BioMedical Engineering OnLine

RESEARCH Open Access

Multi-branch CNNFormer: a novel framework for predicting prostate cancer response to hormonal therapy

Ibrahim Abdelhalim[1], Mohamed Ali Badawy[2], Mohamed Abou El-Ghar[2], Mohammed Ghazal[3], Sohail Contractor[4], Eric van Bogaert[4], Dibson Gondim[5], Scott Silva[6], Fahmi Khalifa[7] and Ayman El-Baz[1*]

*Correspondence:
aselba01@louisville.edu

[1] Department of Bioengineering, University of Louisville, Louisville, KY, USA
[2] Radiology Department, Urology and Nephrology Center, Mansoura, Egypt
[3] Electrical, Computer, and Biomedical Engineering Department, Abu Dhabi University, Abu Dhabi, UAE
[4] Department of Radiology, University of Louisville, Louisville, KY, USA
[5] Department of Pathology and Laboratory Medicine, University of Louisville, Louisville, KY, USA
[6] Department of Radiation Oncology, University of Louisville, Louisville, KY, USA
[7] Electrical and Computer Engineering Department, Morgan State University, Baltimore, MD, USA

Abstract

Purpose: This study aims to accurately predict the effects of hormonal therapy on prostate cancer (PC) lesions by integrating multi-modality magnetic resonance imaging (MRI) and the clinical marker prostate-specific antigen (PSA). It addresses the limitations of Convolutional Neural Networks (CNNs) in capturing long-range spatial relations and the Vision Transformer (ViT)'s deficiency in localization information due to consecutive downsampling. The research question focuses on improving PC response prediction accuracy by combining both approaches.

Methods: We propose a 3D multi-branch CNN Transformer (CNNFormer) model, integrating 3D CNN and 3D ViT. Each branch of the model utilizes a 3D CNN to encode volumetric images into high-level feature representations, preserving detailed localization, while the 3D ViT extracts global salient features. The framework was evaluated on a 39-individual patient cohort, stratified by PSA biomarker status.

Results: Our framework achieved remarkable performance in differentiating responders and non-responders to hormonal therapy, with an accuracy of 97.50%, sensitivity of 100%, and specificity of 95.83%. These results demonstrate the effectiveness of the CNNFormer model, despite the cohort's small size.

Conclusion: The findings emphasize the framework's potential in enhancing personalized PC treatment planning and monitoring. By combining the strengths of CNN and ViT, the proposed approach offers robust, accurate prediction of PC response to hormonal therapy, with implications for improving clinical decision-making.

Keywords: Prostate cancer, Hormonal therapy, Deep learning, Vision transformer

Introduction

Prostate Cancer (PC) is a significant global health concern, being the second most frequently diagnosed cancer. In 2024, the American Society reported 299,010 new PC cases in the United States, leading to approximately 35,250 deaths [1]. PC is characterized by the uncontrolled growth of prostate gland cells, dependent on testicular hormones [2]. Androgen deprivation therapy (ADT), which uses castration and antiandrogens, is a key

treatment for advanced PC [3]. Prostate-specific antigen (PSA) is a crucial biomarker for monitoring disease progression and assessing treatment efficacy [4]. The integration of machine learning (ML) and Deep Learning (DL) techniques in grading, particularly via the Gleason Score (GS), is standard practice for predicting PC aggressiveness [5–10]. In addition, the assessment of PC therapy involves statistical analysis of clinical biomarkers [11–13] and the application of ML/DL on radiomic features or pathological markers [14, 15]. Collins and Cheng [11] provided an overview of PC treatments and highlighted histopathological changes in prostate tissue resulting from these treatments. Osiecki et al. [12] conducted a systematic review exploring the impact of morphological variants on outcomes following radical prostatectomy (RP), establishing a significant correlation. Identifying these high-risk morphologies during prostatectomy could improve prognostic accuracy and refine management strategies. While Guerra et al. [16] developed ML models to predict extracapsular extension in PC patients prior to RP, leveraging clinical, semantic, and radiomic features derived from T2-weighted MRI, decision curve analysis (DCA) and receiver operating characteristic (ROC) metrics were compared to guide feature selection. Using a training cohort of 139 patients and an external validation cohort of 55 patients, the model that integrated all feature types demonstrated the highest net benefit across relevant threshold probabilities. While DCA and the area under the curve (AUC) rankings differed, the combined model showed promise for enhancing predictive accuracy and supporting nerve-sparing surgical decision-making in clinical practice. Saito et al. [13] developed a ML-based prognostic model for PC patients undergoing ADT, demonstrating the ability of ML to provide accurate prognostic predictions for ADT outcomes in metastatic PC. Radiomic features have also been investigated for therapy assessment. Nakata et al. [14] introduced a VGG-16 based model to predict the time to castration-resistant disease in metastatic PC patients with an 80% accuracy rate. Chen et al. [15] evaluated a radiomics approach using biparametric Magnetic Resonance Imaging (bMRI) to detect significant residual PC after ADT, showing potential for improved detection. Furthermore, Spratt et al. [17] designed a ResNet-50-based model to identify patients with predominantly intermediate risk for PC who could benefit from short-term ADT. Zhang et al. [18] developed an ML model utilizing multi-modal ultrasound and PSA measurements to predict clinically significant PC with 80% sensitivity and an Area Under the Curve (AUC) of 85.5%. In summary, the literature emphasizes the utilization of diverse data sources, including clinical biomarkers, multi-modality MRI, pathological markers, and medical records, for managing PC. However, a significant challenge lies in effectively navigating the diagnostic complexities inherent in these multi-modal data sources, particularly during follow-up to evaluate hormonal therapy. The integration of multiple imaging modalities, such as diffusion-weighted imaging (DWI) and T2-weighted MRI, is crucial for a comprehensive diagnosis, as each modality provides unique insights. However, existing literature lacks an efficient methodology for integrating multi-modality MRI data, given their inherent heterogeneity. In addition, conventional ML techniques or CNN-based models often struggle to capture long-range dependence features essential for understanding the anatomical structure of medical images, underscoring the need for modeling global features. To address these limitations, we propose a multi-branch framework designed to effectively integrate multi-modality MRI data to predict PC response to hormonal therpy. This framework

consists of a hybrid approach that combines a 3D CNN encoder and a 3D Vision Transformer (ViT) encoder at each branch (i.e., a branch for each modality) to predict the impact of hormonal therapy on PC lesions. We name this hybrid approach CNNFormer. Our contributions are as follows:

- We proposed a hybrid multi-branch framework named CNNFormer, comprising a 3D CNN encoder and a 3D ViT encoder. This framework is designed to effectively predict the effects of hormonal therapy on PC using multi-modality MRI.
- By integrating multi-modality MRI data, our approach addresses the challenge of data heterogeneity. Each branch within the framework independently learns crucial contextual information from each modality, which collectively refines our understanding of the prostate's anatomical structure from diverse data sources.
- The 3D CNN encoder in our architecture serves two primary functions: extracting key local radiomic features from MRI modalities and reducing data dimensionality to enhance computational efficiency for the 3D ViT encoder. This encoder further refines these features, focusing on salient aspects within the volumetric data. Finally, by merging outputs from each branch and applying average pooling and flattening, we achieve robust classification of responders and non-responders to hormonal therapy.

