scientific reports

Check for updates

OPEN

Bladder lesion detection using EfficientNet and hybrid attention transformer through attention transformation

Poonam Sharma[1], Bhisham Sharma[2]✉, Dhirendra Prasad Yadav[3], Deepti Thakral[4] & Julian L. Webber[5]

Bladder cancer diagnosis is a challenging task because of its intricacy and variation of tumor features. Moreover, morphological similarities of the cancerous cells make manual diagnosis time-consuming. Recently, machine learning and deep learning methods have been utilized to diagnose bladder cancer. However, manual feature requirements for machine learning and the high volume of data for deep learning make them less reliable for real-time application. This study developed a hybrid model using CNN (Convolutional Neural Network) and less attention-based ViT (Vision Transformer) for bladder lesion diagnosis. Our hybrid model contains two blocks of the inceptionV3 to extract spatial features. Furthermore, the global co-relation of the features is achieved using hybrid attention modules incorporated in the ViT encoder. The experimental evaluation of the model on a dataset consisting of 17,540 endoscopic images achieved an average accuracy, precision and F1-score of 97.73%, 97.21% and 96.86%, respectively, using a 5-fold cross-validation strategy. We compared the results of the proposed method with CNN and ViT-based methods under the same experimental condition, and we achieved much better performance than our counterparts.

Keywords Bladder Cancer, Deep learning, Transformer encoder, Self-Attention, Hybrid attention, Vision transformer, InceptionV3

Bladder cancer is a neoplastic disease that arises in the bladder tissues, usually from the urothelial cells lining the inner surface of the organ. It is categorized into subtypes like High-Grade Carcinoma (HGC), Low-Grade Carcinoma (LGC), No Tumor Lesion (NTL), and Non-Suspicion Tissue (NST), based on histopathological characteristics[1]. Bladder cancer is the 10th most frequent cancer globally and around 570,000 new cases and 210,000 deaths in year 2022 reported by the World Health Organization (WHO)[2]. Numerous pathological conditions can impact the structure and function of bladder tissue. More than 90% of bladder cancers are of the histological type known as urothelial carcinoma[3]. Due to the bladder cancer Men are more affected than women[4,5]. Therefore, early detection and classification of bladder lesions are critical for efficient treatment and better patient outcomes. However, standard diagnostic tools like cystoscopy and histopathological examination are labor-intensive, subjective, and susceptible to inter-observer variability, thus less suitable for large-scale or real-time screening[6–8].

Over the last few years, Artificial Intelligence (AI), and more specifically Deep Learning (DL), has been a potential solution for the automated o bladder cancer diagnosis tool using medical imaging data[9–11]. CNNs have shown great strength in local spatial feature extraction, while ViTs are good at capturing long-range dependencies and global contextual information[12]. For example, classifiers based on deep learning have proven to be highly accurate for identifying anomalies in facial features, detecting different skin lesion types, diagnosing novel infectious diseases and analyzing intricate patterns[13]. The incorporation of ViT architectures into standard CNN frameworks has also enhanced model interpretability and performance, rendering them a potential path for future computer-aided diagnosis system improvement[14]. Nevertheless, CNNs miss the edge and boundary

[1]Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab 140401, India. [2]Centre of Research Impact and Outcome, Chitkara University, Rajpura, Punjab 140401, India. [3]Department of Computer Engineering & Applications, G.L.A. University, Mathura, U.P, India. [4]Department of Computer Science and Technology, Manav Rachna University, Faridabad, India. [5]Department of Electronics and Communication Engineering, Kuwait College of Science and Technology (KCST), Doha Area, 7th Ring Road, Kuwait City 13133, Kuwait. ✉email: Bhisham.pec@gmail.com

region, while ViTs are computationally expensive due to attention mechanism and application in clinical environments less scalable[15].

In this study, we designed a hybrid deep learning model that combines a ViT and modified InceptionV3 to diagnose bladder cancer in endoscopic images. Our model contains hybrid attention in the ViT encoder for global feature correlation and two blocks of the InceptionV3 to capture spatial information. In addition, a Grad-CAM is included in the model to ensure the inerrability of the decision process. Furthermore, the results of the proposed method are better compared to state-of-the-art (SOTA) approaches to endoscopic image datasets.

The main contribution of the study is as follows.

(1) We introduce BCHTNet, a hybrid deep neural network that couples modified InceptionV3 for local feature extraction and a Transformer encoder for global context modelling.
(2) In the ViT encoder, a hybrid attention mechanism is designed to improve global contextual information using a more efficient row-column mechanism.
(3) In the model, we included a Grad-CAM to present the model's qualitative interpretability and diagnostic reliability.
(4) We compared BCHTNet results with six state-of-the-art models on a large-scale endoscopic bladder image dataset, under identical experimental conditions.

The remaining sections of the manuscript are organized as follows; Sect. 2 provides an in-depth discussion of the approaches proposed earlier. While Sect. 3 outlines the proposed strategy, Sects. 4 and 5 details and discuss the quantitative results, and Sect. 6 presents the inference of the paper and potential areas for future research.

Literature review

Jiao et al.[16] utilized deep learning techniques to predict HER2 expression status from H&E-stained pathological images of bladder cancer. The corresponding metrics value for the test set were 77.8% F1 score, 0.88 AUC, 0.67 accuracy, 0.56 sensitivity, and 0.75 specificity. Additionally, model was found to statistically outperform pathologists with a p-value less than 0.05, indicating that it has the potential for high diagnostic accuracy. In similar research[17], deep learning, radiomics, and RNA sequencing data were utilized to predict the stage of bladder cancer. Their residual neural network extracts high dimensional spatial features from the CT-scan images. The hybrid model exhibited a high predictive accuracy by differentiating the stages of bladder cancer with an AUC of 0.92. In another research[18] concerning the expression profile of mitochondria-related genes, the authors designed a diagnostic model using machine learning algorithms that detect bladder cancer. Their Support Vector Machine (SVM) algorithm achieved an AUC of 0.95. In addition, discrimination between bladder cancer and normal samples was indicated by sensitivity and specificity values of 92% and 90%, respectively. A novel diagnosis approach[19] was designed for bladder cancer in the early stages by combining the power of machine learning algorithms with Surface-Enhanced Raman Spectroscopy (SERS) in a rat model. The authors demonstrated several classifiers for distinguishing between healthy samples and samples affected by cancerous conditions. The SVM classifier performed the best with an accuracy of 95%, sensitivity of 93%, specificity of 97%, and an AUC of 0.98, showing outstanding differentiation between healthy samples and bladder cancer in its early stages.

Luo et al.[20] developed a Multiview Multi-Scale Graph Attention Network (MVMSGAT) model to predict how bladder cancer patients react to neoadjuvant therapy. The GEO datasets, containing 210 samples, obtained from gene expression profiles of patients. Their MVMSGAT model obtained an AUC value of 0.92. The study[21] adopted a hybrid approach where the feature set was obtained from a pre-trained XceptionNet. After that, Linear Discriminant Analysis (LDA) was used to classify features. The LDA and XceptionNet-based model achieved an F1-score of 89.39% in distinguishing between the cancerous and healthy tissues, F1-score 70.81% in distinguishing between Muscle-Invasive Bladder Cancer (MIBC) and Non-Muscle-Invasive Bladder Cancer (NMIBC), F1-score 74.73% in discriminating between Post-Treatment Changes (PTC) and MIBC, indicating a potential to help evaluate chemotherapy response and recurrence. Lee at al[22] developed a CNN models to classify bladder tumors in cystoscopy images. They compared the results of their models against human experts. Their CNN models outperform compared to human experts. The ResNet50 architecture achieved the best performance with 92% accuracy, 90% sensitivity, 94% specificity, and an AUC of 0.96. Yue et al.[23] to improve bladder tumor diagnosis by incorporating logical clinical knowledge into deep neural networks. The study used an MRI image dataset of the bladder. The precision and recall rates reported in the study were 0.85 and 0.88, respectively, which demonstrated the model's effectiveness in precisely identifying tumor regions.

