Codebook VQ-VAE Approach for Prostate Cancer Diagnosis using Multiparametric MRI

Ekaterina Redekop
University of California, Los Angeles, USA
eredekop@g.ucla.edu

Mara Pleasure
University of California, Los Angeles, USA
mpleasure@g.ucla.edu

Zichen Wang
University of California, Los Angeles, USA
zcwang0702@ucla.edu

Karthik V Sarma
University of California, Los Angeles, USA
ksarma@g.ucla.edu

Adam Kinnaird
University of Alberta, Canada
ask@ualberta.ca

William Speier
University of California, Los Angeles, USA
speier@ucla.edu

Corey W Arnold
University of California, Los Angeles, USA
cwarnold@ucla.edu

Abstract

Multiparametric magnetic resonance imaging (mpMRI) plays an essential role in prostate cancer diagnosis as it can noninvasively localize and grade lesions based on their suspicion of representing clinically significant prostate cancer (csPCa). With the development of deep learning, automatic solutions for csPCa detection based on mpMRI have been developed; however, mpMRI data introduces several difficulties, including data scarcity, heterogeneity in image quality across institutions, and missing modalities. This work addresses these difficulties by building a radiology-based foundational model for prostate cancer mpMRI. Foundation models are deep learning models pretrained on a large-scale dataset and they have recently gained significant interest in computer vision and natural language applications. After pretraining, these models are often adapted for a variety of downstream tasks using smaller datasets from within the same domain. In this work, a large prostate multiparametric MRI (mpMRI) dataset was collected by combining data from our institution with two publicly available datasets. Joint modeling of all mpMRI modalities is essential for accurate prostate cancer diagnosis; however, some of these modalities may be missing. Using unsupervised learning, we pretrained modality-specific vector quantized variational autoencoders (VQ-VAE) to form a radiology foundational model. The learned codebook from VQ-VAE was then used to train a multimodal transformer to perform the diagnosis of clinically significant prostate cancer (csPCa). The proposed multimodal transformer models long-range dependencies between latent representations of input modalities and is augmented with modality-level dropout to increase the model robustness to incomplete modalities. Our framework outperforms previously published work and achieves an average AUC/sensitivity/specificity of $0.764/0.690/0.781$. *Our results show that pretraining on a larger dataset in combination with the power of transformer architecture can improve the accuracy of automatic prostate cancer detection.*

1. Introduction

Prostate cancer (PCa) was the second most deadly and the most frequently diagnosed cancer among American men in 2023 [28]. Traditional prostate cancer screening is primarily based on prostate-specific antigen (PSA) testing and digital rectal examination. Transrectal ultrasound-guided (TRUS) prostate tissue biopsy procedure is typically performed for men with high PSA and/or palpable lesions. The presence of cancer and its aggressiveness is quantified by pathologists with the ISUP Grade Group system (GG) [10, 12]. GG ranges from 1 to 5, with an increasing risk of cancer mortality with increasing GG. Typical ultrasound-guided biopsy is performed by uniform prostate sampling using a predefined grid. However, the ability of this approach to localize lesions is limited by what the operator may see during the procedure. To improve the diagnostic accuracy of prostate biopsy, multiparametric magnetic resonance imaging (mpMRI) can be used as a diagnostic tool for evaluating and grading lesions based on their suspicion of representing clinically significant prostate cancer (csPCa) (with $GG \geq 2$).

Recent advancements in deep learning have led to the development of numerous computer-aided diagnoses (CADs) for csPCa detection, achieving high predictive accuracy. For example, Zhuang et al. [35] proposed a radiomics-based approach for determining the presence of csPCa on mpMRI. The method achieved 73.96% accuracy on the dataset consisting of 26 patients; however, their proposed solution relies on tumor segmentation masks manually drawn by expert radiologists. Obtaining pixel-level annotations is time-consuming and a subjective task.

Another recent work by Zhao et al. [34] developed a deep learning-based algorithm utilizing ShuffleNet3D to detect the presence of csPCa based on mpMRI, which was trained and evaluated on a large cohort of 1,861 patients and resulted in an AUC of 0.896. Similar to the work of Zhuang et al., the model relied on expert-annotated 3D ROIs containing intratumoral and peritumoral tissues extracted from mpMRI for training.