Scope and importance of the study

This research aims to bridge the gap between data heterogeneity and diagnostic precision by introducing a novel framework, CNNFormer, which integrates 3D multi-modal MRI with clinical biomarkers, specifically PSA levels, to predict treatment outcomes for PC patients undergoing hormonal therapy. By leveraging the strengths of 3D CNNs and 3D ViTs, our approach overcomes the key limitations of existing models in capturing both local and global features in medical images.

The scope of this study goes beyond merely predicting treatment outcomes. It aspires to establish a reliable and robust framework that enhances clinical decision-making by equipping clinicians with a tool to monitor and tailor therapy based on predicted responses. This personalized treatment approach has the potential to significantly improve patient care by enabling earlier interventions and reducing the likelihood of ineffective treatments. By accurately identifying which patients are most likely to benefit from hormonal therapy, the framework can help clinicians avoid unnecessary treatments, thereby minimizing side effects and improving patients' quality of life. Moreover, the integration of multiple imaging modalities, such as DWI and T2 MRI images, enables a more comprehensive and nuanced understanding of the disease. This capability distinguishes the proposed model from traditional methods that rely on single-modality data.

The proposed framework could be seamlessly integrated into existing clinical workflows, assisting radiologists and oncologists in their decision-making processes. As the model undergoes further refinement and validation with larger patient cohorts, it holds promise for broader application in clinical settings, potentially transforming the management and monitoring of PC. Furthermore, the flexibility of this framework opens the

possibility for adaptation to other diagnostic tasks, underscoring the broader applicability and significance of this research.

Experiments and results

Data set

In this study, a cohort comprising 39 patients was utilized to evaluate the proposed system. Patients underwent imaging with T2-MRI and DW-MRI, alongside the collection of clinical biomarkers such as PSA levels and GS. Three b values of DW-MRI were used (i.e., b0, b500, b1400). Responders and non-responders to hormonal therapy were identified based on their PSA levels before and after treatment. Among the 39 patients, 23 were classified as non-responders, while 16 were responders. Imaging procedures were carried out using a 3 Tesla MRI scanner equipped with a phased-array body coil, following a specific multiparametric MRI (mp-MRI) protocol. Preprocessing steps were applied to all modalities, as described in "Preprocessing" section.

Setting

The proposed system was trained using the AdamW optimizer, set with a learning rate of 0.001, and a cosine annealing scheduler, along with a batch size of 8. For the purpose of training and testing, we employed a Leave-One-Out Cross-Validation (LOOCV) strategy. In addition, cross-entropy loss was utilized. The implementation was executed via PyTorch, leveraging a single NVIDIA Quadro P5000 GPU with a memory capacity of 16 GB.

CNNFormer's results

We begin our experiments by comparing various Machine Learning (ML) classifiers, including K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), eXtreme Gradient Boosting (XGB), and Fully Connected (FC) classifiers, based on the adaptive average pooling features from a 3D Multi-branch CNN (MCNN). The highest accuracy (ACC) and specificity (SPE) were obtained using the FC classifier, as shown in Table 1, row labeled MCNN. The best sensitivity (SEN) was achieved by both the FC and XGB classifiers. Since the FC classifier outperformed the XGB classifier in all other metrics, we chose to use the FC classifier for the remainder of our experiments.

Moreover, as demonstrated in Table 2, the 3D multi-branch CNNFormer (MCNNFormer) was compared against 3D MCNN and several well-known classification models,

Table 1 A comparison of various ML classifiers was conducted using the flattened concatenated adaptive average pooling features corresponding to T2-MRI and DW-MRI from the 3D MCNN

Model	ACC (%)	SEN (%)	SPE (%)
MCNN+KNN	51.28	31.25	65.22
MCNN+DT	61.54	37.50	78.26
MCNN+RF	56.41	31.25	73.91
MCNN+XGB	71.79	50.00	86.96
MCNN	**79.49**	50.00	**100.00**

As shown in the table, the highest SEN was achieved with the XGB and FC classifiers. However, since the FC classifier outperformed XGB across other metrics, it was selected for use in subsequent experiments

Bold to empahsis that this model or appraoch achieved the best results based on those highlighted values

and it is evident that the 3D MCNNFormer achieved superior results across various metrics, including ACC and SEN. While CNNs excel in modeling local features crucial for understanding the impact of hormonal therapy on PC, the performance of the 3D MCNN lags behind the 3D MCNNFormer in terms of ACC and SEN, although it performs better in terms of SPE, similar to ResNet-18 [19]. The 3D ViT, adept at capturing global features, plays a pivotal role in comprehending the characteristics of the prostate relative to its surrounding tissue. Consequently, the seamless fusion of 3D CNN and 3D ViT enables the integration of their respective strengths, leading to optimal outcomes.

In addition, models such as Wang et al.'s model [20], UniFormer-S [21], PosMLP-Video [22], and I^2GCN [23] demonstrate significant improvements over models such as SqueezeNet [24], EfficientNet-b0 [25], and ResNet-18 [19]. The method of Wang et al. achieves impressive results with an ACC of 90.25% and a high SEN of 92.74%. UniFormer-S further refines these metrics, reaching an ACC of 93.23% and SEN of 94.56%, marking a notable improvement in both metrics. PosMLP-Video and I^2GCN continue to enhance these metrics, with I^2GCN achieving the highest performance among these models, with an ACC of 95.51% and SEN of 96.64%, although it still performs slightly worse than our 3D MCNNFormer. In contrast, ConvNext-T [26] exhibits the poorest performance, with an ACC of only 51.28%, and poor SEN (50%) and SPE (52.17%). For additional details, please refer to Fig. 1, which illustrates the confusion matrix, the ROC curve, and the corresponding AUC for the 3D MCNNFormer. It is worth highlighting that 3D versions of SqueezeNet, EfficientNet-b0, and ResNet-18 were utilized, with smaller variants of competing models being adopted due to memory constraints. Moreover, for the these models, a dual-branch architecture was implemented for processing DW-MRI and T2-MRI data, mirroring the design strategy applied in the 3D MCNNFormer. Also, the 3D MCNNFormer demonstrates inferior performance compared to ResNet-18 and MCNN in terms of SPE. However, in clinical practice, greater emphasis is placed on SEN, as it reflects the model's ability to accurately identify true positive cases, which is of paramount importance.