Khedr et al.[24] performed comparative study of ViT_B32 and ViT_B16 on a bladder cancer dataset. The ViT_B32 and ViT_B16 achieved an accuracy of 99.23% and 99.49% respectively. Using histopathological images, the Shalata et al.[25] designed a multi-scale pyramidal CNN to grade Non-Muscle Invasive Bladder Cancer (NMIBC) accurately. Histopathological images containing various grades of cancer severities were obtained from NMIBC tissue samples. Their model achieved 94.29% F1-score, 94.47% sensitivity, 94.03% specificity, and 94.25% accuracy on test data. Recent studies[26] reported that artificial intelligence can improve speed and accuracy of diagnosis. They developed an algorithm and trained on 925 images. Their model achieved a sensitivity of 72% for muscularis propria and 65% for tumors. Yang et al.[27] performed a comparative study using LeNet, AlexNet and GoogLeNet on the cystoscopic images and obtained the highest accuracy of 96.9%. In another research, using two machine-learning algorithms, a label-free technique is used to detect bladder cancer cells in urine samples. Their method reported that cells were classified with 99% accuracy and 97% AUC for cell lines. The gradient boosting algorithm obtained 95% accuracy and 93% AUC for urine samples, whereas the deep-learning algorithm obtained 96% accuracy and 96% AUC[28]. To aid the preoperative diagnosis of both MIBC and NMIBC, the research[29] proposes a combination models strategy using multi-parametric MRI. The approach

outperformed urologists and matched senior radiologists, achieving an accuracy of 0.869 and an AUC of 0.928 in the internal cohort after being tested on 436 patients using five-fold cross-validation. These results suggest that CMS helps junior clinicians diagnose MIBC before surgery. Jiang et al.[30] presents a urine-based DNA methylation diagnostic panel for detecting bladder cancer (BC). Their method uses the decision tree algorithm and obtained a specificity of 90.9%.

Despite significant advances in bladder cancer detection with deep learning techniques, the real-time robust model is highly demanding. These models, although successful, tend to be based on high computational complexity or multimodal data that restricts their real-time use due to a lack of global correlation of the spatial features. In addition, fewer methods use endoscopic images for classification into clinically meaningful classes such as High-Grade Cancer (HGC), Low-Grade Cancer (LGC), Tumor Lesion (NTL), and Non-Suspicious Tissue (NST). Moreover, several methods cannot balance spatial and global contextual information while being computationally efficient. Thus, there is a need to build lightweight, hybrid transformer models for accurate endoscopic image-based classification with less computational burden and provide interpretability for practical deployment.

Methodology

In the proposed study, we designed a Bladder Cancer Hybrid Transformer Network (BCHTNet) for the bladder disease diagnosis, shown in the Fig. 1. The BCHTNet contains modified InceptionV3 blocks to extract high-dimension spatial features. Furthermore, a hybrid attention module is integrated into the ViT encoder to provide local and global attention. The hybrid attention model provides attention to the feature map using row-wise and column-wise to the spatial features. Moreover, an attention transformation module is available in the hybrid attention to reduce the computation burden and saturation.

Inception module

Let the input image $I \in R^{H \times W \times C}$ is passed to the Inception module to extract the spatial features from the bladder lesion. In the Inception perform convolution using factorized convolutions to improve the spatial feature map. The mathematical the convolution operation is defined as follows.

$$O(q,r,s) = \sum_{x} \sum_{y} \sum_{z} a(q+x, r+y, z).w(x,y,z,s) \tag{1}$$

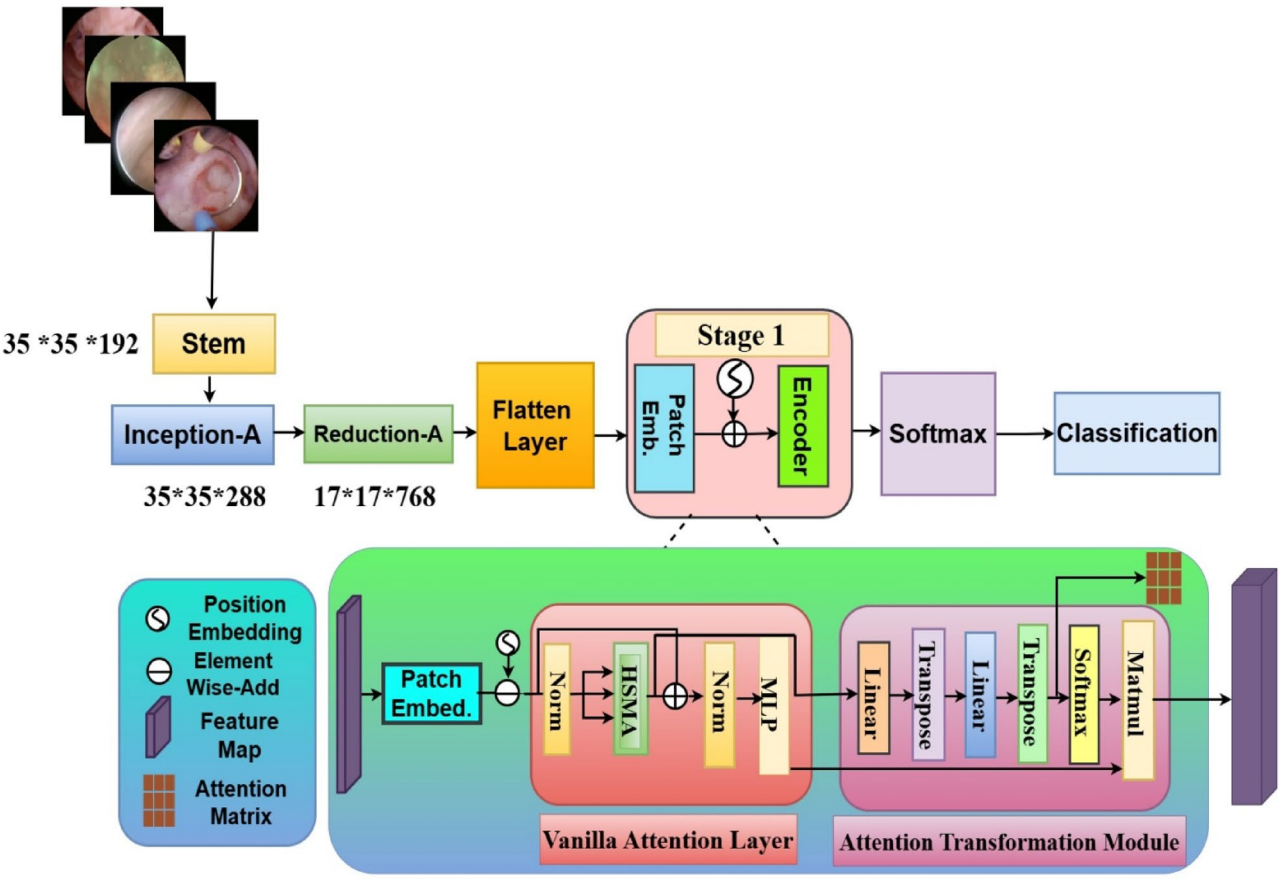


Fig. 1. The BCHTNet for bladder lession diagnosis.

Where, a = input tensor of shape Height(H), Width(W), Channels(C), w = Kernel Tensor of shape (K_H, K_W, C, C^o) , O = Output after convolution, q, r = Spatial indices, s = Output channel and x, y, z = Kernel height, width and input channels. For computational efficiency, in the Inception V3 we substitute combinations of smaller kernels (such as 1 × 5 followed by 5 × 1) and mathematically expressed it as follows.

$$O(q,r,s) = \sum {}_x a(q+x,r,z).w(x,z,s) + \sum {}_y a(q,r+y,z).w(y,z,s) \tag{2}$$

By doing this, the computation complexity decreased from K^2C to $2KC$, where K is the kernel size. In the inception module we applied parallel filters (such as 1 × 1, 3 × 3, and 5 × 5) and concatenating the outputs as in Eq. (3), each module carries out multi-scale processing as follows.

$$O = Concat(O_{1\times 1}, O_{3\times 3}, O_{5\times 5}, O_{pool}) \tag{3}$$

$O_{1\times 1}$ = Output of Convolution 1 × 1, $O_{3\times 3}$ = Output of Convolution 3 × 3, $O_{5\times 5}$ = Output of Convolution 5 × 5 and O_{pool} = Output of pooling. This multi-dimensional tensor will subsequently be transformed into a single long vector by the Flatten layer by using the Eq. (4).

$$Y_{flatten} = Flatten([O_{1\times 1}, O_{3\times 3}, O_{5\times 5}, O_{pool}]) \tag{4}$$

We reshape the flattened output ($Y_{flatten}$) using Eq. 5, back into a 2D spatial patches as follows.

$$X = Reshape(Y_{flatten}, (N, P_h, P_w, C)) \tag{5}$$

Where P_h is height of each patch, P_w is width of each patch, C is number of channels and N is number patches.