Redekop et al. [24] developed a weakly supervised biopsy target detection and prostate cancer diagnosis tool using mpMRI. The model was trained only on image-level annotations and achieved an average 0.75 AUC in csPCa detection. The main limitation was that the model was trained and evaluated on a dataset from a single institution without utilizing any other publicly available dataset. Additionally, the proposed solution does not account for missing modalities, and performance may degrade if one of the modalities is missing.

All the prior solutions are based on the conventional fully supervised deep learning paradigm, where model training relies on large, task-specific, and manually labeled datasets to train individual CADs [4]. A more efficient alternative can be a foundational model (FM) [4], which is a deep learning model pretrained on a large-scale, diverse dataset. Following the initial pretraining, FMs are adapted for a wide range of downstream tasks through fine-tuning. Foundational models are useful when labeled data is scarce, as they learn a more general feature set that can be fine-tuned using a smaller dataset for a specific task. Additionally, foundational models can help address issues with generalizability as they are not fully supervised models trained on only one institution's dataset.

To learn useful representations from a large pretraining dataset, various self-supervised learning techniques have been proposed [6, 9, 16], which can be summarized into three main categories: contrastive, generative, or a combination of both [21]. This work utilizes the Vector Quantised Variational AutoEncoder (VQ-VAE) – a generative model that learns discrete representations [30]. A discrete representation provides a more logical fit for the reasoning of imaging data, as images are often described using discrete language. Additionally, learning a discrete representation allows larger images to be converted to sequences, allowing for the use of established transformer architectures. The VQ-VAE architecture has been previously applied to mpMRI data in the field of brain anomaly detection where the goal is to learn a compact and expressive representation of normal brain [15, 23] that can be used to detect anomalies.

In medical imaging, pretraining has been widely utilized in digital pathology, where extensive collections of digitized slides are publicly available [7, 20]. Due to the availability of large data collections, pan-cancer pretraining leads to better performance in organ-specific downstream tasks [7, 31]. On the other hand, a similar collection doesn't exist in radiology. To our knowledge, the effect of pretraining a vector-quantized multiple modality foundational model for task-specific downstream tasks has not been studied in radiology.

mpMRI typically consists of three modalities, including T2-weighted (T2W), apparent diffusion coefficient (ADC), and high-b value diffusion-weighted images (high-b). High-b value diffusion-weighted image is integral to mpMRI-based prostate cancer diagnosis [3, 5]. It has been shown that ADC maps derived from diffusion-weighted images are able to detect prostate cancer, and ADC values are highly correlated with GG. [14, 18, 32]. However, lesions on ADC maps can be subtle, and it has been observed that incorporating high-b value images enhances the visibility of PCa [13, 25]. It is not always possible to collect all three modalities in clinical practice due to various scanning protocols. For example, due to little agreement on the optimal b-value, data from different institutions could be acquired with different parameters, leading to various image quality and quantitative contrast ratios of lesion to background [3]. In this case, existing solutions may fail to handle an incomplete set of modalities.

Given the benefit of foundational models for small datasets and their expressive power to learn a useful latent representation of the domain [8], we believe training a foundational model for the three radiology modalities could help in instances where data is missing. We trained three VQ-VAE models on each modality - T2W, ADC, and high-b. We then trained a transformer architecture to learn long-range dependencies between the three input modalities on the learned codebooks.

Our main contributions can be summarized as follows:

