scientific reports

OPEN

Feasibility study of AI-assisted multi-parameter MRI diagnosis of prostate cancer

Yibo Xu[1,2], Rongjiang Wang[1,2], Zhihai Fang[1,2] & Jianer Tang[1,2]✉

Distinguishing between benign and malignant prostate lesions in magnetic resonance imaging (MRI) poses challenges that affect prostate cancer screening accuracy. We propose a novel computer-aided diagnosis (CAD) system to extract cancerous lesions from the prostate in multi-parametric MRI (mp-MRI), assessing the feasibility of using artificial intelligence for detecting clinically significant prostate cancer (PCa). A retrospective study was conducted on 106 patients who underwent mp-MRI from 2021 to 2024 at a single center. We analyzed three sequences (T2W, DCE, and DWI) and collected 137 mp-MRI images corresponding to pathological sections. From these, we obtained 274 sets of ROI data, using 206 for training and validation, and 68 for testing. A feature extractor was developed using a pre-trained ResNet50 model combined with a multi-head attention mechanism to fuse modality-specific features and perform classification. The experimental results indicate that our model demonstrates high classification performance, achieving an AUC of 0.89. The PR curve shows high precision across most recall values, with an AUC of 0.91. We developed a novel CAD system based on deep learning ResNet50 models to assess the risk of prostate malignancy in mpMRI images. High classification ability is achieved, and combining the attention mechanism or optimization strategy can improve the performance of the model in medical imaging.

Keywords Prostate cancer, Multi-parameter MRI, Computer-aided diagnosis, Convolutional neural networks

Prostate cancer (PCa) is the second most common malignancy in men globally, accounting for over one-fifth of male cancer diagnoses and posing significant challenges to healthcare systems[1,2]. Early treatment can reduce mortality, prompting widespread prostate cancer screening[3]. Consequently, prostate cancer screening has been widely promoted and implemented over the past decades[2,4]. Currently, prostate biopsy is the standard diagnostic method, but its invasive nature raises concerns about detecting small tumors, leading to overtreatment and psychological distress. Thus, optimizing biopsy protocols to improve detection while minimizing unnecessary procedures has become a critical research focus, aiming to alleviate patient burden and conserve resources.

Over the past decades, researchers have developed various computer-aided diagnosis (CAD) systems to classify malignant and benign prostate lesions[5–8]. However, their effectiveness depends on accurate segmentation and optimal image selection. Until these challenges are addressed, the performance of traditional CAD systems remains questionable. In recent years, rapid advancements in deep learning technology have led to the development of novel CAD systems for prostate cancer diagnosis[7,9]. Deep learning methods can simplify the classification of prostate lesion images and automatically learn complex patterns, reducing manual intervention, provided the dataset is sufficiently large[7,10,11].

Among various deep learning models, convolutional neural networks (CNNs) are the most widely used in medical imaging due to their simplicity and effectiveness[9,12]. It has been applied to in other cancer diagnoses to classify benign and malignant cases[9,12,13]. Compared to the current popular algorithms such as U-Net or DenseNet, the ResNet50 architecture has better performance in alleviating the gradient disappearance problem and achieving deeper network training, which can better capture the detailed features in the image[14]. The U-Net model features a symmetric encoder-decoder structure with downsampling (contracting path) and upsampling (expansive path) parts[15]. It employs skip connection to combine the features of different layers and performs well on small sample datasets, which can effectively capture context information. However, it may consume more memory when processing larger images, making it less suitable for MRI image processing. The DenseNet model uses dense connections, where each layer is connected to all previous layers, resulting in a dense feature

[1]The Department of Urology, The First Affiliated Hospital of Huzhou Normal College, Huzhou 31300, Zhejiang Province, China. [2]Huzhou Key Laboratory of Precise Diagnosis and Treatment of Urinary Tumors, Huzhou 313000, Zhejiang Province, China. ✉email: 50173@zjhu.edu.cn

map[14]. The DenseNet model is widely used in image classification, object detection, and segmentation due to its ability to reduce parameters, improve gradient flow, and enhance gradient reuse. However, its high complexity makes training time-consuming and less accessible for clinicians without a background in convolutional networks. In contrast, ResNet50, a deep residual network, uses residual learning to address training challenges. By introducing skip connections, it alleviates the vanishing gradient problem, allowing deeper networks (e.g., 50 layers) without compromising performance, particularly in image classification tasks like those on the ImageNet dataset. Additionally, the multi-head attention mechanism enhances feature representation by emphasizing key details and reducing redundancy in object recognition tasks[16]. This technique has rarely been applied in developing deep transfer learning models for medical image classification, particularly for distinguishing benign from malignant prostate lesions in MRI images.

We designed a study that integrates a multi-head attention mechanism into the ResNet50 model, leveraging the strengths of both techniques to enhance CAD system performance in classifying prostate MRI lesions. This study aims to validate this hypothesis, with additional details provided in the following sections.

Background

Features were extracted for each modality (T2, DWI, and DCE) using separate pre-trained ResNet50 models. ResNet50 is a deep residual network that addresses vanishing gradients and explosions in deep neural networks through residual blocks. It consists of a 50-layer convolutional architecture with batch normalization and ReLU activation, enabling high-level feature extraction[17]. In this study, we removed the classification layers of ResNet50, retaining only the feature extraction component, which captures rich representations from input images, independent of specific classification tasks. The input data passes through a 7 × 7 convolutional layer, a 3 × 3 max pooling layer, and four residual modules with multiple convolutional layers and skip connections. This architecture enables the extraction of both local and global features, producing a 2048-dimensional feature vector that encapsulates high-level semantic information, including shape, texture, and edge features, which are essential for subsequent classification tasks. The same ResNet50 architecture is consistently used for processing T2, DWI, and DCE images.

The multi-head attention mechanism establishes associations between different modal features and captures dependencies to enhance classification performance. While traditional CNNs effectively extract features from single modalities, they may not fully utilize complementary information in multimodal data[9]. This mechanism improves the mining and fusion of information by weighting combinations of modal features.

In the specific implementation, the 2048-dimensional feature vector extracted from ResNet50 is input into the multi-head attention mechanism. The multi-head attention mechanism comprises multiple parallel attention heads, each independently calculating attention weights and weighted feature representations. For each modality's feature vector, the attention mechanism first computes the similarity with the feature vectors of other modalities to generate the attention weight matrix. These weights are then used to compute a weighted sum of the feature vectors, resulting in the weighted feature representation.

The parallel computation of multiple attention heads captures feature associations across different dimensions, generating richer and more diverse representations. The outputs of all attention heads are concatenated to form a comprehensive feature vector, containing independent information from each modality while fusing complementary information. This processed feature vector is then concatenated with the original feature vectors from the three modalities, creating a high-dimensional representation that encompasses multimodal information.