The hybrid attention

We designed a hybrid attention module to enhance transformers' efficiency, especially when handling high-dimensional input, reducing the quadratic complexity of conventional attention mechanisms. The hybrid attention block calculates attention along specific axes instead of calculating complete pairwise attention along one axis (rows or columns) at a time. This makes it more effective by lowering the computational complexity. The tokens generated from the feature map obtained from the Inception block are used to compute the attention independently along N tokens. For the input feature map $X \in R^{H\times W\times d}$,where H, W is height and width, d is channel dimension. In the row-wise attention, columns separate tokens and calculate attention separately for each row, as shown in Eq. (6).

Let X_h represent the hth row of X, Query, key and Value of the tokens are calculated as follows.

$$Q_h = X_h, W_Q,\ K_h = X_h, W_K,\ V_h = X_h, W_V$$

After, calculation of Q, K, and V for each head attention is defined as follows.

$$\text{Attention row}(Q_h, K_h, V_h) = \text{Softmax}\left(\frac{Q_h \times K_h}{\sqrt{d_k}}\right) V_h \tag{6}$$

Furthermore, column-wise attention treats rows as separate tokens and calculates attention separately for each column as shown in Eq. (7). Where X_w represents w[th] column of X, and query, key, Values are as follows.

$$Q_w = X_h, W_Q,\ K_w = X_w, W_K,\ V_w = X_w, W_V$$

After, calculation of Q, K, and V for each head attention is defined as follows.

$$\text{Attention column}(Q_w, K_w, V_w) = \text{Softmax}\left(\frac{Q_w \times K_w}{\sqrt{d_k}}\right) V_w \tag{7}$$

Row-wise and column-wise attention are applied successively to produce the final Attention result as shown in Eq. (8).

$$X'_{row} = Attention_{row}(Q,K,W),\ X'_{col} = Attention_{column}(Q,K,W) \tag{8}$$

Finally, the results of the column-wise and row-wise attention are concatenated to produce hybrid attention using Eq. (9).

$$X' = Combine\ \left(X'_{row}, X'_{col}\right) \tag{9}$$

For large sequences, calculation of global attention across rows and columns can be computationally costly. Hence attention transformation mechanism on the attention map computed by the model. The overall attention transformation is defined as follows.

$$\widetilde{X} = X' W_L, Attn = Softmax(\widetilde{X}_T)X'\ \ F_{final} = Attn \cdot X' \tag{10}$$

Where, $X' \in R^{N \times D}$= The feature of N tokens of dimension D, $W_L \in R^{D \times d}, d < D$= Linearly reduce the attention map to lower dimension (d) using linear transform and transpose operation and $F_{final} \in R^{N \times d}$= Final projected feature map .After that, we normalize the outputs to reflect probability distributions and apply Softmax to convert logits into corresponding class using Eq. (11).

$$F_{out} = Softmax(W_c F_{final} + b_0) \tag{11}$$

Where, W_c= Classification weight and b_0=Bias. We computed loss of the model using the categorical cross-entropy loss function defined in Eq. (12).

$$L = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_{ij} \log(\widehat{y}_{ij}) \tag{12}$$

Where, N= Number of samples, C= Number of classes, Y_{ij}= Truth value of i_{th} value of j_{th} class, $\widehat{y}_{ij}$= Predicted values of i_{th} value of j_{th} class corresponding to F_{out}.

The suggested method algorithm is as follows.

Input: Image $I \in R^{H \times W \times C}$

1) Resize image to 300x300 pixels and set BS=32, LR=0.0001, Epochs=160
2) Calculate spatial features using Equation 3
3) Create a 1D vector by flattening the concatenated output O:
$$Y_{flatten} = Flatten([O_{1\times1}, O_{3\times3}, O_{5\times5}, O_{pool}])$$
4) To reshape the flattened output (Y) into patches, then restructure it back into a multi-dimensional tensor, similar to 2D spatial patches.
$$X = Reshape(Y_{flatten}, (N, P_h, P_w, C))$$
Where P_h is height of each patch, P_w is width of each patch, C is number of channels and N is number patches.
5) Measure row wise attention:
$$\text{Attention}_{\text{row}}(Q_h, K_h, V_h) = \text{Softmax}(\frac{Q_h \times K_h}{\sqrt{d_k}})V_h$$
Let h represent the h^{th} row of X. where Q_h, K_h, V_h are query, key, value projections respectively and d_k is featuring dimension.
6) Measure column wise attention:
$$\text{Attention}_{\text{column}}(Q_w, K_w, V_w) = \text{Softmax}(\frac{Q_w \times K_w}{\sqrt{d_k}})V_w$$
Where w represents w^{th} column of X, and query, key, Values projections and d_k.
7) Combine the attention results by row and column:
$$X' = \text{Combine}(X'_{\text{row}}, X'_{\text{col}})$$
Where X'row=Attention row (Q, K, W), X' col=Attention column (Q, K, W)
8) Apply an attention transformation mechanism using Equation 10

Output: Class label

Algorithm 1. The proposed model for Image Classification.

Results

This section presents the results of the proposed methods on the endoscopic image dataset.

Dataset description

The dataset contains 1,754 endoscopic images from 23 patients who underwent Trans-Urethral Resection of Bladder Tumor (TURBT)[31]. White Light Imaging (WLI) is used to capture images along with Narrow Band Imaging (NBI); after that, it is labelled according to histopathology analysis. Furthermore, the dataset is categorized into Non-Suspicious Tissue (NST), High-Grade Cancer (HGC), Low-Grade Cancer (LGC), and No Tumor Lesion (NTL). Considering the overfitting problem, we augmented the dataset and 17,540 images were used to evaluate model performance. In the proposed study, we chose to perform 5-fold cross-validation rather than a standard 80 – 20 train-test split to obtain a stronger and more generalizable assessment of the model's performance as dataset contains unequal number of images in each class. Cross-validation moderates this problem by taking the average of the performance over several folds, and hence lowering the variance of the performance measures and preventing any single-data partition bias. Furthermore, the dataset four classes are split using a stratified 5-fold cross-validation approach. Thus, in each fold, ~ 80% of the data (~ 14,032 images) was used for training and ~ 20% (~ 3,508 images) for validation. Summary of images in each class of the dataset is presented in Table 1.

Sr.No.	Class Name	Images
1.	HGC	469
2.	LGC	647
3.	NST	504
4.	NTL	134
Total original images		1754

Table 1. Summary of images in each class.

Experimental settings

We evaluated the BCHTNet on NVIDIA QUADRO RTX-4000 GPU having 128 GB RAM and a dual graphics card of 8GB. Furthermore, the script is written using Python 3.10. The model is trained for 160 epochs in a batch size of 32 using an ADAM optimizer with an initial learning rate of 0.0001. Moreover, we adopted a 5-fold cross-validation technique to avoid bias performance and ensure broad generalizability across various subsets created from the used dataset.

Quantitative results

We resized the images to $300 \times 300 \times 3$ pixels and fed to the model for training using a 5-fold cross validation scheme for 160 epochs. After that for each epoch's confusion matrix is plotted shown in the Fig. 2. In fold 1, our model has 59 false positive and 47 false negative samples. At the same time, fold 2 has 52 false positives and 42 false negatives. Furthermore, fold 3 has 43 false positive and 32 false negative values. In addition, the model obtained 38 false positive and 28 false negative values. Moreover, in fold 5, there are 29 false positive and 28 false negative values.

