times D\times T}$ to denote the Grad-CAM from the temporal branch. We average $\mathcal{G}_t$ over the spa-

Table 6: Efficiency comparisons of PCa-Mamba with Transformer and GRU.

Data	Seq Model	Sensitivity@#FP per patient (95%CI) 1	2	4	GPU Mem (Gb)	Time (Sec)
mpMRI (in-house)	SSM	0.801	0.866	0.894	9.8	5.6
	GRU	0.731	0.812	0.853	10.4	18.4
	Transformer	0.7917	0.856	0.896	21.1	46.1
	Swin Transformer	0.7817	0.840	0.881	16.5	37.2
	UNeTR	0.795	0.859	0.889	22.7	50.8
bpMRI (PI-CAI)	SSM	0.614	0.716	0.778	2.1	3.1
	GRU	0.587	0.674	0.725	2.9	4.6
	Transformer	0.615	0.712	0.779	7.7	11.4
	Swin Transformer	0.605	0.703	0.770	6.5	10.2
	UNeTR	0.617	0.718	0.785	8.1	12.0

Table 7: Performance of PCa-Mamba with and without DCE-MRI dropping during training. The mode is trained on our in-house dataset, and the performance is evaluated with both mpMRI and bpMRI at inference.

Training	Test data	Sensitivity@#FP (95%CI) 1	2	4	8
w drop	bp	0.775	0.831	0.878	0.916
	mp	0.806	0.867	0.896	0.937
w/o drop	bp	0.775	0.832	0.879	0.916
	mp	0.801	0.866	0.894	0.935

Table 8: csPCa detection performance and average L1 loss of using different regularization pharmacokinetic models.

PK Model	Avg L1 Loss	Sensitivity@#False positives per patient 1	2	4	8
Tofts (default)	0.8291	$0.801_{\pm 0.059}$	$0.866_{\pm 0.050}$	$0.894_{\pm 0.075}$	$0.935_{\pm 0.065}$
[32]	0.7801	$0.793_{\pm 0.051}$	$0.851_{\pm 0.052}$	$0.884_{\pm 0.070}$	$0.931_{\pm 0.068}$
[17]	0.7897	$0.795_{\pm 0.062}$	$0.855_{\pm 0.077}$	$0.888_{\pm 0.072}$	$0.931_{\pm 0.081}$
[18]	0.7535	$0.802_{\pm 0.047}$	$0.867_{\pm 0.055}$	$0.896_{\pm 0.060}$	$0.935_{\pm 0.071}$

tial dimensions (H and W) to obtain the temporal Grad-CAMs $\mathcal{G}_{t2} \in \mathbb{R}^T$, and over the temporal dimension (T) to obtain the spatial Grad-CAM $\mathcal{G}_{t1} \in \mathbb{R}^{H \times W \times D}$ of the temporal branch. Fig. 15 illustrates three representative cases from our in-house dataset, showing the histopathological image, T2-weighted image with lesion annotations, Grad-CAM maps from the spatial and temporal branches ($\mathcal{G}_s$, $\mathcal{G}_{t1}$ and $\mathcal{G}_{t2}$), and the predicted csPCa probability maps generated by PCa-Mamba using both bpMRI and mpMRI. As we can see, the spatial Grad-CAMs ($\mathcal{G}_s$ and $\mathcal{G}_{t1}$) highlight different regions of interest in the prostate, and complement each other to provide a more comprehensive understanding of the lesion locations. The temporal Grad-CAMs ($\mathcal{G}_{t2}$) emphasize the peak enhancement phase, indicating that this phase is particularly informative for csPCa detection.

The Grad-CAMs in Fig. 15 show that the spatial and temporal branches contribute uniquely and complementarily to the final csPCa prediction— some lesions are identified by spatial features, while others rely on temporal cues, as highlighted by the red circles. The prediction maps further support this observation: PCa-Mamba de-

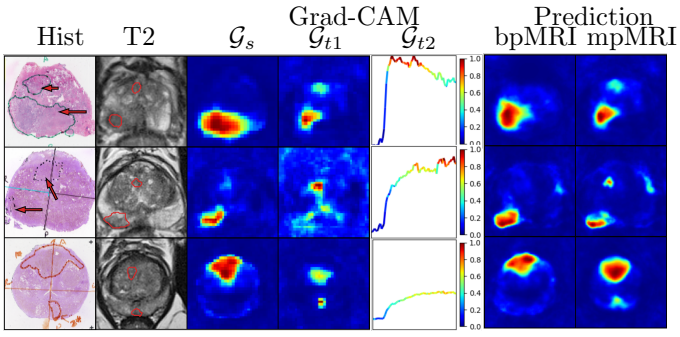

Figure 15: Histopathological image, T2-weighted image with lesion annotations, Grad-CAM maps from the spatial and temporal branches ($\mathcal{G}_s$ and $\mathcal{G}_t$), and the predicted csPCa probability maps generated by PCa-Mamba using both bpMRI and mpMRI.