The parallel computation of multiple attention heads captures feature associations across different dimensions, generating richer representations. The outputs from all heads are concatenated to form a comprehensive feature vector, which contains independent information from each modality while fusing complementary data. This vector is then combined with the original feature vectors from the three modalities, creating a high-dimensional representation that encompasses multimodal information.

Methodology

This study adheres to the Helsinki Declaration (revised in 2013) and does not require ethical approval as the data we utilized was derived from previous image database of the hospitals. In the light of the study's retrospective nature, the requirement for informed consent was waived.

Datasets description

In this study, we collected mp-MRI images corresponding to pathological sections from 106 prostate cancer cases, including T2-weighted (T2WI) and diffusion-weighted imaging. Our inclusion criteria included: (1) undergoing prostate biopsy; (2) Multiparametric magnetic resonance imaging was performed before prostate biopsy. Exclusion criteria: (1) patients who had undergone prostate biopsy within the first 6 weeks of MRI scan; (2) after local prostate surgery, or the patient has received radiotherapy and chemotherapy; (3) Multiple metastasis of prostate cancer in pelvic cavity and obvious destruction of local glands; (4) Patients were pathologically diagnosed with bladder cancer, sarcoma, or other types of malignant tumors that had metastasize to the prostate; (5) Foreign bodies, gas, motion and other artifacts cause poor quality of MRI images, or incomplete sequences. A total of 137 mp-MRI images were gathered, resulting in 274 groups of ROI data from three sequences of prostate mp-MRI images. Of these, 206 sets are designated for training and validation, while 68 sets are reserved for testing. The dataset comprises images from three modalities, each associated with a corresponding ROI (region of interest). The dataset is divided into training, validation, and test sets. A total of 206 samples were used for the training and validation sets, divided in a 6:4 ratio: 124 for training and 82 for validation. The 68 samples in the test set are used to evaluate the model's generalization performance.

Informed written consent was obtained from the patient for the publication of this report and any accompanying images.

Datasets-prostate segmentation

After denoising the three sequences of prostate MRI images, the regions of interest (ROIs) must be further extracted for subsequent input into the neural parameters. In this study, two senior pathologists reviewed the selected pathological sections and annotated the tumor area, prostate capsule, and transition zone boundary. If the pathologists disagree on the image interpretation, they will discuss their findings to reach a consensus.

1. ITK-Snap software was utilized to open the original three sequences of prostate MRI images from selected patients (exported from DICOM format). MRI images of T2WI, DWI, and DCE sequences were collected at intervals of 5–8 mm from the tip of the prostate to the bladder neck and saved.
2. Corresponding prostate specimens were prepared as axial pathological sections at 5–8 mm intervals from the tip of the prostate to the bladder neck. Two senior pathologists independently marked the tumor area, benign area, and transition zone boundary at each level after reviewing the sections, and these annotations were stored in the computer for later use.Both the imaging and pathological sections contain intrinsic markers (e.g., stones, cysts, nodules) that assist in matching them. Validation enables easy identification of corresponding levels in the pathological and imaging sections.
3. Transparent mapping technology is used to virtually overlay the pathological sections onto the three axial MRI sequences at the same level on the computer, with corresponding benign and tumor tissue regions labeled on each sequence.

Details of the experiments

The primary goal of the experiment is to enhance the accuracy and robustness of the classification task by integrating multimodal image data. The pre-trained ResNet50 model serves as a feature extractor, capturing high-level features from the images of each modality. The Multihead Attention mechanism fuses the features from different modalities and captures their correlations. Finally, classification is performed using the Fully Connected Layer to produce predictions of benign or malignant lesions.

Neural network architecture

Neural network architecture diagram (Figs. 1–2), including model construction, training, validation, and testing.

Data preprocessing and loading

Data augmentation techniques, including random cropping, rotation, and flipping, are applied to the training data to enhance the model's robustness and generalization ability. The images are resized to a uniform dimension of 224 × 224 pixels and normalized to ensure their mean and variance align with the input requirements of the pre-trained ResNet50 model. PyTorch's DataLoader is utilized to batch load the training, validation, and test sets, thereby improving training efficiency.

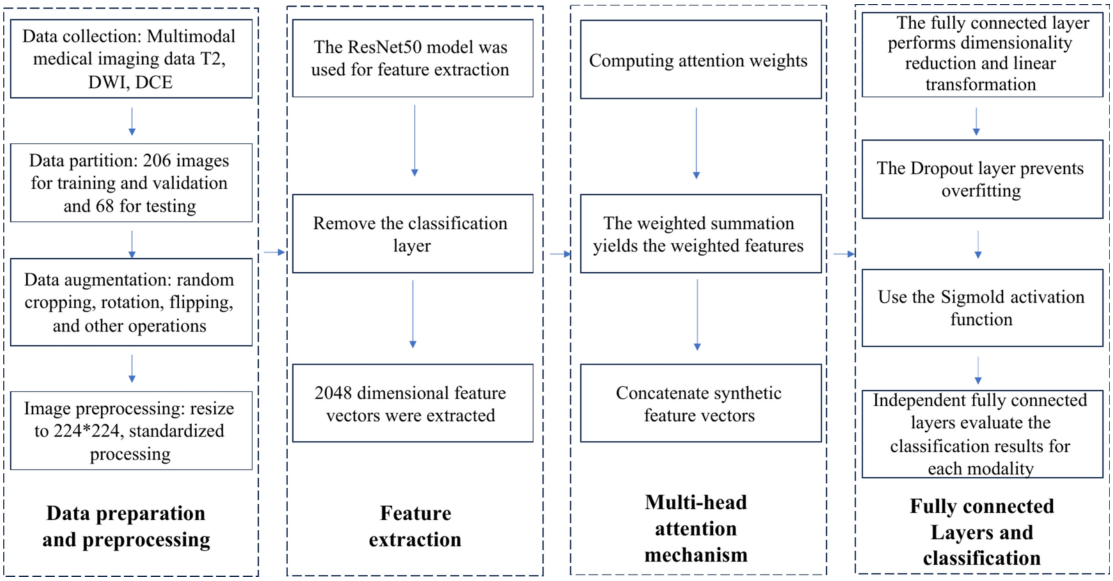
Data collection: Multimodal medical imaging data T2, DWI, DCE
Data partition: 206 images for training and validation and 68 for testing
Data augmentation: random cropping, rotation, flipping, and other operations
Image preprocessing: resize to 224*224, standardized processing
Data preparation and preprocessing
The ResNet50 model was used for feature extraction
Remove the classification layer
2048 dimensional feature vectors were extracted
Feature extraction
Computing attention weights
The weighted summation yields the weighted features
Concatenate synthetic feature vectors
Multi-head attention mechanism
The fully connected layer performs dimensionality reduction and linear transformation
The Dropout layer prevents overfitting
Use the Sigmold activation function
Independent fully connected layers evaluate the classification results for each modality
Fully connected Layers and classification

Fig. 1. Model construction (neural network architecture): data preparation and preprocessing, feature extraction, and weighted combination of different modal features through multi-head attention mechanism for classification.