From the confusion matrices shown in Fig. 2, we calculated performance metrics over five folds shown in Table 2. The performance metrics Kappa, Recall, Precision, F1-score, and Accuracy values are high in all folds. Our model Kappa scores ranging from 0.957 to 0.962,

The range of recall values is 0.9558 to 0.9768, which indicates the percentage of true positive cases that were correctly identified. These high scores imply that the model minimizes false negatives by accurately identifying bladder cancer cases. The model's ability to prevent false positives is further evidenced by precision values, which show the percentage of accurate positive predictions ranging from 0.9624 to 0.9798. The F1-score, which ranges from 0.9582 to 0.9783 and is a balanced indicator of Precision and Recall, is continuously high across all folds.

This suggests that the model is dependable for clinical use because it consistently balances sensitivity and specificity. Accuracy scores range from 0.9697 to 0.9837, which gauges the classifier's overall correctness is remarkably high. This suggests that the model accurately classifies both bladder cancer and non-cancer cases. Overall, the results reveal that the model is quite good in classifying bladder cancer, with very few false positives and negatives and works well across folds of data. Such results indicate that the model is consistent, trustworthy, and, therefore, a good contender for practical use.

Discussion

Bladder cancer early diagnosis is essential to save patient life. The manual cancer detection techniques are time-consuming and require experts. However, ML and DL are widely used to automate and accelerate the diagnosis process. Table 3 compares numerous deep-learning models applied to diagnose bladder cancer using multiple imaging modalities, datasets, and optimization strategies. Each research study has a different model design, including CNNs, ViTs, hybrid transformer models, and Federated Learning techniques.

Hwang et al.[32] applied VGG19 for the 8566 endoscopic images and achieved a classification accuracy of 91.20%. Zhang et al.[33] implemented a Multistage Feature Fusion Network (MSFF) on an endoscopic dataset and achieved an accuracy of 95.17%. Comparatively, Lazo et al.[34] GAN model obtained 90.00% accuracy. Yoo et al.[35] applied ResNeXt-101 to classify cystoscopic images and achieved an accuracy of 94.10%. Liang et al.[36] applied logistic regression and classified bladder cancer with an AUC value of 89%. Alkhalidy et al.[37] utilized the XDL ensemble deep learning model and achieved 95% accuracy. Ye et al.[38] utilized HRNetV2 and obtained an accuracy of 91.30%. El-Atier et al.[39] designed an ensemble-based method and obtained an accuracy of 95%. The BCHTNet utilized a hybrid approach and obtained an accuracy of 97.73%, making it competitive against state-of-the-art techniques and suitable for robust clinical deployment.

Performance comparison

For a fair comparison, we evaluated BCHTNet, ResNet-50[40], Inception V3[41], MobileNetV3[42], YOLOV9[43], ViT[44] and CellViT[45] under the same experimental condition discussed in Sect. 4.2. Table 4 shows that our model performs the best across all metrics, having a Kappa 96.1%, Precision 97.21%, Recall 96.6%, F1-score 96.86%, and Accuracy 97.73%. These results illustrate the model's high accuracy for case classification with minimal false positives and false negatives.

CellViT also performs well with 96.87% accuracy, reflecting an acceptable trade-off between precision and recall. While, it takes more time to train due to complex attention mechanism and lags behind BCHTNet in critical metrics. At the same time, ViT performs well in capturing global dependencies and achieves a moderate accuracy of 95.25%, but it requires large training sets and high computation costs of MHSA (multi-head self-attention). Moreover, YOLOV9 is known for its high speed of inference, making it suitable for real-time use. While having a moderate accuracy of 93.70%, though, it demonstrates poor precision and recall compared to

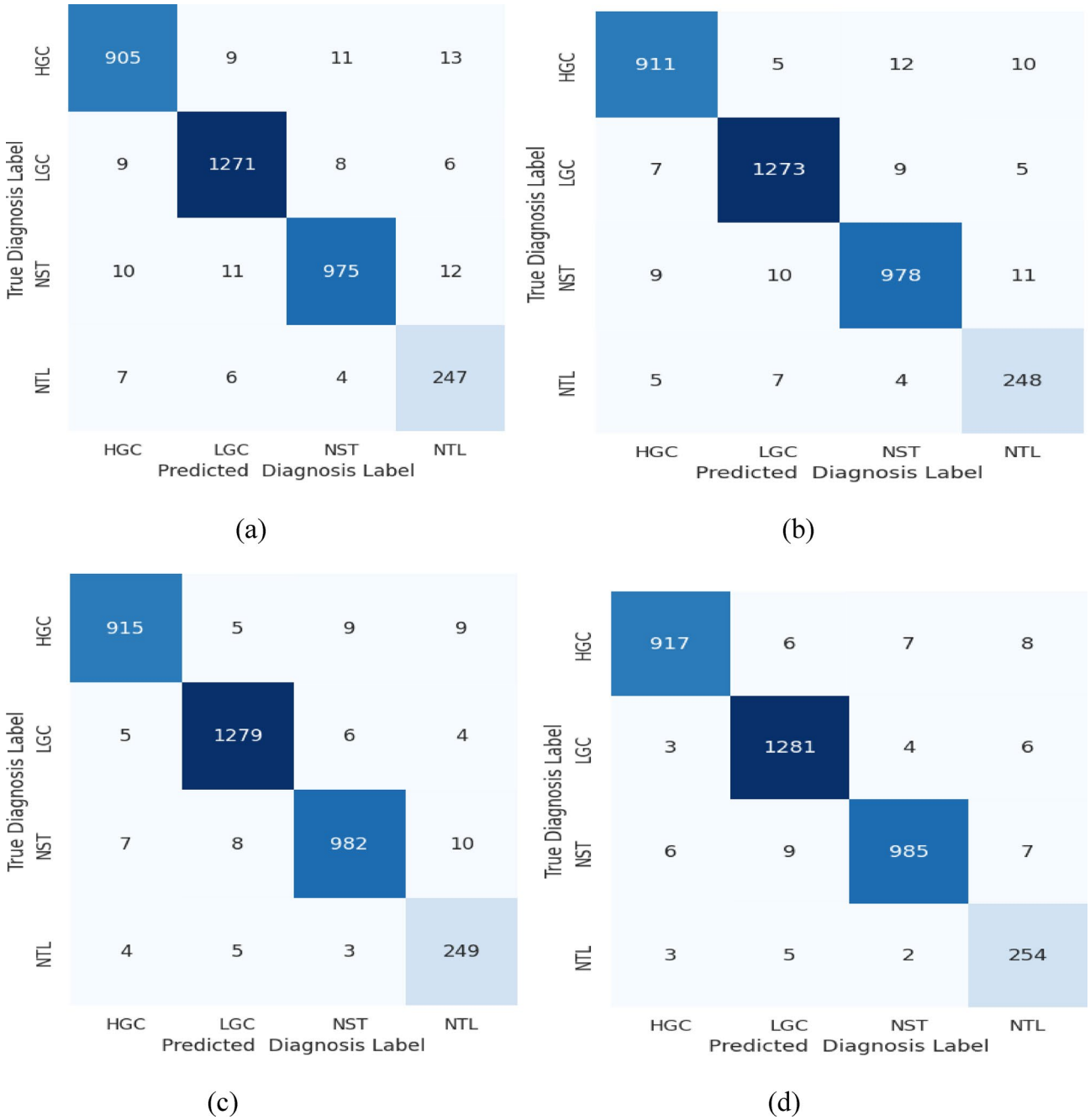
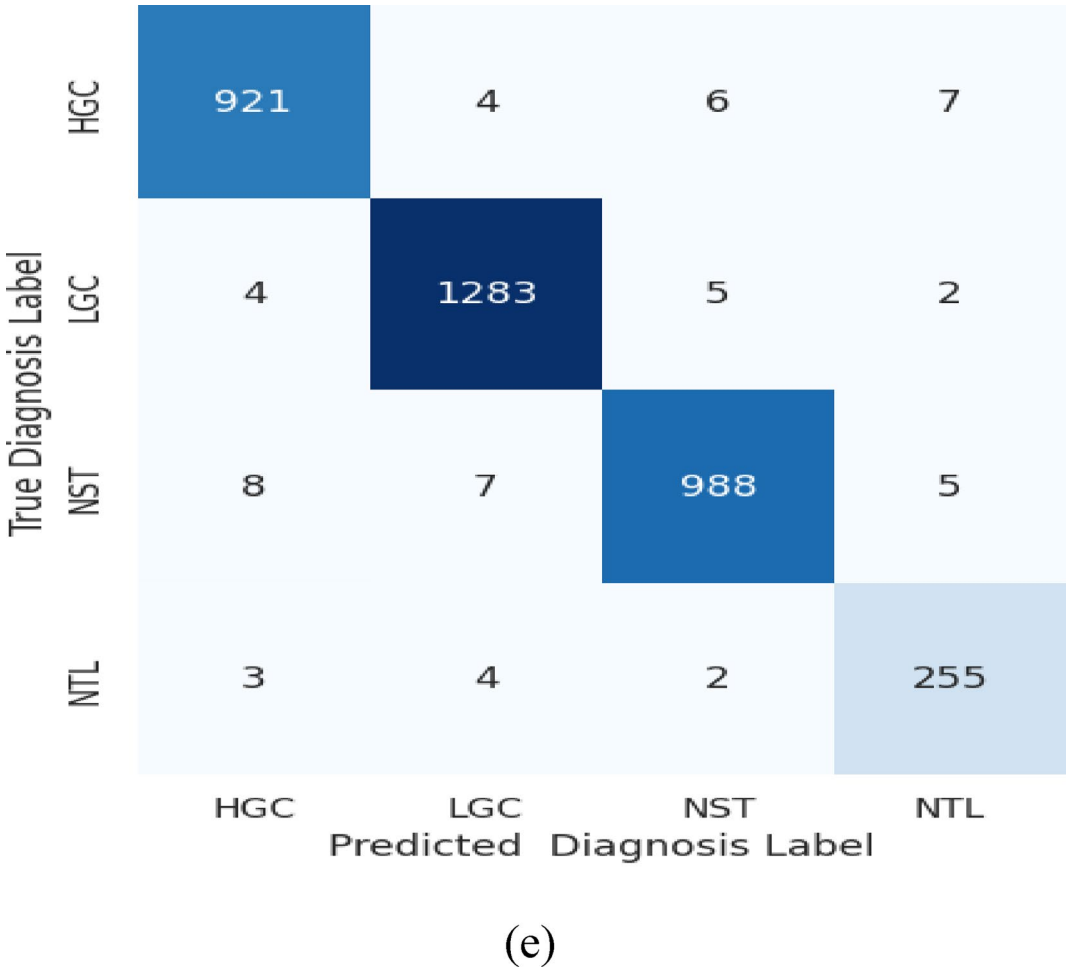