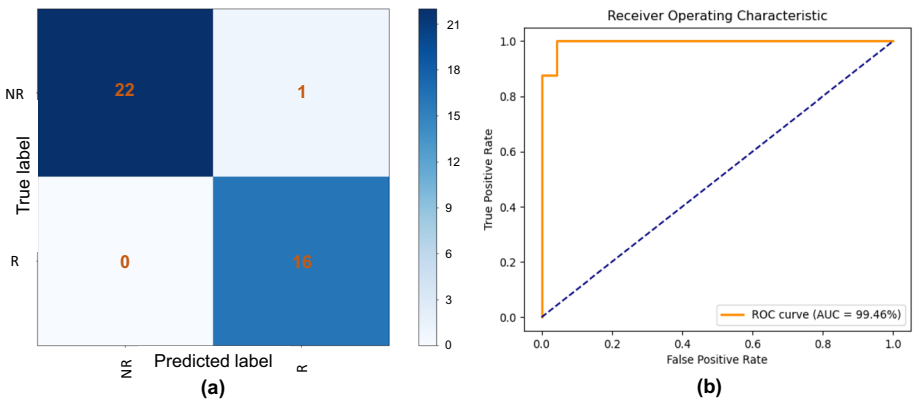

Fig. 1 Confusion matrix alongside the AUC and the receiver operating characteristic (ROC) curve of the 3D MCNNFormer. **a** Confusion matrix, detailing the classification performance, while **b** displays the AUC and ROC curve, providing an overview of the model's discriminatory ability. In the confusion matrix, 'NR' signifies non-responders, while 'R' denotes responders

Furthermore, we provide the number of parameters (in millions) for each model in Table 2, under the column labeled 'Params (M)'. As shown, the MCNNFormer requires 51.95M parameters, which is fewer than ConvNext-T and Wang et al.'s model but greater than most of the other competing models. Despite some models having fewer parameters than the MCNNFormer, it consistently achieves the best results among all models compared. It is also important to note that we used the smaller version of UniFormer, which requires 31.02M parameters, due to memory constraints. Running the larger version, which demands a significantly higher number of parameters, was not feasible in our experimental setup. In contrast, we successfully accommodated the MCNNFormer within the available memory constraints by incorporating a 3D CNN encoder. This design choice effectively reduced the input size and alleviated memory limitations.

To provide additional evidence of the model's performance, we collected another DW-MRI data set consisting of 94 patients, used for diagnosing whether transplanted kidneys were normal (54 patients) or exhibiting acute rejection (40 patients). This data set includes the three b values: 0, 500, and 1000. The same preprocessing steps mentioned in "Preprocessing" section were applied, including concatenation of the b values. For the experiments corresponding to this data set, we did not utilize multi-branches, as only one modality, DW-MRI with three b values, was available. Besides, LOOCV was employed for training and testing. Table 3 presents the results on this data set, further supporting the superiority of the 3D CNNFormer over other models, even in the absence of multi-branching. In addition, the number of parameters is reduced compared to those listed in Table 2, as the multi-branch architecture was not utilized. For a visual example of the kidney data set, refer to Fig. 2.

Ablation study

To evaluate the effectiveness of each component of the 3D MCNNFormer, we conducted an ablation study based on SEN results, which hold significant relevance in clinical contexts, as outlined in Table 4. The study began by utilizing the DW-MRI

Table 2 Comparison between the 3D MCNNFormer, 3D MCNN, and some well-known classification models demonstrates that the superior results in terms of ACC and SEN, highlighted in bold, are achieved by the 3D MCNNFormer

Model	ACC (%)	SEN (%)	SPE (%)	Params (M)
SqueezeNet [24]	82.05	93.75	73.91	3.68
Efficient-b0 [25]	89.74	93.75	86.96	9.38
ResNet-18 [19]	76.92	43.75	100.00	21.69
ConvNext-T [26]	51.28	50.00	52.17	56.00
Wang et al. [20]	90.25	92.74	88.91	72.10
UniFormer-S [21]	93.23	94.56	92.12	31.02
PosMLP-Video [22]	95.00	95.67	92.91	28.40
I^2GCN [23]	95.51	96.64	93.89	6.00
MCNN	79.49	50.00	100.00	0.28
MCNNFormer	**97.50**	**100**	95.83	51.95

Here, MCNN denotes the 3D Multi-branch CNN, while MCNNFormer indicates the 3D Multi-branch CNNFormer, with each branch corresponding to T2-MRI and DW-MRI. T indicates the tiny version of that model while Params represent the number of parameters for each model, which are measured in Millions (M)

Bold to empahsis that this model or appraoch achieved the best results based on those highlighted values

Table 3 Comparison of the 3D CNNFormer, 3D CNN, and several well-known classification models on the kidney data set reveals that the 3D CNNFormer outperforms the others, as indicated by the results highlighted in bold

Model	ACC (%)	SEN (%)	SPE (%)	Params (M)
SqueezeNet [24]	83.12	91.34	75.67	1.84
Efficient-b0 [25]	87.45	91.23	84.56	4.55
ResNet-18 [19]	78.89	45.12	91.76	10.2
ConvNext-T [26]	54.34	48.76	53.45	27.60
Wang et al. [20]	91.12	90.45	89.23	35.64
UniFormer-S [21]	93.67	93.12	94.34	15.20
PosMLP-Video [22]	96.89	90.23	94.56	13.89
I^2GCN [23]	96.12	94.45	96.78	4.8
CNN	80.23	51.67	95.12	0.14
CNNFormer	**98.76**	**96.89**	**97.45**	25.98

In this context, CNN refers to the 3D CNN, while CNNFormer denotes the 3D CNNFormer, with each branch corresponding to T2-MRI and DW-MRI. The notation T represents the tiny version of the model, while S refers to the small version. Params represent the number of parameters for each model, which are measured in Millions (M)

Bold to empahsis that this model or appraoch achieved the best results based on those highlighted values

alone with a 3D CNN encoder, achieving 18.75% (as shown in the first row of Table 4). We attribute this lower performance to the 3D CNN encoder being fed with concatenated different b values (i.e., b0, b500, b1400) of DWI, which could have resulted in distraction due to the variations in these b values. Afterwards, the T2-MRI alone was utilized with the 3D CNN encoder, which achieved 50%. After extracting high-level local features from both DW-MRI and T2-MRI using the 3D MCNN encoder and fusing them through concatenation, we passed them through the classification layer (as shown in the third row of Table 4), achieving a SEN of 50%. However, this did not show any improvement over the T2-MRI alone. In the fourth row of Table 4, we once again used DW-MRI alone with the 3D CNN-Former encoder, achieving 93.75%, a significant improvement compared to the SEN of the first row of Table 4. This suggests that the high-level local radiomic features extracted from the 3D CNN encoder are highly attended by the 3D ViT encoder, resulting in salient informative features that comprehensively represent both the local and global understanding of the prostate. Similarly, the use of T2-MRI alone with a 3D CNNFormer encoder resulted in a substantial improvement, with a SEN of 87.50%, compared to the second row of Table 4. Finally, by concatenating the outputs from each branch of the 3D CNNFormer and passing them through the classification layer (refer to Fig. 4), we were able to leverage the diverse characteristics of the prostate from different modalities, resulting in a SEN of 100%. This improvement indicates that utilizing multi-modal input within the 3D MCNNFormer, along with concatenating the highly informative features of each modality, leads to a comprehensive understanding of the PC response to hormonal therapy. This approach mirrors common clinical practice, where multi-modality is employed for accurate diagnosis. However, our aim is to enhance efficiency, representing a significant improvement over using each modality alone. For completeness, we also include the number of parameters in millions in Table 4, under the column labeled 'Params (M)'.