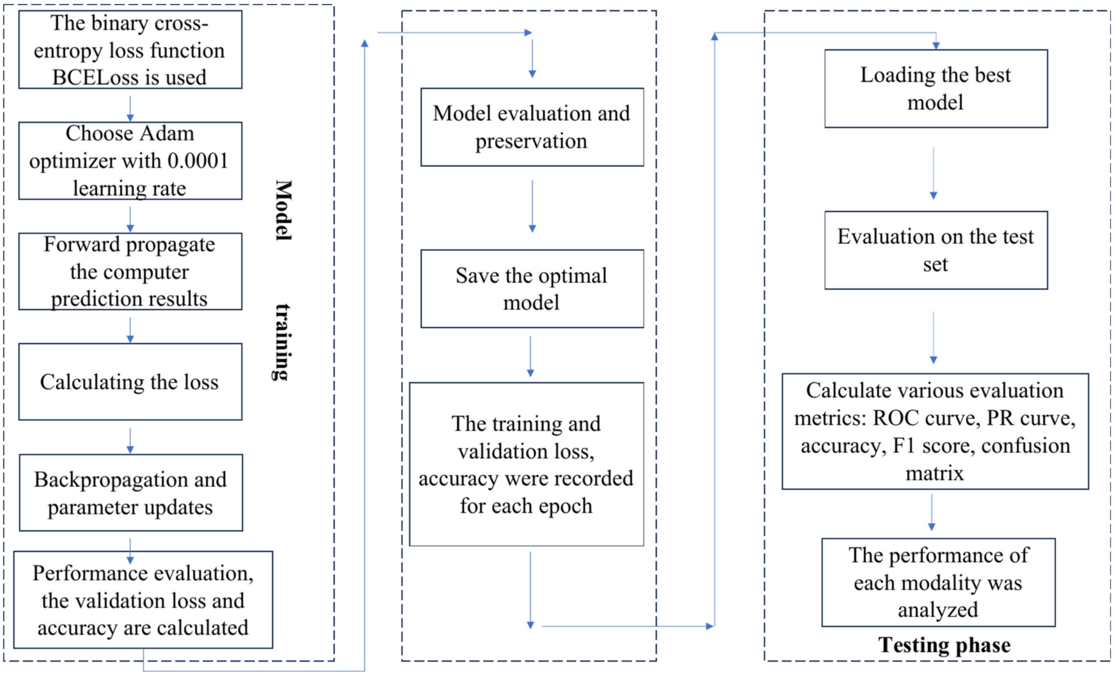
The binary cross-entropy loss function BCELoss is used
Choose Adam optimizer with 0.0001 learning rate
Forward propagate the computer prediction results
Calculating the loss
Backpropagation and parameter updates
Performance evaluation, the validation loss and accuracy are calculated
Model training
Model evaluation and preservation
Save the optimal model
The training and validation loss, accuracy were recorded for each epoch
Loading the best model
Evaluation on the test set
Calculate various evaluation metrics: ROC curve, PR curve, accuracy, F1 score, confusion matrix
The performance of each modality was analyzed
Testing phase

Fig. 2. Model testing and training: evaluate model performance through training, save the optimal model for testing, and analyze the performance of each modality.

Model training

The Binary Cross-Entropy Loss function (BCELoss), suitable for binary classification tasks, measures the difference between predicted results and true labels. The Adam optimizer is selected for its adaptive learning rate adjustment feature and quick convergence, with an initial learning rate set to 0.0001. In each epoch, the model performs forward propagation on the training data, computes the loss, backpropagates, and updates the parameters. Simultaneously, the model's performance is evaluated on the validation set, calculating the validation loss and accuracy. The training and validation loss, along with the accuracy of each modality, are recorded for each epoch.

Model evaluation and preservation

At the end of each epoch, the training and validation loss values and accuracy are recorded and saved as CSV files. If the current validation accuracy is equal to or greater than the historical best value, and the accuracy of at least two of the three modalities—T2, DWI, and DCE—has improved, the optimal model is updated.

Test steps

Load the best model

During the testing phase, the best model parameters saved during training are loaded to ensure that the best model is used for testing.

Evaluate on the test set

The model is evaluated using test set data, calculating the difference between predicted results and true labels. Various evaluation metrics are calculated on the test set, including the ROC curve, PR curve, accuracy, F1 score, and confusion matrix. The evaluation metrics for the three modalities—T2, DWI, and DCE—are calculated separately to analyze each modality's performance in the classification task.

Draw evaluation charts

Draw and save ROC curves, PR curves, and confusion matrices. The prediction results of the total model and each modality are plotted and saved separately.

Ethical approval

This study has been approved by the Ethics Committee of the First Affiliated Hospital of Huzhou Normal University (20,180,017), and all methods were conducted in accordance with relevant guidelines and regulations. This study involved the use of patients' tissue samples in the model, and all patients signed the specific informed consent forms collected.

The results of the experiment

In this section, we analyze the performance of each modality using evaluation metrics such as ROC curve, PR curve, accuracy, confusion matrix, etc. Confusion matrix is an important tool for evaluating the performance of classification models, which shows the relationship between the model prediction results and the actual labels in the form of a matrix.

In a clinical setting, the confusion matrix provides insights into the percentages of false positives and false negatives from the model. False positives occur when healthy patients are incorrectly identified as having a condition, leading to unnecessary treatments and increased psychological burden. False negatives happen when sick patients are misclassified as healthy, delaying treatment and potentially worsening their condition. Such misdiagnoses can have serious consequences, including health risks, financial burdens, reduced trust, and legal implications. Looking at the confusion matrix, we can see that our model has trouble identifying benign samples, which may be caused by unbalanced samples. We can improve this by enhancing feature selection or expanding the dataset.

The training and validation loss plots exhibit significant fluctuations during the initial training phase, particularly with several higher peaks in the validation loss. This suggests that the model may overfit or underfit the data in certain epochs . Overall, both training and validation losses exhibit a gradual downward trend; however, the fluctuations in validation loss suggest that the model's generalization ability requires further optimization (Fig. 3-a).

The figure illustrates that the validation accuracy of the DCE modality exhibits significant fluctuations during training. Accuracy improves markedly in the initial stage, followed by noticeable fluctuations and a downward trend. Overall, the validation accuracy of the DCE modality remains between 0.5 and 0.9 throughout the training process. Possible reasons for this include insufficient adaptation of the model to the DCE features or instability during training due to the complexity of the DCE data (Fig. 3-b).