- We take the first step towards building a radiology-based

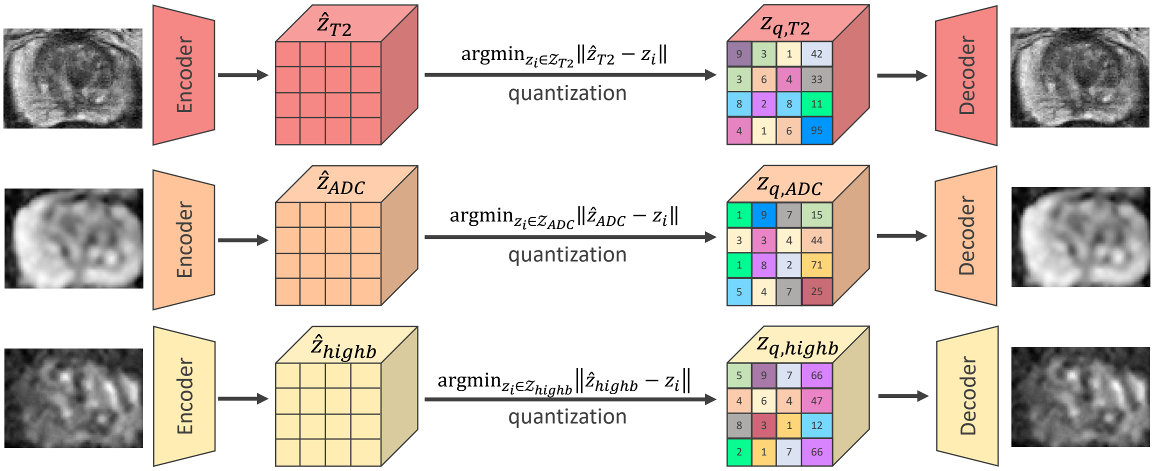

Figure 1. Multimodal VQ-VAE. The encoder output is mapped to the nearest point in the learned codebook.

FM by making a prostate cancer-specific FM. Our FM consists of three modality-specific VQ-VAE models pretrained in an unsupervised manner on a large collection of prostate mpMRI studies. To our knowledge, a large pretraining model has yet to be developed for prostate cancer radiology.

- After VQ-VAE models are pretrained, we train a transformer for the downstream task of predicting clinically significant prostate cancer. The transformer allows us to model long-range dependencies between latent representations of input modalities to achieve better results on the downstream task of csPCa detection.
- Based on the prior work presented in [33], we incorporate a modality-level dropout strategy to the transformer training for the downstream task. We show that utilizing this strategy helps to achieve comparable results when modalities are missing. To our knowledge, the influence of missing modalities on csPCa detection accuracy has not been studied previously.

2. Materials and Methods

2.1. Dataset

Foundational models require a diverse dataset to learn valuable representations; we combined two public and one private dataset to create a robust, multi-institutional dataset. The first public dataset is the PI-CAI challenge dataset, which is comprised of 1,500 studies, including T2W, ADC, and high-b value DWI images acquired using two MRI scanners (Siemens Healthineers and Philips Medical Systems-based scanners with surface coils) [26]. Patients are included only if they do not have a history of treatment or prior $GG \geq 2$ findings. Out of the 1,500 available cases, 1,075 have benign tissue or indolent PCa, and 425 cases have csPCa. The median voxel spacing is $0.5 \times 0.5 \times 3.0$.

The second public dataset was collected and released by Adams et al. [1, 2]. The dataset comprises 158 studies, including T2W, ADC, and high-b value DWI images acquired using two 3.0 Tesla MRI scanners (Siemens VIDA and Skyra, Siemens Healthineers). The dataset contains 102 patients with histologically verified PCa and 56 who served as controls. The median voxel spacing is $0.47mm \times 0.47mm \times 3.0mm$.

Our internal dataset consists of 2,308 studies collected from patients who underwent transrectal ultrasound - MRI fusion biopsy (TRUS biopsy) using the Artemis guided biopsy system (Eigen Systems) between 2010 and 2023 at our institution using a standardized protocol and 3T scanner (Trio, Verio, or Skyra, Siemens Healthineers). The dataset was split on the patient level, and 1,322 studies were held out from task-specific training to train the FM. The remaining 986 studies were used for downstream cancer diagnosis tasks. As part of this clinical process, a radiologist contoured a prostate and any regions of interest (ROIs) for targeted biopsy sampling. Based on pathology examination of biopsied tissue, 494 cases had csPCa. 3D T2W images, ADC maps, and high b-value DWI were available for all studies in the dataset. The median voxel spacing is $0.66mm \times 0.66mm \times 1.5mm$.

Our mpMRI preprocessing pipeline included bias field correction and interquartile range (IQR)-based intra-image normalization to address the relative nature of MRI intensity values [27]. All images from each dataset were resampled to the median voxel spacing of data from our institution ($0.66mm \times 0.66mm \times 1.5mm$).

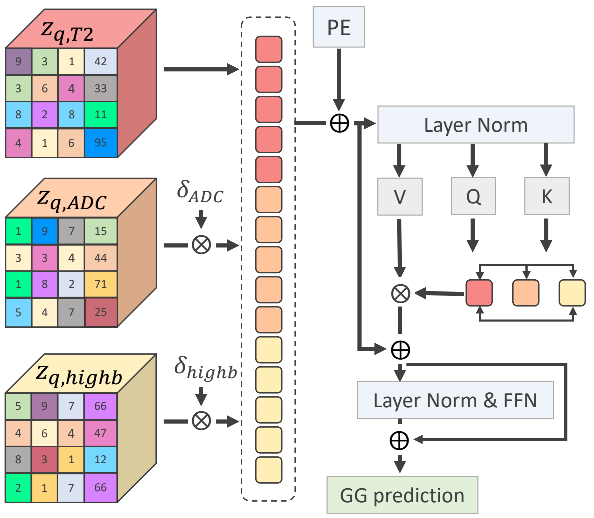

Figure 2. Multimodal transformer to model long-range dependencies between codebook encodings with modality dropout ($\delta_k, k \in \{ADC, high-b\}$). FFN - Feed Forward Network, PE - positional encoding.