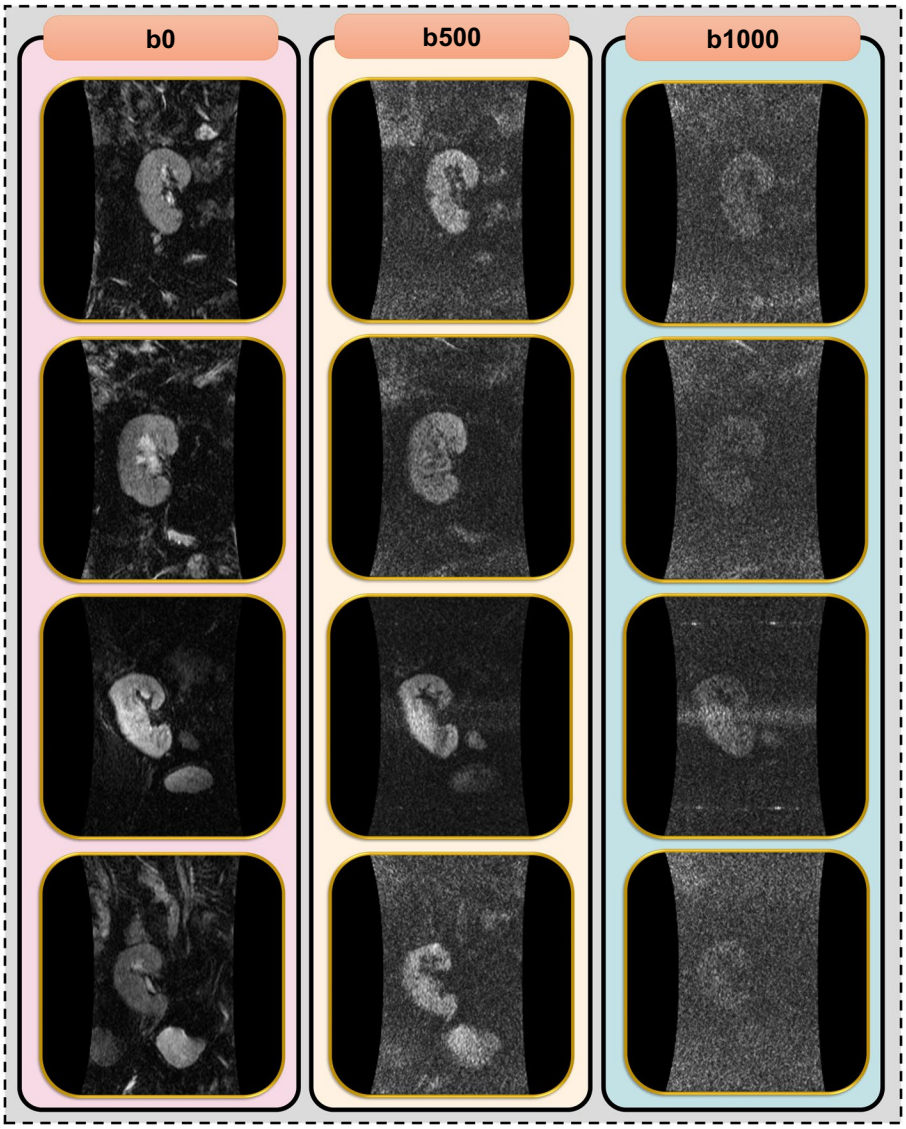

Fig. 2 Examples of DW-MRI images from the kidney data set, with each column labeled according to the corresponding b value of the DW-MRI images

As illustrated, the 3D ViT requires a large number of parameters. This underscores one of the motivations for employing a 3D CNN encoder: to reduce the input size and alleviate memory constraints. In the case of 3D CNNFormer for T2-MRI, the parameter count was 25.97 M, while for DW-MRI, it was 25.98 M, attributable to the difference in the number of channels: one for T2-MRI and three for DW-MRI. Conversely, both T2-MRI and DW-MRI models had identical params in the case of 3D CNN, specifically 0.14 M, owing to the negligible difference in params, which is negligible when rounded. In order to visually illustrate the distinct regions that the 3D MCNNFormer emphasizes while classifying PC patients into responders and non-responders to hormonal therapy, we employ the Class Activation Map (CAM).

Table 4 Ablation study provides a comprehensive evaluation of the SEN achieved by the 3D MCNNFormer model and its individual components

DW-MRI	T2-MRI	3D CNN	3D CNNFormer	SEN (%)	Params (M)
✓		✓		18.75	0.14
	✓	✓		50.00	0.14
✓	✓	✓		50.00	0.28
✓			✓	93.75	25.98
	✓		✓	87.50	25.97
✓	✓		✓	**100.00**	51.95

The study examines the influence of various inputs and architectures, including DW-MRI and T2-MRI data processed through either a 3D CNN encoder or the hybrid 3D CNNFormer encoder. The results reveal that the 3D CNNFormer substantially enhances SEN compared to the 3D CNN alone by effectively leveraging both high-level local and global features. A perfect SEN of 100% is achieved when combining outputs from both DW-MRI and T2-MRI branches of the 3D MCNNFormer, underscoring the model's ability to integrate multi-modal inputs for improved performance. In addition, the number of parameters for each configuration (in millions) is reported (i.e., the column labeled 'Params (M)'), highlighting the trade-offs in computational complexity. These findings emphasize the importance of modality fusion and hybrid architectures in achieving optimal predictions of PC response to hormonal therapy

Bold to empahsis that this model or appraoch achieved the best results based on those highlighted values