Fig. 2. Confusion Matrix (**a**) Fold1 (**b**) Fold2 (**c**) Fold3 (**d**) Fold4 and (**e**) Fold5.

BCHTNet, and it needs optimization of parameters for fine-grained classification tasks. On the other hand, ResNet-50 provides a balanced performance with an accuracy of 92.69% but is poor at extracting global context and tends to have more false negatives. InceptionV3 performs well in capturing spatial features and is computationally more efficient, but is slowed by the lack of a global attention mechanism, thereby performing with slightly less accuracy 89.21%. Finally, MobileNetV3 is the lightest but provides the worst accuracy, 87.98%, and performs the worst on complex classification tasks, hence having limited value in high-precision medical imaging. The fold wise performance measures of each model in presented in Tables 5, 6, 7, 8, 9 and 10. Table 5 shows the evaluation of CellViT's performance on five folds of cross-validation. The Kappa values, which are about 94%, demonstrate an excellent agreement between the predicted labels and the actual labels. The recall is above 96% in all five folds. Precision is in the range of 94 to 95%, indicating a very low false positive rate. The F1-scores vary between about 95.36% and 95.98%. All the folds have a high accuracy, with the highest in Fold 5 at 97.64%.

Table 6 presents fold-wise performance of the ViT model on five cross-validation folds. The Kappa values, at 92 to 93%, reflect substantial agreement between predicted and true labels. Recall improves consistently from



(e)

Fig. 2. (continued)

Folds	Kappa	Recall	Precision	F1-score	Accuracy
Fold1	0.957	0.9558	0.9624	0.9582	0.9697
Fold2	0.962	0.9604	0.96615	0.9630	0.9732
Fold3	0.962	0.9665	0.9732	0.9698	0.9786
Fold4	0.962	0.9705	0.9789	0.9736	0.9812
Fold5	0.962	0.9768	0.9798	0.9783	0.9837

Table 2. Summary of 5-Folds performance.

Author	Model	Dataset	Performance
Hwang et al.[32]	VGG19	Endoscopy Image Dataset	Accuracy: 91.20%
Zhang et al.[33]	Multistage feature fusion network (MSFF)	Endoscopy Image Dataset	Accuracy: 95.17%,
Lazo et al.[34]	GAN-based model	Endoscopy Image Dataset	Accuracy: 90.00%,
Yoo et al.[35]	ResNeXt-101	Cystoscopic	Accuracy:94.1%,
Liang et al.[36]	Logistic regression	Cystoscopic	AUC: 89%,
Alkhalidy et al.[37]	Ensemble deep learning (XDL)	Cystoscopic	Accuracy:95%
Ye et al.[38]	HRNetV2	Endoscopy images	Precision:91.3%
El-Atier et al.[39]	Ensembles	Endoscopy images	Accuracy: 95.00%
Proposed	BCHTNet	Endoscopic bladder dataset	Accuracy: 97.73%

Table 3. Performance comparison with other methods.

Model	Kappa (%)	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
Proposed	96.10	97.21	96.60	96.86	97.73
CellViT	94.07	95.18	96.17	95.67%	96.87
ViT	92.58	93.10	94.92	94.00%	95.25
YOLOV9	90.72	92.84	93.57	93.20%	93.70
MobileNetV3	86.80	87.15	86.30	86.72%	87.98
InceptionV3	88.20	89.04	88.45	88.74%	89.21
ResNet-50	91.13	92.25	91.60	91.92%	92.69

Table 4. Comparison with SOTA methods.

Folds	Kappa (%)	Recall (%)	Precision (%)	F1-score (%)	Accuracy (%)
Fold1	93.97	95.86	94.87	95.36	96.07
Fold2	94.02	96.14	94.93	95.53	96.27
Fold3	94.13	96.22	95.17	95.69	97.12
Fold4	94.20	96.30	95.35	95.82	97.15
Fold5	94.23	96.36	95.60	95.98	97.64

Table 5. Performance measures of cellvit for each fold.

Folds	Kappa (%)	Recall (%)	Precision (%)	F1-score (%)	Accuracy (%)
Fold1	92.17	93.90	92.60	93.25	94.67
Fold2	92.29	94.23	92.87	93.55	95.08
Fold3	92.50	94.46	93.09	93.77	95.33
Fold4	92.93	95.70	93.34	94.51	95.54
Fold5	92.96	96.35	93.64	94.98	95.62

Table 6. Performance measures of ViT for each fold.

Folds	Kappa (%)	Recall (%)	Precision (%)	F1-score (%)	Accuracy (%)
Fold1	89.27	92.45	91.52	91.98	92.56
Fold2	90.14	93.60	92.06	92.82	93.70
Fold3	90.76	93.42	93.16	93.29	93.74
Fold4	91.05	93.56	93.28	93.42	93.80
Fold5	92.13	94.83	93.98	94.40	94.70

Table 7. Performance measures of YOLO v9 for each fold.