2.2. VQ-VAE Multimodal Transformer

Our proposed approach consists of two steps. In the first step, three modality-specific VQ-VAE models are pretrained on a large pretraining dataset (see Fig. 1, Sec. 2.2.1). In the second step, the multimodal transformer is trained to predict the presence of csPCa based on encodings learned by the VQ-VAE models (see Fig. 2, Sec. 2.2.2). We compared our approach to several previously published baseline models (described in section Sec. 2.2.3).

2.2.1 Vector quantized variational autoencoder

The VQ-VAE model comprises an encoder E and decoder D, such that together they are trained to represent input as a set of codes from a learned discrete codebook $Z = \{Z_k\}_{k=1}^{K}$, where K is the vocabulary size [30]. First, E projects input $x \in \mathbb{R}^{H \times W \times D}$ onto a latent embedding space $\hat{z} \in \mathbb{R}^{h \times w \times d \times n_z}$, where n_z is the dimensionality of latent codes. z_q is obtained using element-wise quantization $q(\cdot)$ of each latent code $\hat{z}_{ijk}$ onto its closest codebook element z_k. Decoder D then reconstructs $\hat{x} \in \mathbb{R}^{H \times W \times D}$ from the quantized latent space. According to [30], the overall loss function consists of three components: VQ objective, commitment loss, and reconstruction loss. Following the later work [11], we replace L_2 reconstruction loss with perceptual loss [19].

In this work, three VQ-VAE models were separately trained on each modality (T2, ADC, and high-b), which resulted in three modality-specific codebooks.

2.2.2 Multimodal Transformer

Following ideas presented in the work [33], we utilize the transformer model to learn long-range dependencies between modality-specific latent encodings.

With E and D pretrained, images can now be represented in terms of the codebook indices and their embeddings. Specifically, quantized encodings of input image x are given by $z_q = q(E(x)) \in \mathbb{R}^{h \times w \times d \times n_z}$. Our Multimodal transformer consists of a Multi-head Self Attention (MSA) and a FeedForward Network (FFN). The encodings are processed sequentially by flattennig into a $1D$ sequence and transforming into token space by a linear projection with matrix W. A learnable positional embedding P is added to the input sequence to preserve location information. Modality level dropout is performed by randomly setting δ_n to 0 during training. $\delta_n \in \{0, 1\}$ is a Bernoulli indicator to add robustness while modeling long-range dependencies, even when some modalities are missing. During evaluation, the encoding vector corresponding to missing modalities was replaced with a zero vector.

$$z = [\delta_i z_{q,i}]W + P, i \in T2, ADC, high-b \qquad (1)$$

where W - weights of linear projection and P - learnable positional embedding, $[\cdot,\cdot]$ -concatenation operation. The sequence z is processed by MSA and FFN as follows:

$$z_{global} = FFN(LN(p)) + p, p = MSA(LN(z)) + z, \quad (2)$$

The MSA is formulated as follows:

$$head_m^i = softmax(\frac{Q_m^i K_m^{iT}}{\sqrt{d_k}})V_m^i, \qquad (3)$$

$$MSA = [head_m^1, ..., head_m^N]W_m^o, \qquad (4)$$

where $Q_m^i = LN(z)W_m^{Q^i}$, $K_m^i = LN(z)W_m^{K^i}$, $V_m^i = LN(z)W_m^{v^i}$, $LN(\cdot)$ - layer normalization, d_k - dimensionality of K_m, $N = 8$ - number of attention heads. FFN - two-layer perceptron with GELU activation [17].

2.2.3 Baseline solutions

We used two baselines LoGo MIL proposed by [24] and 3D ViT proposed by [22]. LoGo MIL, is a model consisting of two branches: an attention-based multiple-instance learning framework in the local branch and a CNN feature extractor in the global branch. The local branch operates on the overlapping 3D patches extracted from the input image, and the global branch operates on the entire image. 3D ViT is a 3D Vision Transformer model that operates on 3D patches and uses a self-attention mechanism to learn dependencies.