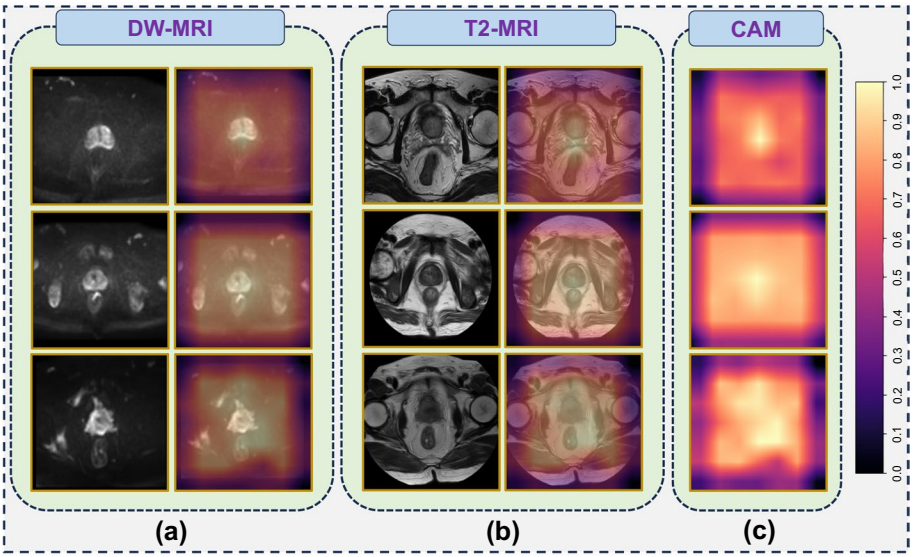

Fig. 3 Areas emphasized by the 3D MCNNFormer in classifying the impact of hormonal therapy on PC using Class Activation Maps (CAM). **a** shows a cross section of DW-MRI with an overlaid CAM, where the activation intensity highlights the model's focus on the prostate region while de-emphasizing less relevant anatomical areas. **b** depicts a cross section of T2-MRI with a CAM overlay, demonstrating a similar pattern of high activation in the prostate region, underscoring the model's consistency in identifying clinically significant regions across different imaging modalities. As inferred from **c**, the model primarily focuses on the prostate region, with intensity gradually decreasing as the distance from this region increases. This behavior illustrates the model's ability to target regions of interest while minimizing the influence of irrelevant areas, ensuring accurate and reliable feature extraction

Please refer to Fig. 3 for the CAM overlay on the cross sections of DW-MRI and T2-MRI.

Conclusions and future work

In this paper, we present a new framework designed to predict the response of PC to hormonal therapy. Our method leverages the unique characteristics revealed by various imaging modalities, allowing for a thorough evaluation of PC lesions and enhancing

the assessment of outcomes from hormonal therapy. Specifically, we extract high-level local radiomic features using a 3D CNN encoder for each modality (i.e., T2-MRI and DW-MRI). These features are then passed through a 3D ViT, one for each modality, to augment our system's ability to model global features. Then, we concatenate the highly informative features from each branch and pass them through a classification layer to obtain class logits, achieving SEN of 100%. As a result, our fusion strategies showcase the potential of the proposed framework to improve diagnostic performance, thereby serving as an auxiliary tool to optimize treatment planning and facilitate enhanced monitoring of patient responses to hormonal therapy. To further validate the improved performance of our approach compared to other models, we collected an additional data set, demonstrating that our model outperforms others in terms of ACC, SEN, and SPE. In future work, we aim to collect more data and explore other imaging modalities and data sets to increase the accuracy and reliability of the 3D MCNNFormer in assessing the outcomes of hormonal therapy in PC patients. This will enable a more comprehensive evaluation of its effectiveness. In addition, we plan to investigate more state-of-the-art models, including larger ones, and compare them with our approach. Furthermore, we intend to investigate the integration of radiomic features with biomedical markers such as GS and demographic information.

Methodology

The workflow of the proposed framework for hormonal therapy assessment is illustrated in Fig. 4. The input data comprises MRI scans (T2 and DWI) along with PSA levels. Regarding the imaging input, the framework initiates with a preprocessing step aimed at preparing the data for our 3D multi-branch CNNFormer. Then, the PSA levels are used to categorize the PC's data into responders and non-responders to hormonal therapy. Each branch consists of a 3D CNNFormer that leverages a 3D CNN encoder and 3D

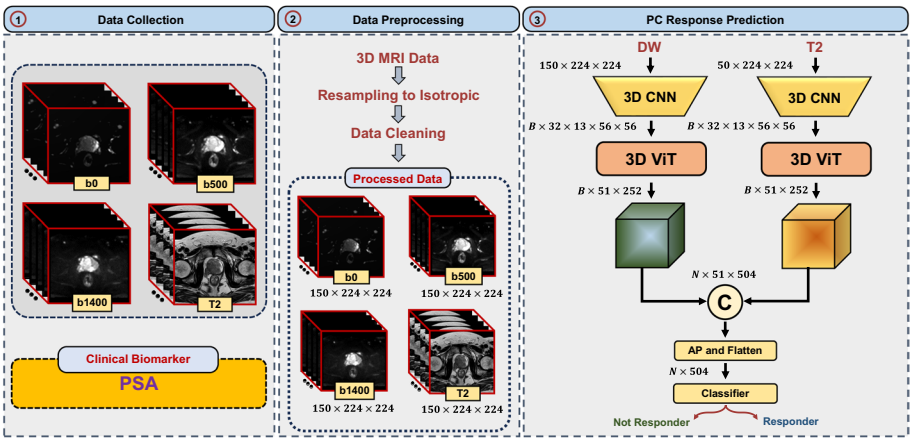

Fig. 4 Proposed framework for predicting the impact of hormonal therapy on PC begins with preprocessing input data, including T2 and DW MRI images. These images are resampled to an isotropic format and cleaned to generate volumes with a consistent resolution of 1 × 50 × 224 × 224. For DW-MRI, the b values of 500, 1400, and the baseline b0 are concatenated before being fed into a 3D CNN encoder, resulting in volumes of 3 × 50 × 224 × 224. PSA levels are then used to categorize PC data into responders and non-responders. Each branch of the model incorporates a 3D CNNFormer, which combines a 3D CNN encoder to extract local radiomic features and a 3D ViT encoder to capture global information from the features

ViT encoder. For the 3D CNN encoder, a sequence of convolutional layers is employed to extract high-level local radiomic features from volumetric MRI scans of PC. Furthermore, to capture global information from these features, 3D ViT is employed.

Preprocessing

To accurately predict the response of PC to hormonal therapy, the input data is subjected to preprocessing to ensure its seamless integration into our network. The input data comprises T2 and DW MRIs. Initially, these images are resampled to an isotropic form followed by a data cleaning step to generate volumes with a uniform resolution of $1 \times 50 \times 224 \times 224$ for each T2 and DW MRI image. For the DW-MRI images, we consider the b values of 500, 1400, and the baseline b0, which are concatenated prior to being input into the 3D CNN encoder, resulting in $3 \times 50 \times 224 \times 224$. Examples of DW-MRI and T2-MRI can be found in Fig. 5.