Folds	Kappa (%)	Recall (%)	Precision (%)	F1-score (%)	Accuracy (%)
Fold1	85.57	85.40	86.24	85.85	86.59
Fold2	85.85	85.98	86.26	86.08	86.94
Fold3	86.93	86.30	87.02	86.71	87.85
Fold4	87.19	86.37	88.10	87.20	88.68
Fold5	88.48	87.49	88.14	87.81	89.84

Table 8. Performance measures of MobileNetV3 for each fold.

93.90% in Fold 1 to 96.35% in Fold 5. Precision varying between 92.60% and 93.64%. The F1-scores also increase steadily from 93.25 to 94.98%. Accuracy also increases from 94.67% in Fold 1 to 95.62% in Fold 5.

Table 7 shows the fold-wise performance statistics of the YOLO v9 model. The Kappa statistics vary between 89.27% and 92.13. Recall improves from 92.45 to 94.83% in Fold 5. Precision is also steadily increasing, from 91.52% to 93.98. F1-scores also increase from 91.98 to 94.40%. Accuracy also increases with each fold, reaching

Folds	Kappa (%)	Recall (%)	Precision (%)	F1-score (%)	Accuracy (%)
Fold1	87.25	87.95	88.21	88.05	88.39
Fold2	87.49	88.02	88.75	88.35	89.12
Fold3	88.02	88.53	89.32	88.84	89.40
Fold4	88.10	88.72	89.40	89.02	89.53
Fold5	90.16	89.03	89.51	89.31	89.65

Table 9. Performance measures of InceptionV3 for each fold.

Folds	Kappa (%)	Recall (%)	Precision (%)	F1-score (%)	Accuracy (%)
Fold1	90.56	90.46	91.30	90.90	91.66
Fold2	90.80	91.08	91.47	91.31	92.07
Fold3	91.14	92.05	92.23	92.20	92.22
Fold4	91.39	92.18	92.60	92.41	93.31
Fold5	91.76	92.23	93.67	92.94	94.19

Table 10. Performance measures of ResNet-50 for each fold.

a maximum of 94.70%. All these results indicate that YOLO v9 has stable and increasingly better performance across folds.

Table 8 presents the performance of the MobileNetV3 model on five cross-validation folds. The Kappa values range from 85.57 to 88.48%. Recall increases consistently from 85.40% for Fold 1 to 87.49% for Fold 5. Precision varies, from 86.24 to 88.14%, indicating steady control of false positives. F1-scores also consistently increase from 85.85 to 87.81%, validating overall performance improvement. Accuracy also increases across the folds, beginning at 86.59% and ending at 89.84% in the last fold.

Table 9 illustrates the performance of the InceptionV3 model in five cross-validation folds. The Kappa values range between 87.25% and 90.16%. Recall consistently improves from 87.95% for Fold 1 to 89.03% for Fold 5. Precision also indicates an increase from 88.21 to 89.51%. F1-scores move from 88.05 to 89.31%, confirming the model's balanced execution. Accuracy increases over the folds, from 88.39 to 89.65%.

Table 10 shows the fold-wise performance metrics for the ResNet-50 model. The Kappa values vary between 90.56% and 91.76%. Recall increases from 90.46% in Fold 1 to 92.23% in Fold 5. Precision also increases across the folds, from 91.30 to 93.67%. F1-scores consistently grow from 90.90 to 92.94%, indicating enhanced general predictive performance. Accuracy also has consistent growth from 91.66% up to 94.19% by Fold 5.

The accuracy and loss analysis

We plotted the training and validation loss of the BCHTNet on the endoscopic bladder cancer dataset, as depicted in the Fig. 3. The training accuracy rapidly converges to more than 98% in the first 20 epochs and remains close to 99% afterwards. Analogously, validation accuracy increases sharply in initial epochs, reaching over 90% and stabilizing at around 97%. Training loss begins moderately low and reduces steadily. Validation loss first rises above 1.4 but then falls in the early epochs. Following this initial drop, it settles between 0.1 and 0.2, with slight fluctuation till 160 epochs. In fold 2, training accuracy increases quickly in the early epochs to almost 99%, which means the model fits the training data well. Validation accuracy also increases rapidly to about 97% and stays constant throughout the training, which implies good generalization to new data. As per the loss curve, training loss drops rapidly in the early epochs and levels off at a low value. Validation loss begins high, sharply drops, and reaches between 0.1 and 0.3.

Fold 3, accuracy, and loss curves represent a well-tuned training process over 160 epochs. The training accuracy sharply rises, around 99%, whereas the validation accuracy gradually converges to 97%, signifying successful generalization. On the loss aspect, the training loss decreases from about 0.45 towards less than 0.05, which implies that the model rapidly reduces errors over the training set. The validation loss decreased from around 0.7 to below 0.1 with slight fluctuations. These indicate that the model consistently performs on training data and unseen data without overfitting. Fold 4 exhibits robust model performance with little overfitting. In the accuracy plot, training accuracy begins roughly at 0.85 and increases quickly in the first 20 epochs. Validation accuracy starts roughly at 0.15 but reaches around 0.95 by epoch 20 and holds steady at about 0.98 through the rest of the training. For the loss plot, the training loss starts at around 0.4 and decreases to below 0.05 in epoch 20. Then, it drops slightly more, with a stabilizing trend of nearing 0.01 in later epochs. Validation loss starts at 1.05; it goes down to about 0.1 during the first 20 epochs. It then changes between 0.05 and 0.1 throughout the rest of the training and shows the model has stable generalization.

Fold 5 demonstrates that the model is doing well with strong convergence and little overfitting. The training accuracy starts at about 0.75 and rises by about epoch 20. The validation accuracy also rises from around 0.25 to well over 0.90 during the same timeframe, then slowly improving, levelling at around 0.99 for the remainder of the training. In the loss curve, the training loss starts around 0.5 and decreases rapidly below 0.05 in the initial 20 epochs, reducing to around 0.01. The validation loss begins around 0.8, decreases sharply in the initial epochs,