The validation accuracy of the T2 modality also exhibits significant fluctuations during training, particularly in the early and middle stages. Overall, accuracy fluctuates between 0.2 and 0.7, indicating the model's unstable performance with T2 modal data. Enhancing the model's adaptation to T2 data may necessitate improved feature extraction and fusion techniques (Fig. 3-c).

The validation accuracy of the DWI modality fluctuates significantly during the initial training phase before gradually stabilizing; however, overall accuracy remains low, primarily fluctuating between 0.4 and 0.55. This suggests that the model faces challenges in handling the modal features of DWI, indicating that further optimization may be required to enhance the classification performance of DWI data(Fig. 3-d).

The ROC curve for the training set nearly completely encompasses the ideal (0, 1) point, achieving an AUC value of 1.00. This suggests that the model demonstrates an exceptionally high classification ability on the training set, allowing it to nearly perfectly distinguish between benign and malignant samples(Fig. 3-e).

The ROC curve for the validation set closely resembles that of the training set, with an AUC value near 1.00, indicating that the model's classification ability on the validation set is also robust. This demonstrates that the model performs well on both the training and validation sets, reflecting its solid discrimination ability and indicating a degree of generalization capability (Fig. 3-f).

The confusion matrix for the DCE modality indicates that all samples are classified as malignant (label 1), with none correctly identified as benign (label 0). This suggests that the model exhibits poor classification performance for the DCE modality, failing to accurately distinguish between benign and malignant samples (Fig. 4-a).

The PR curves for the DCE modality indicate that while the model achieves high precision at certain recall values, its overall performance is lacking. The AUC value of 0.99 suggests that, despite the model maintaining high accuracy with DCE data, it struggles to effectively distinguish between sample categories (Fig. 4-b).

The ROC curve for the DCE modality displays an AUC value of 0.99, approaching 1.0, which indicates the model's high classification ability on the test set. However, the results from the confusion matrix reveal that the actual classification performance is suboptimal (Fig. 4-c).

The confusion matrix for the DWI modality indicates that all 68 malignant samples were correctly classified; however, only 3 benign samples were accurately identified, while the remaining 65 were misclassified as malignant. This suggests that the model encounters significant challenges in distinguishing benign samples (Fig. 5-a).

The PR curve for the DWI modality indicates that the model maintains high precision across most recall values, with an AUC value of 0.96, suggesting that it can effectively distinguish between benign and malignant samples in most instances (Fig. 5-b).

The ROC curve for the DWI modality displays an AUC value of 0.95, indicating the model's high classification ability on the test set. However, the results from the confusion matrix reveal that the model struggles to perform effectively with benign samples (Fig. 5-c).

The confusion matrix for the T2 modality indicates that all samples are classified as malignant (label 1), with none correctly identified as benign (label 0). This resembles the results for the DCE modality, reflecting the model's very poor classification performance on T2 data (Fig. 6-a).

The PR curve for the T2 modality indicates that the model performs poorly on T2 data, with an AUC value of 0.31, suggesting that it is largely unable to distinguish between sample categories (Fig. 6-b).

The ROC curve for the T2 modality displays an AUC value of 0.01, significantly lower than the ideal value, further confirming the model's very poor classification performance on T2 data (Fig. 6-c).

The confusion matrix for the overall test reveals:

45 benign samples correctly classified (True Negatives).

23 benign samples misclassified as malignant (False Positives).

All 68 malignant samples were correctly classified (True Positives).

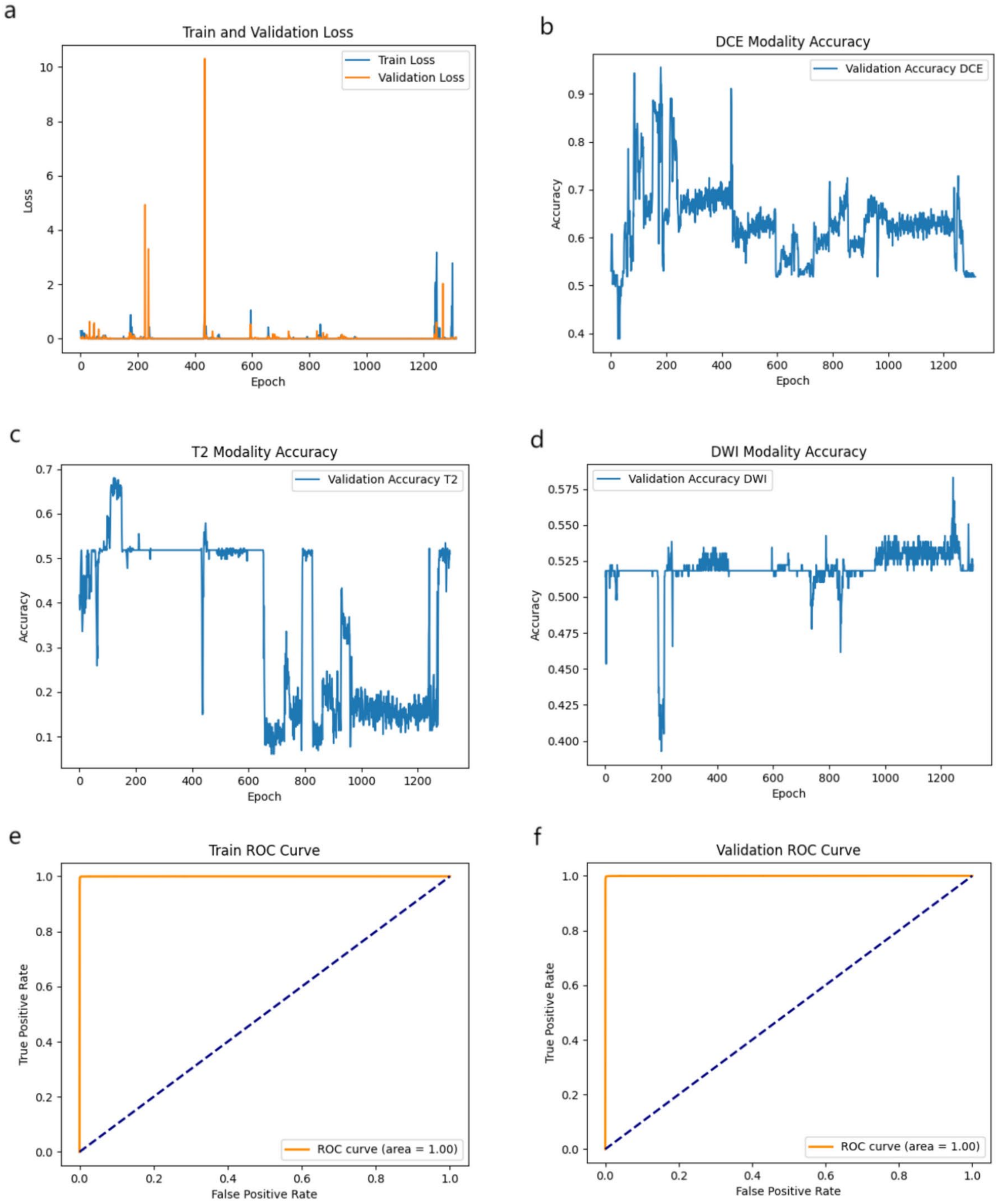

Fig. 3. (**a**) The training and validation loss plots; (**b**) The validation accuracy of the DCE modality; (**c**) The validation accuracy of the T2 modality; (**d**) The validation accuracy of the DWI modality; (**e**) ROC curve for the training set; (**f**) ROC curve for the validation set.