CNNFormer

The inherent limitation of convolutional filters lies in their inability to capture global information within images, as they primarily focus on local features. While local radiomic features are pivotal for understanding how PC responds to hormonal therapy,

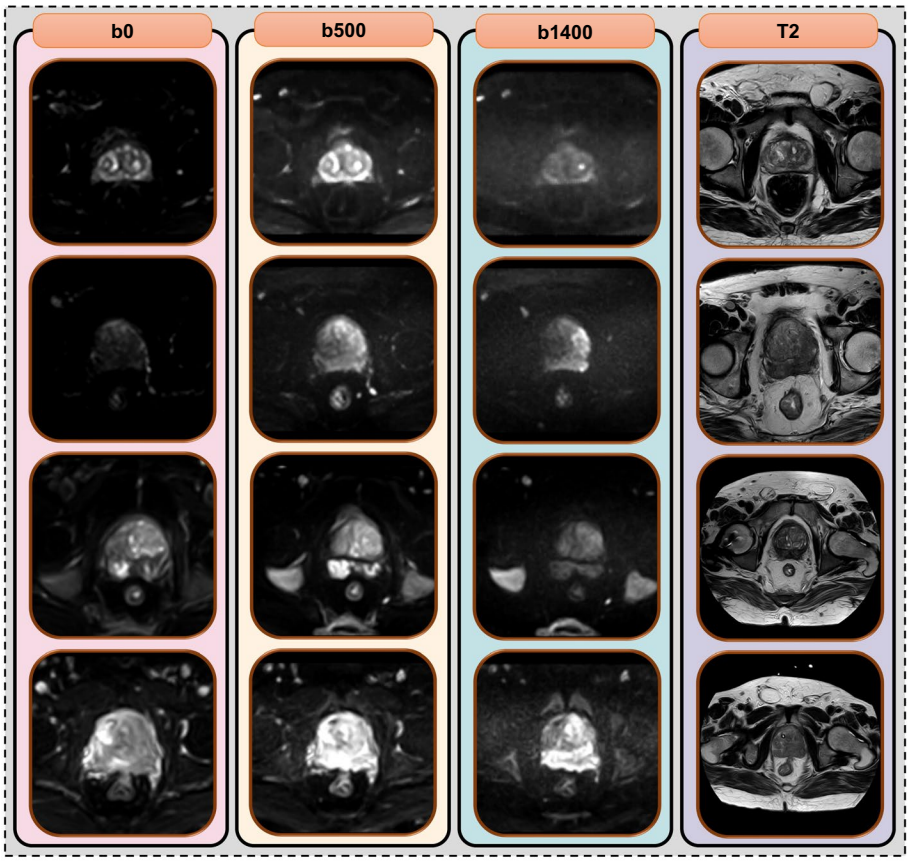

Fig. 5 Examples of DW-MRI and T2-MRI. The first three columns from the left represent examples of b values (b0, b500, and b1400), as labeled on each column of the DW-MRI. The final column illustrates examples of T2-MRI

Table 5 Detailed architecture of the 3D CNN encoder

Layer	Settings	DW-MRI	T2-MRI
Input	–	$B \times 3 \times 50 \times 224 \times 224$	$B \times 1 \times 50 \times 224 \times 224$
DoubleConv3D	3 × 3, 1, 1	$B \times 16 \times 50 \times 224 \times 224$	$B \times 16 \times 50 \times 224 \times 224$
Conv3D	3 × 3, 2, 1	$B \times 16 \times 25 \times 112 \times 112$	$B \times 16 \times 25 \times 112 \times 112$
DoubleConv3D	3 × 3, 1, 1	$B \times 32 \times 25 \times 112 \times 112$	$B \times 32 \times 25 \times 112 \times 112$
Conv3D	3 × 3, 2, 1	$B \times 32 \times 13 \times 56 \times 56$	$B \times 32 \times 13 \times 56 \times 56$
DoubleConv3D	3 × 3, 1, 1	$B \times 32 \times 13 \times 56 \times 56$	$B \times 32 \times 13 \times 56 \times 56$

'Settings' represent the kernel size, stride, and padding, respectively. The term 'DoubleConv3D' denotes a sequential list of operations, which includes two 3D convolution layers (Conv3D) and Rectified Linear Units (ReLU), arranged in the following order: Conv3D, ReLU, Conv3D, and ReLU. The symbol *B* indicates the batch size

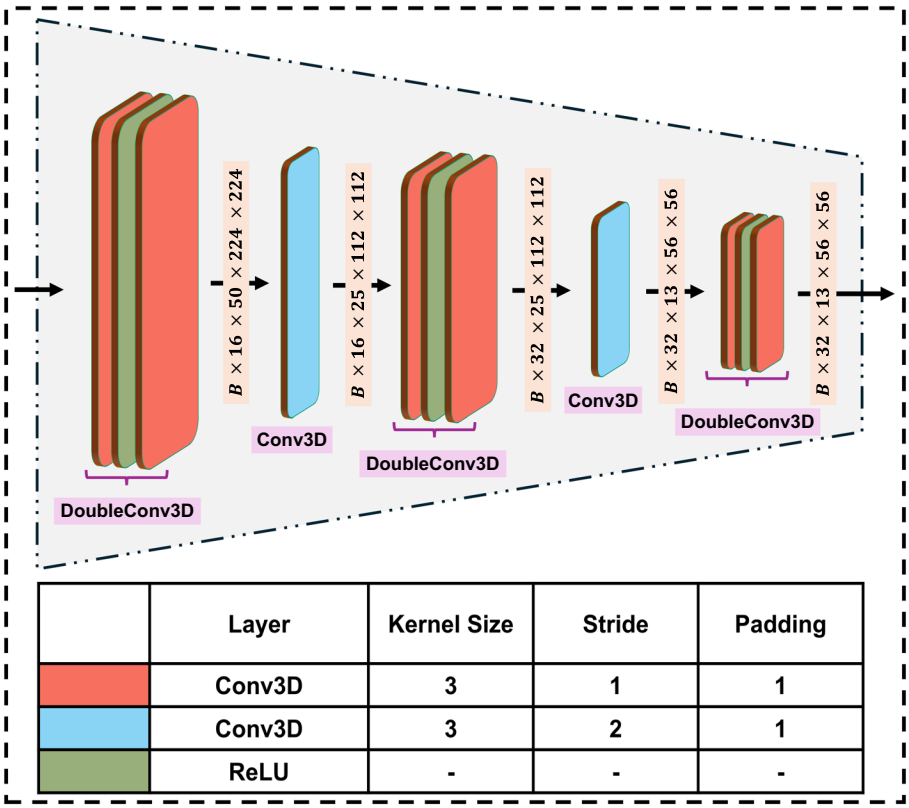

Fig. 6 Architecture of the 3D CNN model. The DoubleConv3D component consists of a 3D convolutional layer followed by a Rectified Linear Unit (ReLU), which is then succeeded by another 3D convolutional layer. Each convolutional layer utilizes a kernel size of 3 × 3, a stride of 1, and padding of 1. In addition, to reduce the input size by half at each step, a 3D convolutional layer with a kernel size of 3 × 3, a stride of 2, and padding of 1 is employed for downsampling. It is important to note that the number of channels is adjusted depending on whether the input is DW-MRI or T2-MRI. Specifically, DW-MRI comprises three channels corresponding to b0, b500, and b1400, whereas T2-MRI consists of a single channel