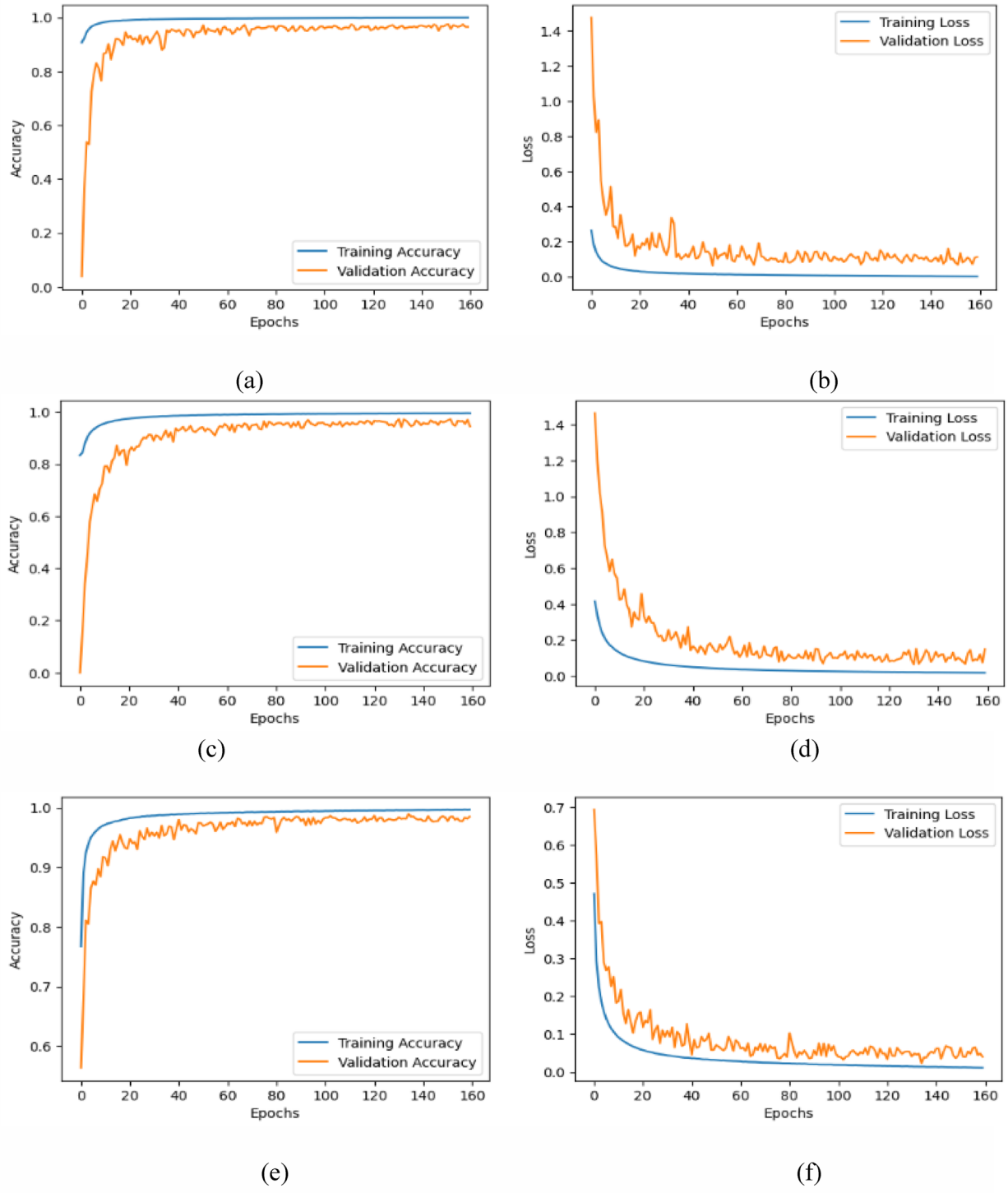


Fig. 3. The training and loss curves for Fold 1, Fold 2, Fold 3, Fold 4, and Fold 5 are shown in (**a**–**j**), respectively.

and settles between 0.05 and 0.1, but without any overfitting trend. These loss values are consistent with high accuracy, verifying successful model training and robust validation performance in fold 5.

ROC plot based comparison

We performed an ROC plot-based comparison of the ResNet-50[33], Inception V3[34], MobileNetV3[35], YOLOV9[36], ViT[37] and CellViT[38] with the proposed BCHTNet shown in the Fig. 4. With the highest average Area Under

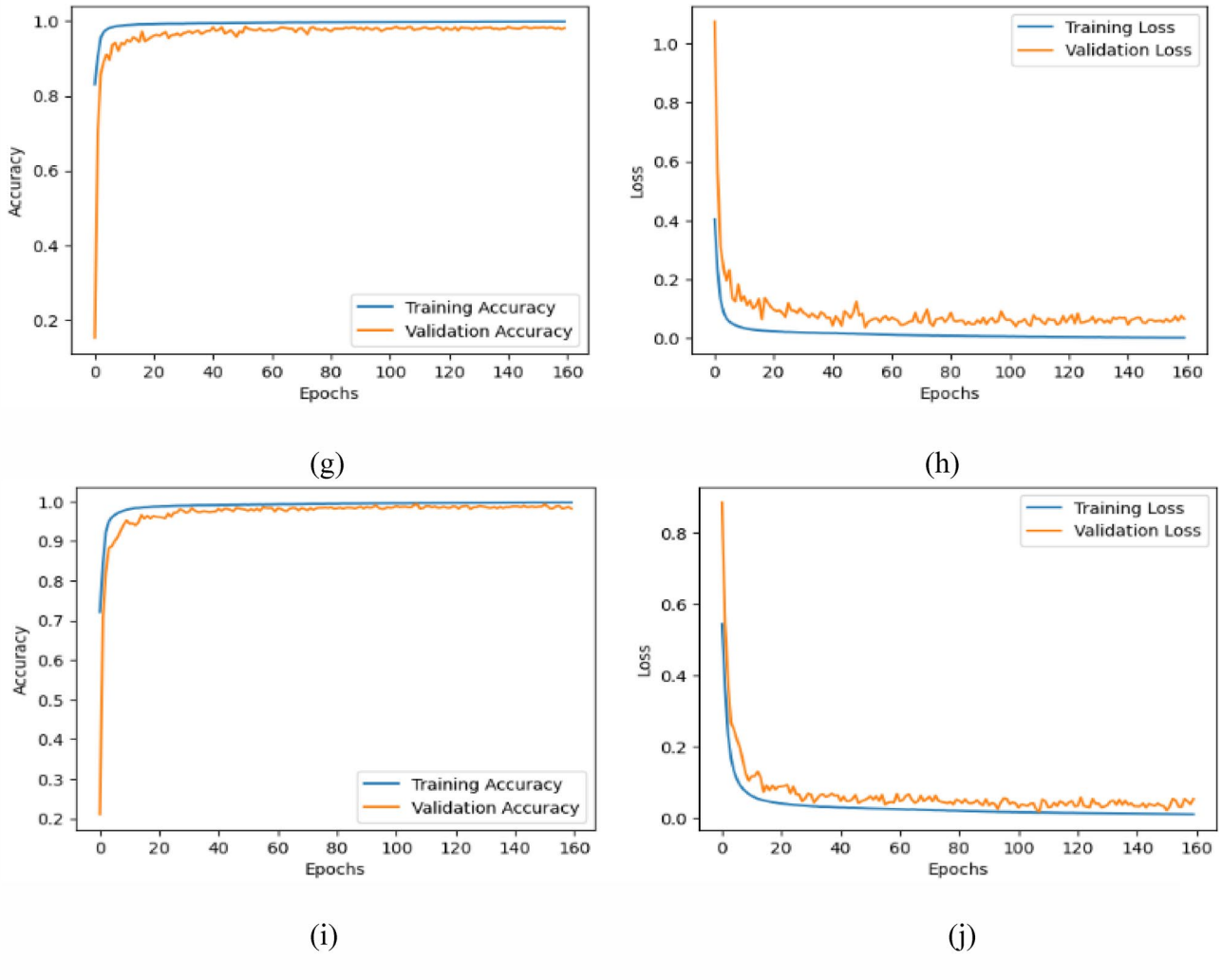


Fig. 3. (continued)

the Curve (AUC) value of 0.9903, the proposed model performs exceptionally well in classification, striking an almost ideal balance between sensitivity (False Positive Rate) and specificity (True Positive Rate). The curve's close adherence to the plot's upper-left corner demonstrates its ability to attain a high true positive rate while keeping a low false positive rate. ViT and CellViT are the best-performing models, with an average AUC of 0.9640 and 0.9712, respectively. Despite their excellent efficacy, these models perform marginally worse than the Proposed Model, with slightly higher false positive rates. Although YOLOV9 performs reasonably well, with an average AUC of 0.9517, it deviates more noticeably from the upper-left corner of the plot, suggesting a marginally worse balance between sensitivity and specificity.

The ResNet-50 obtained an average AUC of 0.9418 and lags further behind the top-performing models, demonstrating respectable but subpar performance. On the other hand, InceptionV3 and MobileNetV3, whose average ROC curves are located farthest from the top-left corner, perform the worst, with average AUC values of 0.9016 and 0.8928, respectively. This implies that they are less dependable for the classification task due to higher false positive and lower true positive rates. The proposed model is the most dependable and efficient classifier with a substantial performance advantage over all other models. Although they are good substitutes, CellViT and ViT perform marginally worse. While InceptionV3 and MobileNetV3 must be optimized for comparable results, models such as YOLOV9 and ResNet50 provide moderate performance.