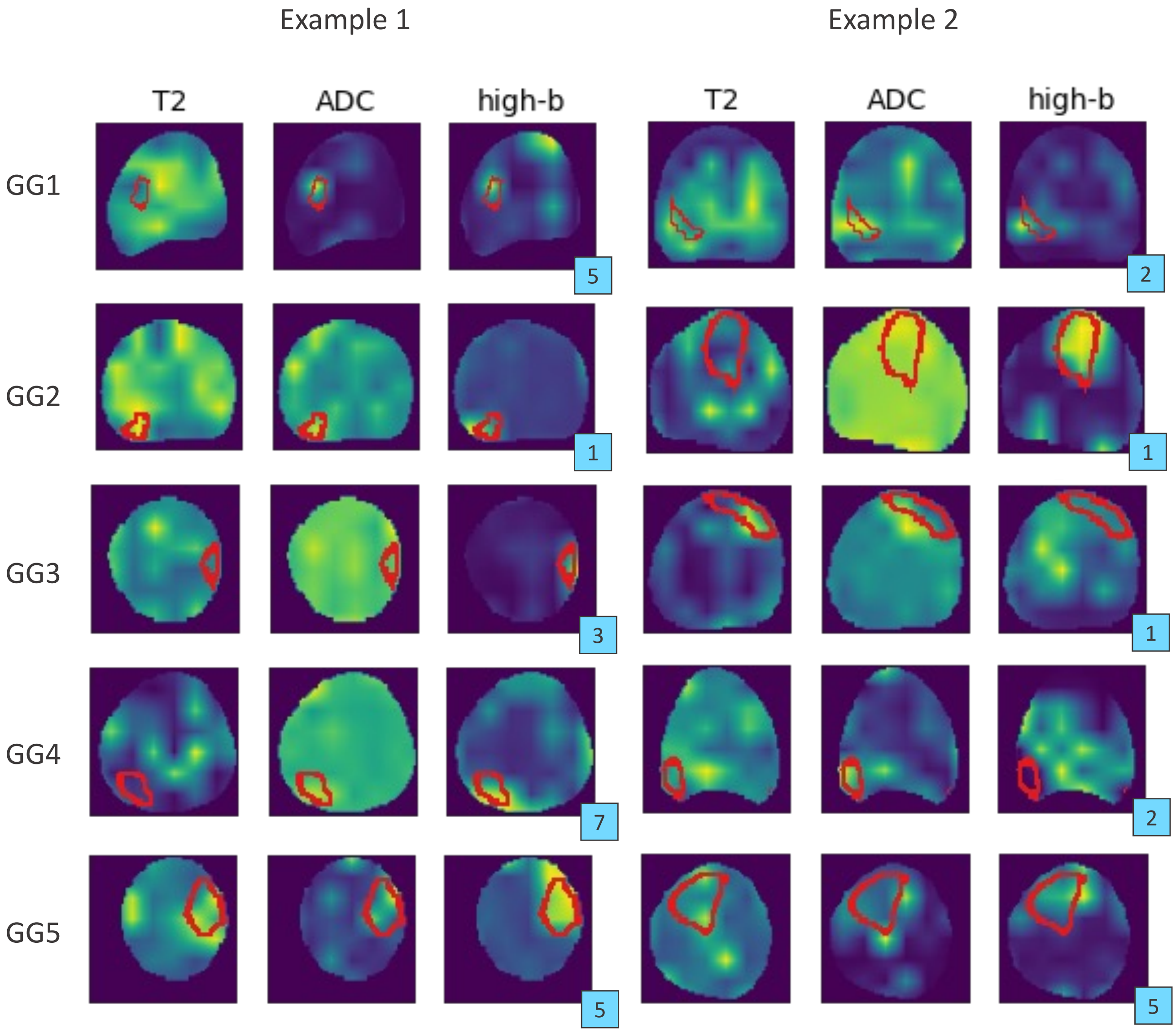

Figure 3. Attention maps from MSA block of the multimodal transformer. Attentions are presented separately for each modality. Groundtruth ROIs are indicated with red contour. A blue square with a number inside it on the bottom right corner of each example corresponds to the number of attention heads used to generate visualization.