tects more lesions when using mpMRI (with DCE), but misses some when using only bpMRI, underscoring the critical role of DCE-MRI in capturing temporal information essential for identifying certain lesions.

5 Discussion

5.1 Principal Findings and Interpretation

This study presents PCa-Mamba, a spatiotemporal SSM framework designed to integrate spatial information from bpMRI with temporal enhancement dynamics from DCE-MRI for csPCa detection. Across experiments on both an in-house dataset and the public PI-CAI dataset, the results demonstrate that explicitly modeling temporal enhancement patterns provides complementary information beyond conventional bpMRI-based approaches. The observed improvements are particularly evident in challenging cases, including small lesions, peripheral zone lesions, and MRI-negative lesions, which are known to be difficult to detect using bpMRI alone. These findings suggest that effective integration of temporal DCE-MRI information can enhance the sensitivity of csPCa detection without sacrificing robustness across datasets.

5.2 Role of DCE-MRI and Spatiotemporal Modeling

The added value of DCE-MRI observed in this work is consistent with prior clinical and computational studies reporting that contrast enhancement dynamics reflect vascular properties associated with tumor angiogenesis. Unlike approaches that rely on early- or late-phase DCE images or precomputed pharmacokinetic parametric maps, PCa-Mamba directly learns from original DCE-MRI time–concentration curves (TCCs), allowing the model to exploit the full temporal evolution of contrast uptake.

The proposed spatiotemporal modeling strategy plays a critical role in this integration. SSMs capture long-range anatomical context across multi-slice bpMRI volumes, while temporal state-space models encode voxel-wise enhancement dynamics. The interactive fusion mechanism enables bidirectional information exchange between spatial and temporal representations, allowing temporal cues to be interpreted within their anatomical context and improving overall detection performance.

5.3 Pharmacokinetic Regularization

PK regularization serves as a key component for stabilizing temporal feature learning from noisy and high-dimensional DCE-MRI TCCs. By constraining the temporal SSM to approximate physiologically meaningful PK behavior, the network is encouraged to learn representations that are both diagnostically relevant and physiologically plausible. However, our comparative experiments with alternative PK models reveal that improved curve-fitting accuracy does not necessarily translate into superior csPCa detection performance. This observation highlights an important distinction: within the proposed framework, PK models are used as a regularization mechanism rather than as direct diagnostic predictors. More complex and non-convex PK formulations may introduce additional optimization challenges, which can limit the effectiveness of regularization and ultimately affect downstream detection performance.

5.4 Clinical Implications, Limitations, and Future Work

From a clinical perspective, the proposed framework supports the growing evidence that DCE-MRI can provide meaningful diagnostic information when appropriately integrated with bpMRI. The improved detection of small and MRI-invisible lesions is particularly relevant for early diagnosis and risk stratification. Importantly, PCa-Mamba is designed to operate with or without DCE-MRI at inference time through modality dropout during training, enhancing its practical applicability in real-world clinical settings where DCE-MRI may be unavailable or omitted. Several limitations should be noted. First, temporal modeling operates independently at the voxel level, which may increase computational cost and limit spatial contextualization of temporal dynamics. Second, PK regularization relies on a population-based arterial input function, which may not fully capture inter-patient vascular variability. Future work will explore more spatially aware temporal modeling strategies, adaptive or patient-specific PK formulations, and prospective multicenter validation to further assess generalizability and clinical utility.

6 Conclusion

In this work, we proposed PCa-Mamba, a spatiotemporal state-space model framework for automated clinically significant prostate cancer (csPCa) detection from multiparametric MRI. The proposed approach jointly models spatial information from bpMRI and temporal enhancement dynamics from DCE-MRI, and integrates these complementary cues through interactive feature fusion within an end-to-end learning framework. Pharmacokinetic regularization based on the Tofts model was incorporated to impose physiologically motivated constraints on temporal feature learning and to improve training stability.