the absence of broader contextual understanding hinders the interpretation of the anatomical structure of the prostate concerning its surroundings. Moreover, applying ViT directly to full-resolution volumetric images results in significant computational complexity. To address these challenges, we propose a hybrid approach, so-called CNN-Former, that merges the strengths of both a CNN-based and a ViT-based models. The

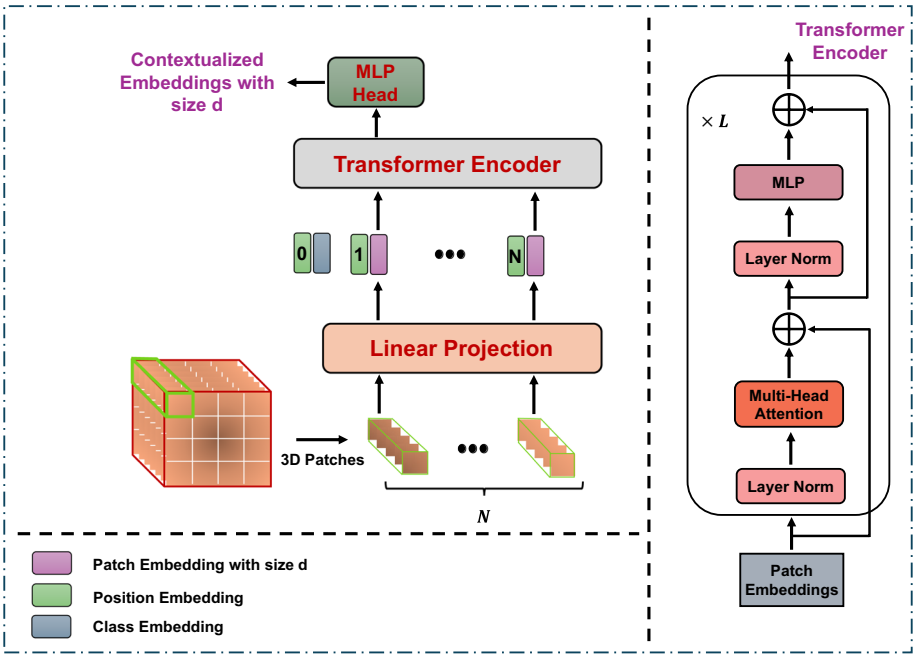

Fig. 7 Architecture of the 3D ViT model. Initially, the output from the 3D CNN is divided into 3D patches, which are subsequently flattened and transformed into patch embeddings of size $d = 256$. These patch embeddings are augmented with positional embeddings and class embeddings before being fed into the transformer encoder. The encoder comprises L layers, where $L = 12$, to generate contextualized embeddings

developed 3D CNN encoder (see Table 5 and Fig. 6) serves two crucial functions. Firstly, it encodes MRI scans (specifically, T2 and DWI) into high-level local radiomic feature representations. This is essential because ViT tends to prioritize capturing broad visual features over fine local details, mainly due to its iterative reduction of image resolution. This characteristic of ViT, wherein image resolution is repeatedly downsampled, leads to a diminished ability to precisely identify and localize specific details within images. Secondly, it helps alleviate to some extent the significant computational complexity associated with ViT, which is known to be computationally intensive. In addition, the CNNFormer's 3D ViT plays a vital role in directing attention towards global information, which aids in identifying the prostate structure among surrounding regions, particularly due to the visual similarities between the prostate and adjacent tissues. To further explain, for each branch (see Fig. 4), an MRI scan $x \in \mathbb{R}^{D\times W\times H}$ is encoded using a 3D CNN encoder, where D, W, and H represent the depth, width, and height of the MRI scan, respectively. This results in feature maps $F_{T2}, F_{DW} \in \mathbb{R}^{B\times C\times P_D\times P_W\times P_H}$, where P_D, P_W, and P_H denote the depth, width, and height, respectively, of the output from the 3D CNN encoder. B represents the batch size, and C represents the number of channels received by 3D ViT. In the 3D ViT (represented by an orange box), the high-level local radiomic features are divided into N vectorized $P \times C$ patches, where $P = P_D \times P_W \times P_H$, and N represents the number of patches. Following this, the patches are mapped to a latent D-dimensional space using a trainable linear projection (i.e., patch embedding) $e = [e_1, ..., e_N] \in R^{N\times P\times C}$. Learnable position embeddings $p = [p_1, ..., p_N] \in R^{N\times P\times C}$ are then added to the patch embeddings to retain positional information of the patches, resulting in the input sequence of tokens $\delta = e + p$. Next, a

set of K learnable class embeddings $c = [c_1, ..., c_K] \in \mathbb{R}^{N \times K \times C}$, where K corresponds to the number of classes indicating responder and non-responder PC, is processed alongside δ by the 3D ViT encoder, which consists of L layers. We denote the concatenation of c and δ as Δ. Each layer comprises a multi-headed self-attention (MSA) block followed by a pointwise MLP block, with Layer Normalization (LN) applied before and residual connections added after each block:

$$a_{i-1} = \mathrm{MSA}(\mathrm{LN}(\Delta_{i-1})) + \Delta_{i-1},$$
$$\Delta_i = \mathrm{MLP}(\mathrm{LN}(a_{i-1})) + a_{i-1},$$

where $i \in 1, ..., L$. The self-attention mechanism computes queries $\mathbf{Q} \in \mathbb{R}^{N \times d}$, keys $\mathbf{K} \in \mathbb{R}^{N \times d}$, and values $\mathbf{V} \in \mathbb{R}^{N \times d}$ via three pointwise linear layers, followed by self-attention calculation:

$$\mathbf{MSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}.$$

The transformer encoder maps the input sequences to a contextualized encoding sequence containing rich, salient information, $\Delta_L = [\Delta_{L,1}, ..., \Delta_{M,N}]$. Subsequently, this contextualized sequence is then average-pooled and reshaped for classification by a classification head comprising one LN and a fully connected classification layer (see Fig. 4 and Fig. 7).