2.3. Implementation Details

The input image size to the VQ-VAE model is $64 \times 64 \times 32$ voxels, the median size of images across utilized datasets. Random flip, rotation, and intensity shifts are employed for data augmentation. Our VQ-VAE implementation is similar to the one reported in the work by Pinaya et al. [23]. The encoder consists of three strided 3D convolutional layers with stride 2 and 256 hidden units. Resulted latent representations have size $8 \times 8 \times 4$. We used a codebook with 256 unique codes.

The multimodal transformer consists of 16 layers with an embedding size of 256. MSA block consists of 8 attention heads. To train these models, we used the Adam optimizer with an initial learning rate of $1e-4$, which was decreased by a factor of 10 if the validation loss did not improve in the last five epochs. Implementation details of the baseline solutions were taken from the original papers.

To evaluate model performance in detecting csPCa, we utilize sensitivity, specificity, F1 Score, and the area under the receiving operator characteristic curve (AUC). Models were compared using the Wilcoxon signed-rank test at a

	AUC	Sensitivity	Specificity	F1 Score
3D ViT	0.726±0.01	**0.719±0.051**	0.653±0.046	0.705±0.041
LoGo MIL	0.751±0.035	0.712±0.046	0.682±0.073	0.720±0.043
Ours	**0.764±0.019**	0.690±0.016	**0.781±0.013**	**0.751±0.017**

Table 1. Performance of prostate $GG \geq 2$ vs. $GG < 2$ classification. 3D ViT - visual transformer model, LoGo MIL - CNN-based model incorporating attention-based multiple instance learning framework.

0.05 significance level.

3. Results

In this work, we utilized encodings learned by an VQ-VAE model to perform downstream analysis of csPCa detection and compared our method to two baseline solutions described in Sec. 2.2.3. The results are presented in Tab. 1. Our solution achieved 0.764±0.019 AUC, which significantly outperformed the CNN (LoGo MIL) (AUC = 0.751±0.035, $p < 0.001$) and vision transformer (3D ViT) (AUC = 0.726±0.01, $p < 0.001$) baselines. Our method achieved the highest specificity, resulting in a lower sensitivity than other solutions. The optimal threshold weighted true positives and true negatives differently than in the baseline solutions. Our method achieved a higher F1 Score, a combined score that considers both sensitivity and specificity.

We evaluated the performance of our model on incomplete multimodal classification by setting δ_i to zero during inference for ADC, high-b, and a combination of both modalities. The dropout in the T2 modality was not applied as it was presumed to be available due to a more standard protocol. According to the results presented in Tab. 2, dropping ADC modality during evaluation results in a slight drop in AUC value (0.757±0.025 vs. 0.764±0.019), similar average sensitivity (0.690), a5.2% drop in average specificity (0.740±0.048 vs. 0.781±0.013) and a 6.4% drop in F1 score. High-b modality dropout leads to a larger drop in AUC by 6%, 9.7% drop in sensitivity, 6.7% drop in specificity, and F1 score. Dropping ADC and high-b modalities simultaneously leads to AUC, sensitivity, and specificity similar to the ones obtained by dropping high-b only.

4. Discussion

For a multimodal transformer trained to predict the presence of csPCa, we visualize the different attention heads from MSA modules and reveal that the proposed approach can localize regions suspicious of cancer in line with ROIs drawn by the radiologist. Similar observations appeared in previous research that showed that MSA blocks can be used as a method for object localization [6, 29]. In Fig. 3, we provide two examples of localization maps for each GG 1-5. Based on the visual assessment, the proposed multimodal transformer localized ROIs areas for most cases. However, we noticed that localization quality varies between modalities and between attention head numbers. Provided examples include the best localization results among all 8 heads and it can be noticed that the head number varies. While high-b modality leads to a better localization in most of the examples, ADC and T2 provide better results when high-b localization fails. Therefore, future work is needed to optimize the choice of head and modality to provide unsupervised localization results along with the prediction of the csPCa presence.

A strength of our approach is that our model is trained solely based on image-level labels, providing a more generalizable approach than methods requiring pixel-level annotations, allowing us to easily accommodate clinically-generated data. However, given that our target population for this study was all patients with biopsy and mpMRI, we do not have surgical specimens to provide a ground truth for csPCa prediction. Future work could include a dataset with matched radical prostatectomy specimens to identify a stronger ground truth of csPCa.

While modality-level dropout incorporated into our model architecture leads to robust model performance when ADC modality is missing, the absence of high-b modality leads to decreased model performance. This can be explained by the fact that lesions are typically visible on high-b images when ADC and T2 signal is subtle [13, 25]. In our future work, we will improve the technique to handle missing modalities and increase robustness to address scenarios when high-b is missing.