Comprehensive experiments on an in-house cohort and the public PI-CAI dataset demonstrate that PCa-Mamba effectively leverages spatiotemporal mpMRI information

for csPCa detection. These results support the potential of explicitly integrating DCE-MRI temporal dynamics with bpMRI for improving prostate cancer detection in challenging clinical scenarios. Future work will focus on further improving computational efficiency, enhancing spatiotemporal context modeling, and validating the proposed framework in larger, prospective, and multicenter clinical studies.

References

[1] Rebecca L Siegel, Kimberly D Miller, Nikita Sandeep Wagle, and Ahmedin Jemal. Cancer statistics, 2023. *Ca Cancer J Clin*, 73(1):17–48, 2023.

[2] B Turkbey, AB Rosenkrantz, and MA et al Haider. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *European Urology*, 76(3):340–351, 2019.

[3] Thomas E Yankeelov and John C Gore. Dynamic contrast enhanced magnetic resonance imaging in oncology: theory, data acquisition, analysis, and examples. *Current Medical Imaging*, 3(2):91–107, 2007.

[4] Nicolai Alexander Huebner, Stephan Korn, Irene Resch, Bernhard Grubmüller, Tobias Gross, Robert Gale, Gero Kramer, Nina Poetsch, Paola Clauser, Andrea Haitel, et al. Visibility of significant prostate cancer on multiparametric magnetic resonance imaging (mri)—do we still need contrast media? *European Radiology*, 31:3754–3764, 2021.

[5] Jiahui Zhang, Lili Xu, Gumuyang Zhang, Xiaoxiao Zhang, Xin Bai, Hao Sun, and Zhengyu Jin. Effects of dynamic contrast enhancement on transition zone prostate cancer in prostate imaging reporting and data system version 2.1. *Radiology and Oncology*, 57(1):42–50, 2023.

[6] Mehdi Taghipour, Alireza Ziaei, Francesco Alessandrino, Elmira Hassanzadeh, Mukesh Harisinghani, Mark Vangel, Clare M Tempany, and Fiona M Fennessy. Investigating the role of dce-mri, over t2 and dwi, in accurate pi-rads v2 assessment of clinically significant peripheral zone prostate lesions as defined at radical prostatectomy. *Abdominal radiology*, 44:1520–1527, 2019.

[7] Oscar J Pellicer-Valero, Jose L Marenco Jimenez, Victor Gonzalez-Perez, Juan Luis Casanova Ramon-Borja, Isabel Martín García, Maria Barrios Benito, Paula Pelechano Gomez, José Rubio-Briones, María José Rupérez, and José D Martín-Guerrero. Deep learning for fully automatic detection, segmentation, and gleason grade estimation of prostate cancer in multiparametric magnetic resonance images. *Scientific reports*, 12(1):2975, 2022.

[8] Yoh Matsuoka, Yoshihiko Ueno, Sho Uehara, Hiroshi Tanaka, Masaki Kobayashi, Hajime Tanaka, Soichiro Yoshida, Minato Yokoyama, Itsuo Kumazawa, and Yasuhisa Fujii. Deep-learning prostate cancer detection and segmentation on biparametric versus multiparametric magnetic resonance imaging: Added value of dynamic contrast-enhanced imaging. *International Journal of Urology*, 30(12):1103–1111, 2023.

[9] Ruiming Cao, Amirhossein Mohammadian Bajgiran, Sohrab Afshari Mirak, Sepideh Shakeri, Xinran Zhong, Dieter Enzmann, Steven Raman, and Kyunghyun Sung. Joint prostate cancer detection and gleason score prediction in mp-mri via focalnet. *IEEE transactions on medical imaging*, 38(11):2496–2506, 2019.

[10] Haoxin Zheng, Alex Ling Yu Hung, Qi Miao, Weinan Song, Fabien Scalzo, Steven S Raman, Kai Zhao, and Kyunghyun Sung. Atpca-net: anatomical-aware prostate cancer detection network on multiparametric mri. *Scientific Reports*, 14(1):5740, 2024.

[11] Coen De Vente, Pieter Vos, Matin Hosseinzadeh, Josien Pluim, and Mitko Veta. Deep learning regression for prostate cancer detection and grading in bi-parametric mri. *IEEE Transactions on Biomedical Engineering*, 68(2):374–383, 2020.