No malignant samples were misclassified as benign (False Negatives).

The confusion matrix indicates that the model performs exceptionally well in distinguishing malignant samples, with no malignant samples misclassified. However, the model struggles to distinguish benign samples, misclassifying 23 as malignant. This suggests that the model has some limitations in handling benign samples and may require further optimization (Fig. 7-a).

The PR curve for the overall test indicates that the model maintains high precision across most recall values, achieving an AUC value of 0.91. This suggests that the model can effectively distinguish between benign and malignant samples in the majority of cases. The decreasing trend of the PR curve suggests that precision declines as recall increases; however, overall precision remains high (Fig. 7-b).

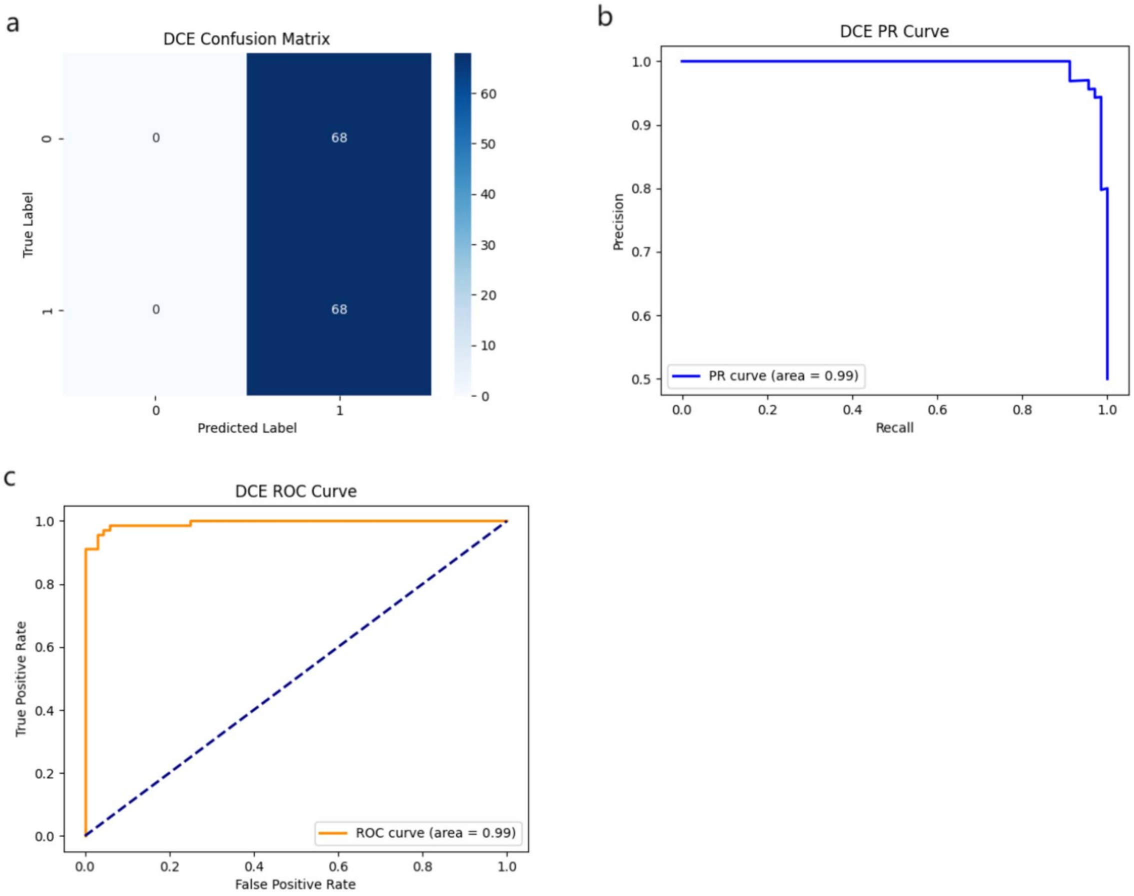

Fig. 4. The DCE modes were analyzed. (**a**) the confusion matrix suggested poor classification performance; (**b**) the PR curve showed that the overall was not ideal; (**c**) the ROC curve suggested that the actual classification performance was poor.

The ROC curve for the overall test displays an AUC value of 0.89, suggesting that the model demonstrates high classification ability on the test set. The ROC curve is positioned close to the top left, indicating that the model effectively distinguishes between positive and negative samples. However, the AUC value is below 1, suggesting that the model may still experience misclassification in certain instances (Fig. 7-c).

According to the description in Table 1, we can obtain the following information. The overall accuracy is 0.8309, indicating that 83.09% of the tested samples were correctly classified by the model. Overall accuracy is a crucial metric for assessing the model's classification performance on the entire test set, demonstrating its ability to distinguish between benign and malignant samples in most instances.

The overall F1 score, defined as the harmonic mean of precision and recall, is 0.8553. This score comprehensively reflects the model's ability to classify correctly and recognize samples from different classes, making it particularly suitable for datasets with class imbalance. The model demonstrates strong performance in balancing precision and recall.

In summary, the model's performance on the overall test set is as follows:

Classification Ability: The overall ROC curve indicates that the model has high classification ability, with an AUC value of 0.89, demonstrating its effectiveness in distinguishing between positive and negative samples on the test set.

Misclassification of Benign Samples: The confusion matrix indicates that the model struggles to distinguish benign samples, misclassifying 23 benign samples as malignant. This could result in a high false positive rate, necessitating further optimization of the model to enhance the recognition accuracy of benign samples.

Identification of Malignant Samples: The model demonstrates excellent performance in distinguishing malignant samples, accurately classifying all malignant samples, which indicates a strong ability to identify such cases.

Overall Precision and Recall: The PR curve indicates that the model maintains high precision across most recall values, achieving an AUC value of 0.91. This suggests that the model can maintain high accuracy and recall for most samples; however, precision decreases at higher recall levels.