Lastly, although three datasets were combined for the pretraining dataset, resulting in a total of 2,980 studies used to pretrain FM, a pretraining dataset with more institutions could further benefit the training.

5. Conclusions

In this work, we presented a novel csPCa diagnosis framework which consists of a novel prostate radiology foundational model based on a multimodal VQ-VAE. Our FM was pretrained on a large set of mpMRI images from two public and one private dataset. After the multimodal VQ-VAE was pretrained, we finetuned the model for csPCa prediction utilizing a multimodal transformer with modality dropout to account for the possibility of missing modalities. The proposed solution significantly outperformed two baselines: 3D ViT and attention-based LoGo-MIL.

Included modality	AUC	Sensitivity	Specificity	F1 Score
T2, high-b	0.757±0.025	0.690±0.035	0.740±0.048	0.703±0.038
T2, ADC	0.713±0.023	0.623±0.051	0.732±0.076	0.700±0.038
T2	0.712±0.02	0.626±0.045	0.724±0.057	0.701±0.039

Table 2. Evaluation of our model performance for various modality dropout settings. T2 is always included

Modality dropout showed the robustness of the model to incomplete modality input. Our results show that automatic prostate cancer detection can be improved by utilizing a large dataset for pretraining modality-specific FM. Additionally, we found potential in using a VQ-VAE as the base for a non-prostate cancer-specific radiology FM model.

References

[1] Lisa C Adams, Marcus R Makowski, Günther Engel, Maximilian Rattunde, Felix Busch, Patrick Asbach, Stefan M Niehues, Shankeeth Vinayahalingam, Bram van Ginneken, Geert Litjens, et al. Dataset of prostate mri annotated for anatomical zones and cancer. *Data in Brief*, 45:108739, 2022. 3

[2] Lisa C Adams, Marcus R Makowski, Günther Engel, Maximilian Rattunde, Felix Busch, Patrick Asbach, Stefan M Niehues, Shankeeth Vinayahalingam, Bram van Ginneken, Geert Litjens, et al. Prostate158-an expert-annotated 3t mri dataset and algorithm for prostate cancer detection. *Computers in Biology and Medicine*, 148:105817, 2022. 3

[3] Harsh K Agarwal, Francesca V Mertan, Sandeep Sankineni, Marcelino Bernardo, Julien Senegas, Jochen Keupp, Dagane Daar, Maria Merino, Bradford J Wood, Peter A Pinto, et al. Optimal high b-value for diffusion weighted mri in diagnosing high risk prostate cancers in the peripheral zone. *Journal of Magnetic Resonance Imaging*, 45(1):125–131, 2017. 2

[4] Bobby Azad, Reza Azad, Sania Eskandari, Afshin Bozorgpour, Amirhossein Kazerouni, Islem Rekik, and Dorit Merhof. Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689*, 2023. 2

[5] Jelle O Barentsz, Jonathan Richenberg, Richard Clements, Peter Choyke, Sadhna Verma, Geert Villeirs, Olivier Rouviere, Vibeke Logager, and Jurgen J Fütterer. Esur prostate mr guidelines 2012. *European radiology*, 22:746–757, 2012. 2

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 6

[7] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 2

[8] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, pages 1–13, 2024. 2

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[10] Jonathan I Epstein, Lars Egevad, Mahul B Amin, Brett Delahunt, John R Srigley, Peter A Humphrey, Grading Committee, et al. The 2014 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma: definition of grading patterns and proposal for a new grading system. *The American journal of surgical pathology*, 40(2):244–252, 2016. 1

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 4

[12] Donald F Gleason and George T Mellinger. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *The Journal of urology*, 111(1):58–64, 1974. 1

[13] Kinzya B Grant, Harsh K Agarwal, Joanna H Shih, Marcelino Bernardo, Yuxi Pang, Dagane Daar, Maria J Merino, Bradford J Wood, Peter A Pinto, Peter L Choyke, et al. Comparison of calculated and acquired high b value diffusion-weighted imaging in prostate cancer. *Abdominal imaging*, 40:578–586, 2015. 2, 6

[14] Thomas Hambrock, Diederik M Somford, Henkjan J Huisman, Inge M van Oort, J Alfred Witjes, Christina A Hulsbergen-van de Kaa, Thomas Scheenen, and Jelle O Barentsz. Relationship between apparent diffusion coefficients at 3.0-t mr imaging and gleason grade in peripheral zone prostate cancer. *Radiology*, 259(2):453–461, 2011. 2