[12] Ahmet Karagoz, Deniz Alis, Mustafa Ege Seker, Gokberk Zeybel, Mert Yergin, Ilkay Oksuz, and Ercan Karaarslan. Anatomically guided self-adapting deep neural network for clinically significant prostate cancer detection on bi-parametric mri: a multi-center study. *Insights into Imaging*, 14(1):110, 2023.

[13] Audrey Duran, Gaspard Dussert, Olivier Rouvière, Tristan Jaouen, Pierre-Marc Jodoin, and Carole Lartizien. Prostattention-net: A deep attention model for prostate cancer segmentation by aggressiveness in mri scans. *Medical Image Analysis*, 77:102347, 2022.

[14] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[16] Paul S Tofts, Gunnar Brix, David L Buckley, Jeffrey L Evelhoch, Elizabeth Henderson, Michael V Knopp, Henrik BW Larsson, Ting-Yim Lee, Nina A Mayr, Geoffrey JM Parker, et al. Estimating kinetic parameters from dynamic contrast-enhanced t1-weighted mri of a diffusable tracer: standardized quantities and symbols. *Journal of Magnetic Resonance Imaging*, 10(3):223–232, 1999.

[17] Kyunghyun Sung. Modified mr dispersion imaging in prostate dynamic contrast-enhanced mri. *Journal of Magnetic Resonance Imaging*, 50(4):1307–1317, 2019.

[18] Kai Zhao, Kaifeng Pang, Alex LingYu Hung, Haoxin Zheng, Ran Yan, and Kyunghyun Sung. A deep learning-based framework for highly accelerated prostate mr dispersion imaging. *Cancers*, 16(17):2983, 2024.

[19] Eline K Vos, Geert JS Litjens, Thiele Kobus, Thomas Hambrock, Christina A Hulsbergen-van de Kaa, Jelle O Barentsz, Henkjan J Huisman, and Tom WJ Scheenen. Assessment of prostate cancer aggressiveness using dynamic contrast-enhanced magnetic resonance imaging at 3 t. *European urology*, 64(3):448–455, 2013.

[20] Yang Song, Yu-Dong Zhang, Xu Yan, Hui Liu, Minxiong Zhou, Bingwen Hu, and Guang Yang. Computer-aided diagnosis of prostate cancer using a deep convolutional neural network from multiparametric mri. *Journal of Magnetic Resonance Imaging*, 48(6):1570–1577, 2018.

[21] Atilla P Kiraly, Clement Abi Nader, Ahmet Tuysuzoglu, Robert Grimm, Berthold Kiefer, Noha El-Zehiry, and Ali Kamen. Deep convolutional encoder-decoders for prostate cancer detection and classification. In *International conference on MICCAI*, pages 489–497. Springer, 2017.

[22] Arun Seetharaman, Indrani Bhattacharya, Leo C. Chen, Christian A. Kunder, Wei Shao, Simon J. C. Soerensen, Jeffrey B. Wang, Nikola C. Teslovich, Richard E. Fan, Pejman Ghanouni, James D. Brooks, Katherine J. Too, Geoffrey A. Sonn, and Mirabela Rusu. Automated detection of aggressive and indolent prostate cancer on magnetic resonance imaging. *Medical Physics*, 48(6):2960–2972, 2021.

[23] Yuchun Li, Yuanyuan Wu, Mengxing Huang, Yu Zhang, and Zhiming Bai. Attention-guided multiscale learning network for automatic prostate and tumor segmentation on mri. *Computers in Biology and Medicine*, 165:107374, 2023.

[24] Matin Hosseinzadeh, Anindo Saha, Patrick Brand, Ilse Slootweg, Maarten de Rooij, and Henkjan Huisman. Deep learning–assisted prostate cancer detection on bi-parametric MRI: minimum training data size requirements and effect of prior knowledge. *European Radiologuy*, 32(4):2224–2234, 2022.

[25] Coen de Vente, Pieter Vos, Matin Hosseinzadeh, Josien Pluim, and Mitko Veta. Deep Learning Regression for Prostate Cancer Detection and Grading in Bi-Parametric MRI. *IEEE Transactions on Biomedical Engineering (TBME)*, 68(2):374–383, 2020.

[26] Audrey Duran, Gaspard Dussert, Olivier Rouvière, Tristan Jaouen, Pierre-Marc Jodoin, and Carole Lartizien. Prostattention-net: A deep attention model for prostate cancer segmentation by aggressiveness in mri scans. *Medical Image Analysis*, 77:102347, 2022.