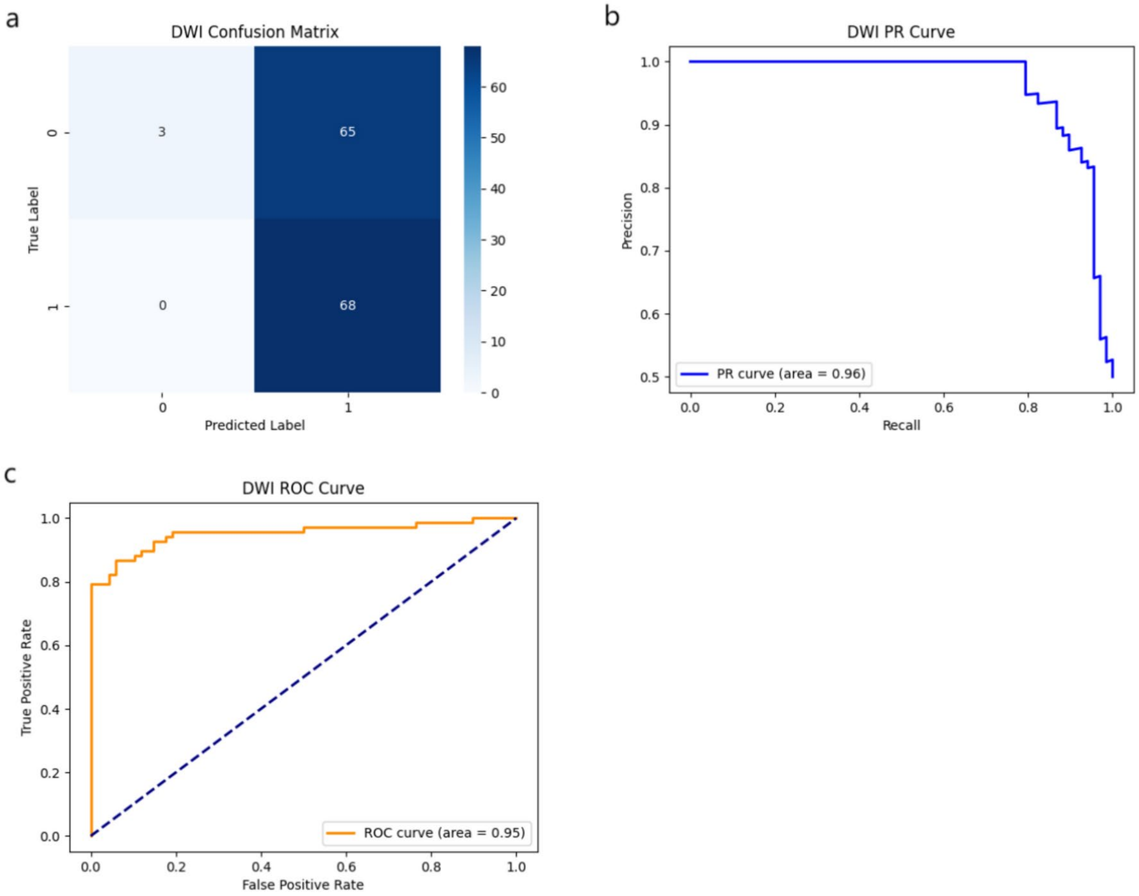

Fig. 5. The DWI modality was analyzed, (**a**) the confusion matrix showed that it was difficult to classify benign samples; (**b**) the PR curve showed that the model could effectively distinguish benign and malignant samples; (**c**) the ROC curve showed that the model had a high classification ability.

Discussion

Accurate evaluation of MRI-detected prostate cancer (PCa) is essential for early diagnosis and management, but it remains a challenge in clinical practice. This study employs a pre-trained ResNet50 model as a feature extractor, using a multi-head attention mechanism to fuse features from different modalities and capture their interrelationships. Classification is completed through a fully connected layer to predict benign or malignant lesions. Our study shows that the proposed model, integrating T2-weighted, DCE, and DWI MRI sequences, can identify interpretable and diagnostically relevant changes by analyzing regions of interest. The overall AUC of our AI model is 0.89, indicating strong classification ability. While it effectively identifies malignant samples, its performance in distinguishing benign samples is suboptimal, highlighting the need for further optimization. The findings of this study lay the groundwork for future research on using AI to detect MRI lesion changes and assess clinical relevance, ultimately improving prostate cancer diagnostic rates and reducing misdiagnoses and missed diagnoses, while providing clearer guidance for patients.

In recent years, ResNet50 has become a prominent CNN architecture for disease diagnosis. Compared to traditional deep networks, it effectively alleviates the vanishing gradient problem and enables deeper training, allowing for better feature capture in images[17]. Using the ResNet50 model directly to classify benign and malignant lesions in prostate MRI may not be the most effective approach, as prostate tumors are subtler and harder to differentiate than other cancers. To address this challenge and enhance performance, we incorporate a multi-head attention mechanism to optimize the ResNet50 model for classifying prostate cancer lesions. Compared to traditional CNNs, the multi-head attention mechanism effectively handles multi-modal data. By using a weighted combination of modal features, it better extracts and fuses complementary information across modalities[18]. We compare our model with previous AI algorithms that evaluate features from three complete MRI sequences, optimized to identify the best machine learning pipeline. Other models mainly use one or two sequences, such as T2-weighted imaging (T2WI) and diffusion-weighted imaging (DWI), whereas we innovatively incorporate DCE-MRI sequences[19]. DCE-MRI sequences can quantitatively and semi-quantitatively evaluate tumor blood flow and energy changes. The density of blood vessels around prostate cancer (PCa) tissue is twice that of normal cells, showing higher permeability and rapid enhancement, indicating an "outflow type." In contrast, early-stage tumors exhibit an “inflow type” enhancement, aiding in distinguishing prostate lesions and improving diagnostic accuracy. Compared to T2 sequences, DCE-MRI better reveals tumor location.