[15] Ravi Hassanaly, Camille Brianceau, Olivier Colliot, and Ninon Burgos. Unsupervised anomaly detection in 3d brain fdg pet: A benchmark of 17 vae-based approaches. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 110–120. Springer, 2023. 2

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[17] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[18] Yasushi Itou, Katsuyuki Nakanishi, Yoshifumi Narumi, Yasuko Nishizawa, and Hideaki Tsukuma. Clinical utility of apparent diffusion coefficient (adc) values in patients with prostate cancer: can adc values contribute to assess the aggressiveness of prostate cancer? *Journal of Magnetic Resonance Imaging*, 33(1):167–172, 2011. 2

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 4

[20] Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416, 2018. 2

[21] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021. 2

[22] Eva Pachetti, Sara Colantonio, and Maria Antonietta Pascali. On the effectiveness of 3d vision transformers for the prediction of prostate cancer aggressiveness. In *International Conference on Image Analysis and Processing*, pages 317–328. Springer, 2022. 4

[23] Walter Hugo Lopez Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sébastien Ourselin, and M Jorge Cardoso. Unsupervised brain anomaly detection and segmentation with transformers. *arXiv preprint arXiv:2102.11650*, 2021. 2, 5

[24] Ekaterina Redekop, Karthik V Sarma, Adam Kinnaird, Anthony Sisk, Steven S Raman, Leonard S Marks, William Speier, and Corey W Arnold. Attention-guided prostate lesion localization and grade group classification with multiple instance learning. In *International Conference on Medical Imaging with Deep Learning*, pages 975–987. PMLR, 2022. 2, 4

[25] Andrew B Rosenkrantz, Lorenzo Mannelli, Xiangtian Kong, Ben E Niver, Douglas S Berkman, James S Babb, Jonathan Melamed, and Samir S Taneja. Prostate cancer: Utility of fusion of t2-weighted and high b-value diffusion-weighted images for peripheral zone tumor detection and localization. *Journal of Magnetic Resonance Imaging*, 34(1):95–100, 2011. 2, 6

[26] Anindo Saha, Joeran Bosma, Jasper Twilt, Bram van Ginneken, Derya Yakar, Mattijs Elschot, Jeroen Veltman, Jurgen Fütterer, Maarten de Rooij, et al. Artificial intelligence and radiologists at prostate cancer detection in mri—the pi-cai challenge. In *Medical Imaging with Deep Learning, short paper track*, 2023. 3

[27] Karthik V Sarma, Alex G Raman, Nikhil J Dhinagar, Alan M Priester, Stephanie Harmon, Thomas Sanford, Sherif Mehralivand, Baris Turkbey, Leonard S Marks, Steven S Raman, et al. Harnessing clinical annotations to improve deep learning performance in prostate segmentation. *Plos one*, 16 (6):e0253829, 2021. 3

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

[29] Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. *arXiv preprint arXiv:2109.14279*, 2021. 6

[30] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2, 4

[31] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, et al. Virchow: A million-slide digital pathology foundation model. *arXiv preprint arXiv:2309.07778*, 2023. 2

[32] Kengo Yoshimitsu, Keijiro Kiyoshima, Hiroyuki Irie, Tsuyoshi Tajima, Yoshiki Asayama, Masakazu Hirakawa, Kousei Ishigami, Seiji Naito, and Hiroshi Honda. Usefulness of apparent diffusion coefficient map in diagnosing prostate carcinoma: correlation with stepwise histopathology. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(1):132–139, 2008. 2

[33] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 107–117. Springer, 2022. 3, 4

[34] Litao Zhao, Jie Bao, Xiaomeng Qiao, Pengfei Jin, Yanting Ji, Zhenkai Li, Ji Zhang, Yueting Su, Libiao Ji, Junkang Shen, et al. Predicting clinically significant prostate cancer with a deep learning approach: a multicentre retrospective study. *European Journal of Nuclear Medicine and Molecular Imaging*, 50(3):727–741, 2023. 2

[35] Haoming Zhuang, Aritrick Chatterjee, Xiaobing Fan, Shouliang Qi, Wei Qian, and Dianning He. A radiomics based method for prediction of prostate cancer gleason score using enlarged region of interest. *BMC Medical Imaging*, 23(1): 205, 2023. 2