[27] Xin Yu, Bin Lou, Bibo Shi, David Winkel, Nacim Arrahmane, Mamadou Diallo, Tongbai Meng, Heinrich von Busch, Robert Grimm, Berthold Kiefer, Dorin Comaniciu, Ali Kamen, Henkjan Huisman, Andrew Rosenkrantz, Tobias Penzkofer, Ivan Shabunin, Moon Hyung Choi, Qingsong Yang, and Dieter Szolar. False Positive Reduction Using Multiscale Contextual Features for Prostate Cancer Detection in Multi-Parametric MRI Scans. *IEEE ISBI*, pages 1355–1359, 2020.

[28] Anindo Saha, Matin Hosseinzadeh, and Henkjan Huisman. End-to-end prostate cancer detection in bpMRI via 3D CNNs: Effects of attention mechanisms, clinical priori and decoupled false positive reduction. *Medical Image Analysis*, 73:102155, 2021.

[29] Alex Ling Yu Hung, Haoxin Zheng, Kai Zhao, Kaifeng Pang, Demetri Terzopoulos, and Kyunghyun Sung. Cross-slice attention and evidential critical loss for uncertainty-aware prostate cancer detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 113–123. Springer, 2024.

[30] Sadhna Verma, Baris Turkbey, Naira Muradyan, Arumugam Rajesh, Francois Cornud, Masoom A Haider, Peter L Choyke, and Mukesh Harisinghani. Overview of dynamic contrast-enhanced mri in prostate cancer diagnosis and management. *American Journal of Roentgenology*, 198(6):1277–1288, 2012.

[31] Jeffrey C. Weinreb, Jelle O. Barentsz, Peter L. Choyke, Francois Cornud, Masoom A. Haider, Katarzyna J. Macura, Daniel Margolis, Mitchell D. Schnall, Faina Shtern, Clare M. Tempany, Harriet C.

Thoeny, and Sadna Verma. Pi-rads prostate imaging–reporting and data system: 2015, version 2. *European Urology*, 69:16–40, 2016.

[32] Massimo Mischi, Simona Turco, Cristina Lavini, Kyveli Kompatsiari, Jean JMCH de la Rosette, Marcel Breeuwer, and Hessel Wijkstra. Magnetic resonance dispersion imaging for localization of angiogenesis and cancer growth. *Investigative radiology*, 49(8):561–569, 2014.

[33] Simona Turco, Cristina Lavini, Stijn Heijmink, Jelle Barentsz, Hessel Wijkstra, and Massimo Mischi. Evaluation of dispersion mri for improved prostate cancer diagnosis in a multicenter study. *American journal of Roentgenology*, 211(5):W242–W251, 2018.

[34] Tim Ottens, Sebastiano Barbieri, Matthew R Orton, Remy Klaassen, Hanneke WM van Laarhoven, Hans Crezee, Aart J Nederveen, Xiantong Zhen, and Oliver J Gurney-Champion. Deep learning dce-mri parameter estimation: Application in pancreatic cancer. *Medical image analysis*, 80:102512, 2022.

[35] Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

[36] Jimmy T.H. Smith, Andrew Warrington, and Scott Linderman. Simplified state space layers for sequence modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[37] Daniel Y Fu, Tri Dao, Khaled Kamal Saab, Armin W Thomas, Atri Rudra, and Christopher Re. Hungry hungry hippos: Towards language modeling with state space models. In *The Eleventh International Conference on Learning Representations*, 2023.

[38] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Forty-first International Conference on Machine Learning*, 2024.

[39] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 578–588, Cham, 2024. Springer Nature Switzerland.

[40] Alex Ling Yu Hung, Haoxin Zheng, Qi Miao, Steven S Raman, Demetri Terzopoulos, and Kyunghyun Sung. Cat-net: A cross-slice attention transformer model for prostate zonal segmentation in mri. *IEEE transactions on medical imaging*, 42(1):291–303, 2022.

[41] Geoff JM Parker, Caleb Roberts, Andrew Macdonald, Giovanni A Buonaccorsi, Sue Cheung, David L Buckley, Alan Jackson, Yvonne Watson, Karen Davies, and Gordon C Jayson. Experimentally-derived functional form for a population-averaged high-temporal-resolution arterial input function for dynamic contrast-enhanced mri. *Magnetic Resonance in Medicine*, 56(5):993–1000, 2006.