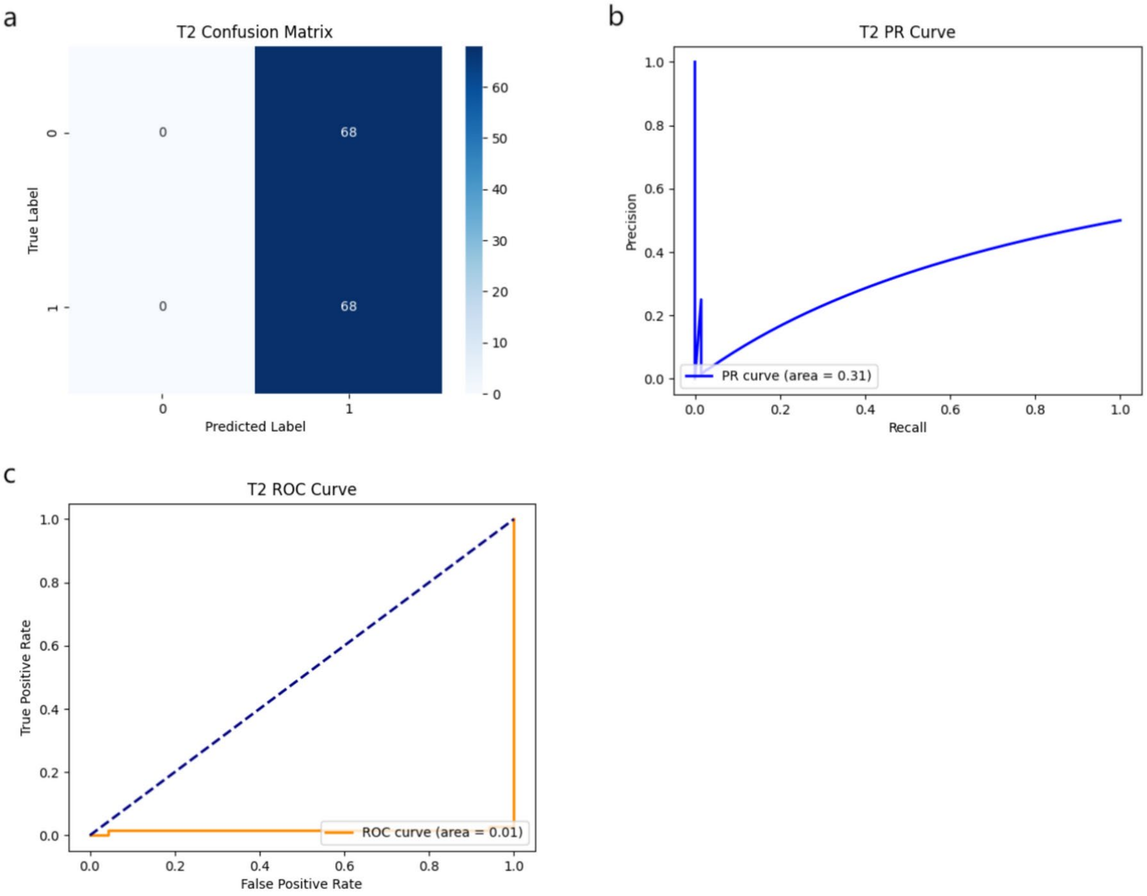

Fig. 6. The T2 modality was analyzed. (**a**) the confusion matrix showed poor classification performance; (**b**) the PR curve showed extremely poor performance; (**c**) the ROC curve indicated that the model was much lower than expected.

Additionally, unlike other studies, we use pathological slices to align imaging data. This verification allows us to easily identify corresponding levels in both pathological and imaging slices, enhancing our model's accuracy.

In terms of cancer detection performance, most reported AUC values range from 0.82 to 0.89, while our model achieves an AUC of 0.89, indicating its superiority in diagnostic accuracy compared to most other models.[8,20–23].

While we observe encouraging results, we acknowledge several limitations. The dataset used to train and validate the classifier is relatively small, necessitating larger datasets for further validation. Our single-center study lacks ethnic and geographic diversity, which may introduce bias, highlighting the need for nationwide studies to confirm generalizability. Future research will involve generating new samples through transformations to enhance diversity and improve model generalization. Additionally, we employ cross-validation to address small sample size challenges, primarily using multi-parametric MRI for prostate cancer diagnosis, achieving high sensitivity, specificity, and accuracy[23]. However, grading malignancies remains a challenge. Blessin et al. developed machine learning models for automatic grading by combining imaging features from whole slide images with the Ki-67PI marker for brain tumors[24]. Ying et al. used traditional machine learning techniques, specifically SVM, as the classifier[12]. This approach achieved high classification accuracy in grading brain tumors. Our study builds on their method by combining MRI and pathological images to enhance our understanding of tumor characteristics. Kwak et al. digitized pathological images and compared them with standardized MRI images, significantly improving machine learning efficiency and enabling accurate detection of clinically significant prostate cancer (csPCa) that may be missed with sequential MRI features alone[20]. Thirdly, our study relied on expert manual segmentation to extract radiological features from MRI scans, while Mehralivand et al. proposed utilizing the system to achieve similar performance[11]. Nonetheless, current machine learning models have not yet matched the accuracy of senior radiologists, requiring some manual calibration. In the future, we plan to optimize our model using ensemble methods, combining ResNet50 with U-Net, DenseNet, and others to enhance diagnostic accuracy. We will also employ automated methods, such as grid search and Bayesian optimization, to fine-tune hyperparameters like learning rate and batch size. Additionally, we are considering multi-center data collection in collaboration with various medical institutions to gather diverse prostate cancer imaging data, ensuring better dataset diversity and representation to enhance our study.