Acknowledgements
This study was made possible through the generous support of the American people via the United States Agency for International Development (USAID). The authors are solely responsible for the content, which does not necessarily reflect the views of USAID or the United States Government.

Author contributions
I.A. (Ibrahim Abdelhalim) conceptualized the proposed approach, developed the methodology, executed the coding, and wrote the main manuscript text. A.E. (Ayman El-Baz) and F.K. (Fahmi Khalifa) provided supervision and were also involved in manuscript writing and revisions. M.G. (Mohammed Ghazal) assisted in writing the manuscript. M.A.B. (Mohamed Ali Badawy), M.A.E. (Mohamed Abou El-Ghar), S.C. (Sohail Contractor), E.V.B. (Eric VAN Bogaert), D.G. (Dibson Gondim), and S.S. (Scott Silva) gathered and ensured the integrity of the data. All authors reviewed the manuscript before submission.

Funding
This research is supported by the Science and Technology Development Fund (STDF), Egypt (grant 45889).

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
All procedures conducted in studies involving human participants adhered to the ethical standards of the institutional and/or national research committees, as well as the 1964 Helsinki Declaration and its subsequent amendments or equivalent ethical guidelines. Informed consent was obtained from all individual participants included in the study.

Competing interests
The authors declare that they have no competing financial or non-financial interests related to this work.

Received: 2 October 2024 Accepted: 11 December 2024
Published online: 23 December 2024

References

1. Society AC. Key statistics for prostate cancer. https://www.cancer.org/cancer/types/prostate-cancer/about/key-statistics.html.

2. Tanaka G, Tsumoto K, Tsuji S, Aihara K. Bifurcation analysis on a hybrid systems model of intermittent hormonal therapy for prostate cancer. Physica D: Nonlinear Phenomena. 2008;237(20):2616–27.
3. Yanagisawa T, Rajwa P, Thibault C, Gandaglia G, Mori K, Kawada T, et al. Androgen receptor signaling inhibitors in addition to docetaxel with androgen deprivation therapy for metastatic hormone-sensitive prostate cancer: a systematic review and meta-analysis. Eur Urol. 2022;82:584–98.
4. Crawford ED, Rosenblum M, Ziada AM, Lange PH. Overview: hormone refractory prostate cancer. Urology. 1999;54(6):1–7.
5. Özhan O, Yağin FH. Machine learning approach for classification of prostate cancer based on clinical biomarkers. J Cogn Syst. 2022;7(2):17–20.
6. Hassan MR, Islam MF, Uddin MZ, Ghoshal G, Hassan MM, Huda S, et al. Prostate cancer classification from ultrasound and MRI images using deep learning based Explainable Artificial Intelligence. Future Generation Comput Syst. 2022;127:462–72.
7. Ordones FV, Kawano PR, Vermeulen L, Hooshyari A, Scholtz D, Gilling PJ, et al. A novel machine learning-based predictive model of clinically significant prostate cancer and online risk calculator. Urology. 2024. https://doi.org/10.1016/j.urology.2024.11.001.
8. Cheng G, Xu J, Wang H, Chen J, Huang L, Qian ZR, et al. mtPCDI: a machine learning-based prognostic model for prostate cancer recurrence. Fronti Genet. 2024;15:1430565.
9. Müller D, Meyer P, Rentschler L, Manz R, Hieber D, Bäcker J, et al. Assessing the performance of deep learning for automated gleason grading in prostate cancer; 2024. arXiv preprint arXiv:2403.16695.
10. Talaat FM, El-Sappagh S, Alnowaiser K, Hassan E. Improved prostate cancer diagnosis using a modified ResNet50-based deep learning architecture. BMC Med Inf Decis Making. 2024;24(1):23.
11. Collins K, Cheng L. Morphologic spectrum of treatment-related changes in prostate tissue and prostate cancer: an updated review. Human Pathol. 2022;127:56–66.
12. Osiecki R, Kozikowski M, Sarecka-Hujar B, Pyzlak M, Dobruch J. Prostate cancer morphologies: cribriform pattern and intraductal carcinoma relations to adverse pathological and clinical outcomes-systematic review and meta-analysis. Cancers. 2023;15(5):1372.
13. Saito S, Sakamoto S, Higuchi K, Sato K, Zhao X, Wakai K, et al. Machine-learning predicts time-series prognosis factors in metastatic prostate cancer patients treated with androgen deprivation therapy. Sci Rep. 2023;13(1):6325.
14. Nakata W, Mori H, Tsujimura G, Tsujimoto Y, Gotoh T, Tsujihata M. Pilot study of an artificial intelligence-based deep learning algorithm to predict time to castration-resistant prostate cancer for metastatic hormone-naïve prostate cancer. Jpn J Clin Oncol. 2022;52(9):1062–6.
15. Chen ZZ, Gu WJ, Zhou BN, Liu W, Gan HL, Zhang Y, et al. Radiomics based on biparametric MRI for the detection of significant residual prostate cancer after androgen deprivation therapy: using whole-mount histopathology as reference standard. Asian J Androl. 2023;25(1):86.
16. Guerra A, Orton MR, Wang H, Konidari M, Maes K, Papanikolaou NK, et al. Clinical application of machine learning models in patients with prostate cancer before prostatectomy. Cancer Imaging. 2024;24(1):24.
17. Spratt DE, Tang S, Sun Y, Huang HC, Chen E, Mohamad O, et al. Artificial intelligence predictive model for hormone therapy use in prostate cancer. NEJM Evidence. 2023;2(8):EVIDoa2300023.
18. Zhang M, Liu Y, Yao J, Wang K, Tu J, Hu Z, et al. Value of machine learning-based transrectal multimodal ultrasound combined with PSA-related indicators in the diagnosis of clinically significant prostate cancer. Front Endocrinol. 2023;14:1137322.
19. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–8.
20. Wang J, Yang X, Li H, Liu L, Wu Z, Jiang YG. Efficient video transformers with spatial-temporal token selection. In: European conference on computer vision. Springer; 2022. p. 69–86.
21. Li K, Wang Y, Gao P, Song G, Liu Y, Li H, et al. Uniformer: Unified transformer for efficient spatiotemporal representation learning; 2022. arXiv preprint arXiv:2201.04676.
22. Hao Y, Zhou D, Wang Z, et al. PosMLP-Video: spatial and temporal relative position encoding for efficient video recognition. Int J Comput Vis. 2024;132:5820–40. https://doi.org/10.1007/s11263-024-02154-z.
23. Chen Z, Liu J, Zhu M, Woo PY, Yuan Y. Instance importance-aware graph convolutional network for 3D medical diagnosis. Med Image Anal. 2022;78: 102421.
24. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size; 2016. arXiv preprint arXiv:1602.07360.
25. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. PMLR; 2019. p. 6105–14.
26. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022. p. 11976–86.

Publisher's Note