[42] Anindo Saha, Joeran S Bosma, Jasper J Twilt, Bram van Ginneken, Anders Bjartell, Anwar R Padhani, David Bonekamp, Geert Villeirs, Georg Salomon, Gianluca Giannarini, Jayashree Kalpathy-Cramer, Jelle Barentsz, Klaus H Maier-Hein, Mirabela Rusu, Olivier Rouvière, Roderick van den Bergh, Valeria Panebianco, Veeru Kasivisvanathan, Nancy A Obuchowski, Derya Yakar, Mattijs Elschot, Jeroen Veltman, Jurgen J Fütterer, Constant R. Noordman, Ivan Slootweg, Christian Roest, Stefan J. Fransen, Mohammed R.S. Sunoqrot, Tone F. Bathen, Dennis Rouw, Jos Immerzeel, Jeroen Geerdink, Chris van Run, Miriam Groeneveld, James Meakin, Ahmet Karagöz, Alexandre Bône, Alexandre Routier, Arnaud Marcoux, Clément Abi-Nader, Cynthia Xinran Li, Dagan Feng, Deniz Alis, Ercan Karaarslan, Euijoon Ahn, François Nicolas, Geoffrey A. Sonn, Indrani Bhattacharya, Jinman Kim, Jun Shi, Hassan Jahanandish, Hong An, Hongyu Kan, Ilkay Oksuz, Liang Qiao, Marc-Michel Rohé, Mert Yergin, Mohamed Khadra, Mustafa E. Şeker, Mustafa S. Kartal, Noëlie Debs, Richard E. Fan, Sara Saunders, Simon J.C. Soerensen, Stefania Moroianu, Sulaiman Vesal, Yuan Yuan, Afsoun Malakoti-Fard, Agnė Mačiūnien, Akira Kawashima, Ana M.M. de M.G. de Sousa Machadov, Ana Sofia L. Moreira, Andrea Ponsiglione, Annelies Rappaport, Arnaldo Stanzione, Arturas Ciuvasovas, Baris Turkbey, Bart de Keyzer, Bodil G. Pedersen, Bram Eijlers, Christine Chen, Ciabattoni Riccardo, Deniz Alis, Ewout F.W. Courrech Staal, Fredrik Jäderling, Fredrik Langkilde, Giacomo Aringhieri, Giorgio Brembilla, Hannah Son, Hans Vanderlelij, Henricus P.J. Raat, Ingrida Pikūnienė, Iva Macova, Ivo Schoots, Iztok Caglic, Jeries P. Zawaideh, Jonas Wallström, Leonardo K. Bittencourt, Misbah Khurram, Moon H. Choi, Naoki Takahashi, Nelly Tan, Paolo N. Franco, Patricia A. Gutierrez, Per Erik Thimansson, Pieter Hanus, Philippe Puech, Philipp R. Rau, Pieter de Visschere, Ramette Guillaume, Renato Cuocolo, Ricardo O. Falcão, Rogier S.A. van Stiphout, Rossano Girometti, Ruta Briediene, Rūta Grigienė, Samuel Gitau, Samuel Withey, Sangeet

Ghai, Tobias Penzkofer, Tristan Barrett, Varaha S. Tammisetti, Vibeke B. Løgager, Vladimír Černý, Wulphert Venderink, Yan M. Law, Young J. Lee, Maarten de Rooij, and Henkjan Huisman. Artificial intelligence and radiologists in prostate cancer detection on mri (pi-cai): an international, paired, non-inferiority, confirmatory study. *The Lancet Oncology*, 25(7):879–887, 2024.

[43] Hao Chen, Qi Dou, Lequan Yu, Jing Qin, and Pheng-Ann Heng. Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images. *NeuroImage*, 170:446–455, 2018.

[44] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[45] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[46] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

[47] Chengdong Yan, Jurong Ding, Hui Zhang, Ke Tong, Bo Hua, and Shaolong Shi. Seresu-net for multimodal brain tumor segmentation. *IEEE Access*, 10:117033–117044, 2022.

[48] Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman. Computer-aided detection of prostate cancer in mri. *IEEE transactions on medical imaging*, 33(5):1083–1092, 2014.

[49] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

[50] MK Scherr, M Seitz, UG Müller-Lisse, M Ingrisch, MF Reiser, and UL Müller-Lisse. Mr-perfusion (mrp) and diffusion-weighted imaging (dwi) in prostate cancer: quantitative and model-based gadobenate dimeglumine mrp parameters in detection of prostate cancer. *European journal of radiology*, 76(3):359–366, 2010.

[51] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[52] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, October 2014.

[53] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.