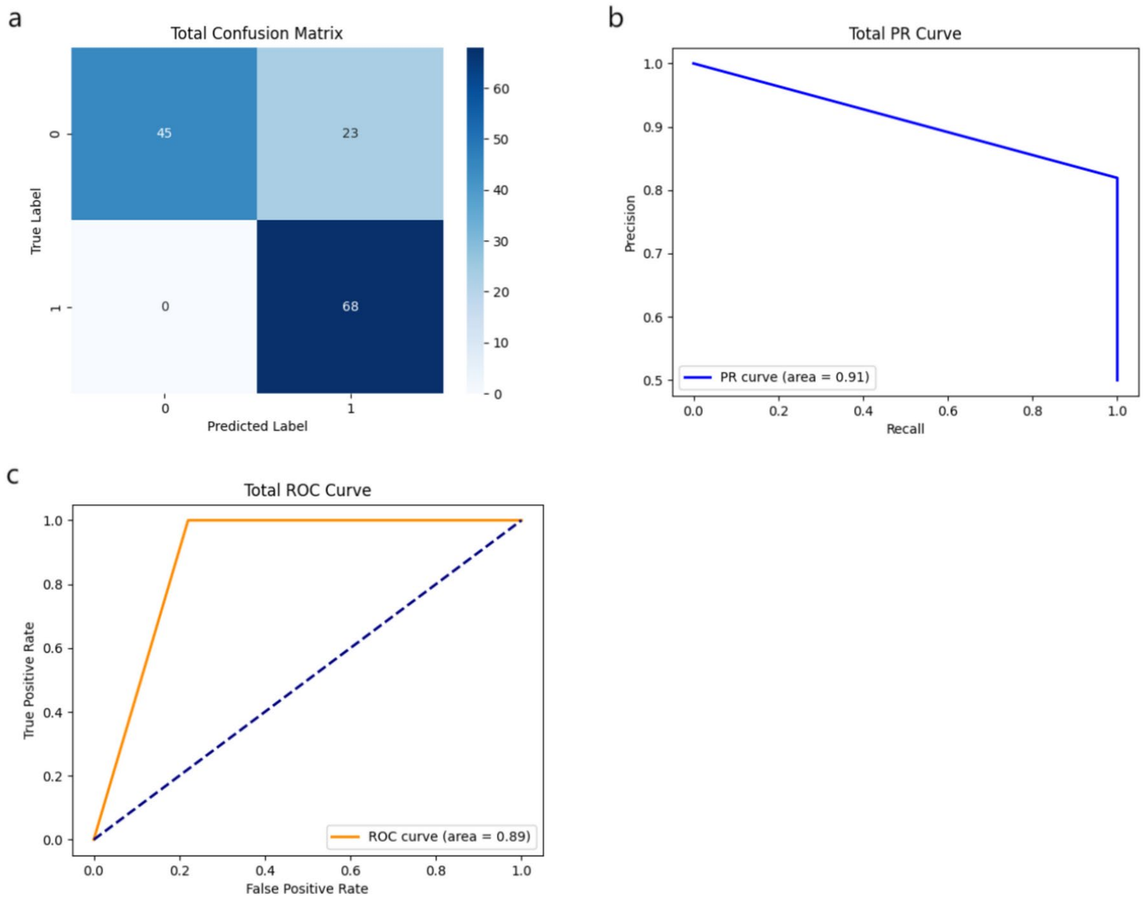

Fig. 7. To evaluate the overall test; (**a**) the confusion matrix shows that the model performs well in classifying malignant samples; (**b**) the PR curve shows that the model performs well under most recall values; (**c**) the ROC curve indicates that the model has a high classification ability.

	Total accuracy	Total F1 Score
Best	0.9926	0.9927
Final	0.8309	0.8553

Table 1. Overall accuracy and overall F1 scores.

Conclusions

This paper presents a novel study investigating the feasibility and advantages of a new CAD scheme based on a deep learning ResNet50 model, enhanced by a multi-head attention mechanism for fine-tuning. This approach aims to classify malignant and benign lesions in prostate MRI examinations. Data analysis shows that the lesion classification performance, indicated by the AUC value, significantly exceeds that of other models. Incorporating attention mechanisms or optimization strategies can enhance deep learning model performance in medical imaging.

Data availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 21 October 2024; Accepted: 24 December 2024
Published online: 27 March 2025

References

1. Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424 (2018).
2. GBD 2019 Stroke Collaborators. Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Neurol.* **20**, 795–820 (2021).

3. Sun, D. Q. et al. Quality assessment of global prostate cancer screening guidelines. *Zhonghua Liu Xing Bing Xue Za Zhi* **42**, 227–233 (2021).
4. Qi, D. et al. Trends of prostate cancer incidence and mortality in Shanghai, China from 1973 to 2009. *PROSTATE* **75**, 1662–1668 (2015).
5. Panebianco, V. et al. Multiparametric magnetic resonance imaging vs. standard care in men being evaluated for prostate cancer: a randomized study. *Urol. Oncol.* **33**, 17.e1-17.e7 (2015).
6. Moore, C. M. et al. Reporting magnetic resonance imaging in men on active surveillance for prostate cancer: The PRECISE recommendations-a report of a European School of Oncology Task Force. *Eur Urol* **71**, 648–655 (2017).
7. Reda, I. et al. Deep learning role in early diagnosis of prostate cancer. *Technol. Cancer Res. Treat* https://doi.org/10.1177/1533034618775530 (2018).
8. Antonelli, M. et al. Machine learning classifiers can predict Gleason pattern 4 prostate cancer with greater accuracy than experienced radiologists. *Eur. Radiol.* **29**, 4754–4764 (2019).
9. Zhang, Y. et al. Neural network-based approaches for biomedical relation classification: A review. *J. Biomed. Inform.* **99**, 103294 (2019).
10. Aldoj, N., Lukas, S., Dewey, M. & Penzkofer, T. Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network. *Eur. Radiol.* **30**, 1243–1253 (2020).
11. Mehralivand, S. et al. A cascaded deep learning-based artificial intelligence algorithm for automated lesion detection and classification on biparametric prostate magnetic resonance imaging. *Acad. Radiol.* **29**, 1159–1168 (2022).
12. Zhuge, Y. et al. Automated glioma grading on conventional MRI images using deep convolutional neural networks. *Med. Phys.* **47**, 3044–3053 (2020).
13. Lin, C. L. & Wu, K. C. Development of revised ResNet-50 for diabetic retinopathy detection. *BMC Bioinform.* **24**, 157 (2023).
14. Ray, I., Raipuria, G. & Singhal, N. Rethinking ImageNet pre-training for computational histopathology. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2022**, 3059–3062 (2022).
15. Rajamani, K. T., Rani, P., Siebert, H., ElagiriRamalingam, R. & Heinrich, M. P. Attention-augmented U-Net (AA-U-Net) for semantic segmentation. *Signal Image Video Process.* **17**, 981–989 (2023).
16. Islam, W. et al. Improving performance of breast lesion classification using a ResNet50 model optimized with a novel attention mechanism. *Tomography* **8**, 2411–2425 (2022).
17. Islam, R., Imran, A. & Rabbi, M. F. Prostate cancer detection from MRI using efficient feature extraction with transfer learning. *Prostate Cancer* **2024**, 1588891 (2024).
18. Yang, J. Y., Lee, T. C., Liao, W. T. & Hsu, C. C. Multi-head self-attention mechanism enabled individualized hemoglobin prediction and treatment recommendation systems in anemia management for hemodialysis patients. *Heliyon* **9**, e12613 (2023).
19. Roest, C. et al. AI-assisted biparametric MRI surveillance of prostate cancer: feasibility study. *Eur. Radiol.* **33**, 89–96 (2023).
20. Kwak, J. T. et al. Automated prostate cancer detection using T2-weighted and high-b-value diffusion-weighted magnetic resonance imaging. *Med. Phys.* **42**, 2368–2378 (2015).
21. Wibmer, A. G., Vargas, H. A. & Hricak, H. Role of MRI in the diagnosis and management of prostate cancer. *Future Oncol.* **11**, 2757–2766 (2015).
22. Akatsuka, J. et al. Illuminating clues of cancer buried in prostate MR Image: Deep learning and expert approaches. *Biomolecules* **9**, 673 (2019).
23. Stabile, A. et al. Multiparametric MRI for prostate cancer diagnosis: Current status and future directions. *Nat. Rev. Urol.* **17**, 41–61 (2020).
24. Blessin, N. C. et al. Automated Ki-67 labeling index assessment in prostate cancer using artificial intelligence and multiplex fluorescence immunohistochemistry. *J. Pathol.* **260**, 5–16 (2023).

Author contributions

Contributions YX and JT contributed to the conception and design of the review. YX wrote the manuscript. ZF validate the manuscript. RW contribute to the establishment of the model. All authors have read and agreed to the published version of the manuscript.

Funding

The authors declare that no funding was received.

Declarations

Competing interests

The authors declare no competing interests.

Informed consent

Informed written consent was obtained from the patient for the publication of this report and any accompanying images.

Additional information

Correspondence and requests for materials should be addressed to J.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.