Ablation study

We performed an ablation study of different components of the BCHTNet on the bladder image dataset, shown in Table 11. Based on the findings, the CNN + ViT with attention transformation has the best F1-score of 96.86%, accuracy of 97.73%, and precision of 97.21%. This shows that combining both CNNs for local feature extraction and ViT for global feature modelling while reducing attention complexity achieved the best balance between performance and computational efficiency. A solo CNN, however, achieves 94.32% accuracy while performing poorly, so transformer-based models are probably a better way to capture long-range relationships. The accuracy of the ViT with full multi-head self-attention (MHSA) is 95.12%, which is worse than the ViT with

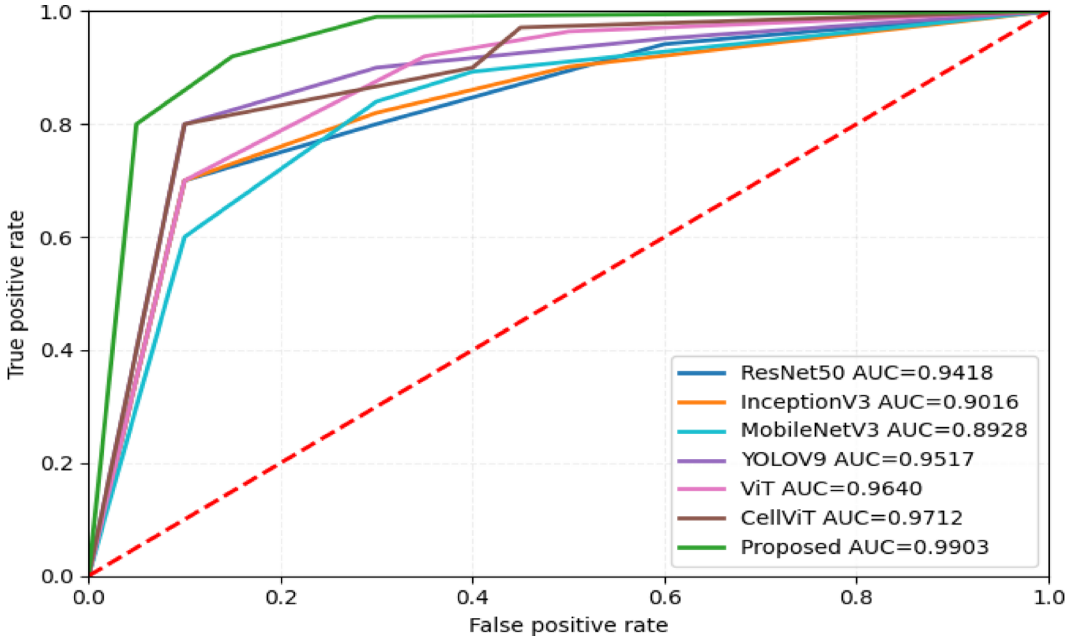


Fig. 4. ROC plot-based comparison with SOTA methods.

Components	Accuracy (%)	Precision (%)	F1-score (%)
CNN + ViT + Attention transformation	97.73	97.21	96.86
CNN	94.32	94.08	93.61
ViT + MHSA	95.12	94.85	94.59
ViT + transformation	96.48	96.11	95.81

Table 11. Effect of different components.

attention transformation (96.48%) but still better than CNN. This implies that although MHSA improves feature representation, it could add too much complexity, resulting in less-than-ideal generalization. Interestingly, ViT with attention transformation achieves 96.48% accuracy, which is better than ViT with full MHSA. This indicates that, without sacrificing computing efficiency, reducing the number of attention heads or layers helps improve generalization. Overall, the results indicate that the best-performing method is a hybrid CNN + ViT model with reduced attention, which circumvents the drawbacks of excessive attention complexity while retaining the benefits of transformers for long-range dependencies and CNNs for local feature extraction.

Training and validation time comparison

We compared the training and validation time of the BCHTNet, ResNet-50[33], Inception V3[34], MobileNetV3[35], YOLOV9[36], ViT[37] and CellViT[38], and the results are depicted in the Fig. 5. Figure 5 shows that transformer-based methods CellViT and ViT took the highest training and validation time. At the same time, YOLOV9 and ResNet-50 have less training and validation time. Furthermore, MobileNetV3 has the least computation time. Moreover, BCHTNet and InceptionV3 training and validation times are close.

The Grad-CAM based analysis

The expert can utilize the Grad-CAM results to visualize the decision process, and it helps the oncologist to locate the region highlighted by the model. Figure 6 shows that without attention, the transformation module model is not able to focus on the region of interest. At the same time, including the attention transformation in the model produces better Grad-cam results and can reach more bladder cancer regions.

Parameters and flops comparison

We calculated each model's parameters (in millions) and Gflops and presented them in Table 12. Table 10 compares various models' computational time (GFLOPs) and number of parameters (in millions). The proposed model has 3.8 GFLOPs and 20.63 M parameters, which is more efficient compared to ResNet-50 with 4.1 GFLOPs, 23.51 M, and InceptionV3 with 5.3 GFLOPs, 23.85 M. It is also lighter compared to YOLOv9 with 22.5 GFLOPs, 25.30 M, and ViT with 48 GFLOPs, 86 M, and CellViT with 57 GFLOPs, 21.7 M. Though MobileNetV3 has the lowest 0.22 GFLOPs and 5.47 M parameters, the proposed model provides a good trade-off between low cost and high accuracy.

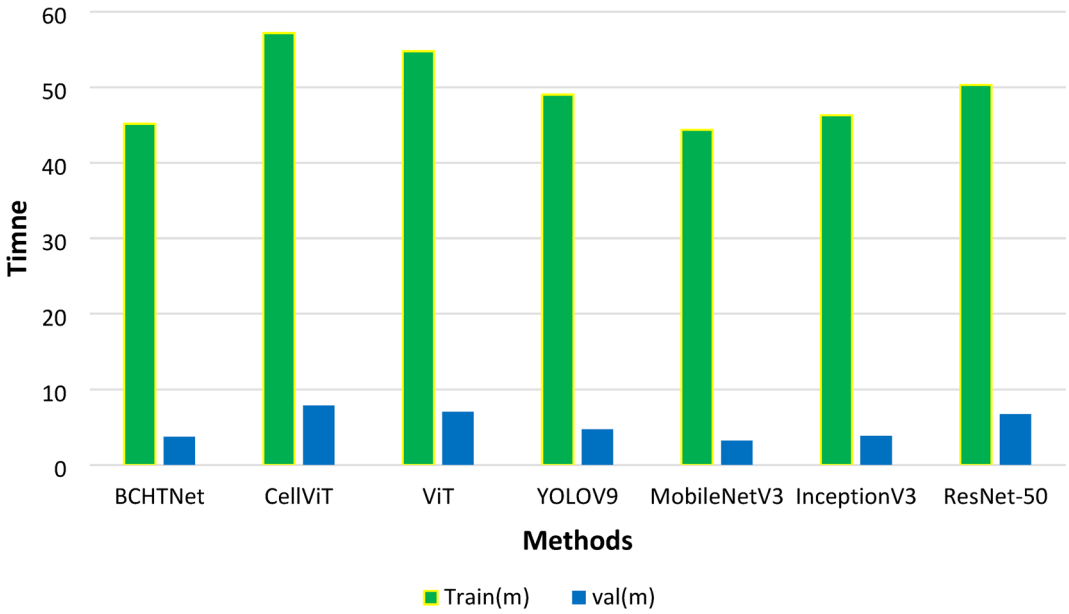


Fig. 5. Training and validation time comparison with SOTA methods.

Conclusion

In this study, we designed a hybrid deep learning model that combines the attention transformation-based-ViT with an Inception V3 module. The quadratic computation costs of the classical MHSA in the ViT make it less applicable in real-time applications. We designed a hybrid multi-head self-attention using an attention transformation module and incorporated it into the model for long-range dependencies. In addition, a lightweight Inception V3 module is utilized for collecting fine-grained spatial features. The proposed method is more effective and scalable for real-time applications that reduce complexity without losing important contextual information. The model produced impressive results that consistently outperformed state-of-the-art methods with an average accuracy of 97.73%, an F1-score of 96.86%, and an AUC of 0.9903. These results confirm that the suggested approach is dependable in correctly dividing bladder tissue into several diagnostic groups. Though the model shows robust performance on an endoscopic bladder image dataset. However, potentially influencing generalizability across varied clinical environments or imaging sources conditions need to be tested. In addition, the hybrid architecture consumes significant computational resources, which can restrict its real-time usage in low-resource settings. Moreover, the model in its present form is image-based feature-focused, and its performance can be improved further by incorporating multimodal data like clinical history or genomic markers. In addition, nature inspired algorithm such as Grey wolf, Ant colony, Swarm optimization can be used for the optimization of the spatial features.

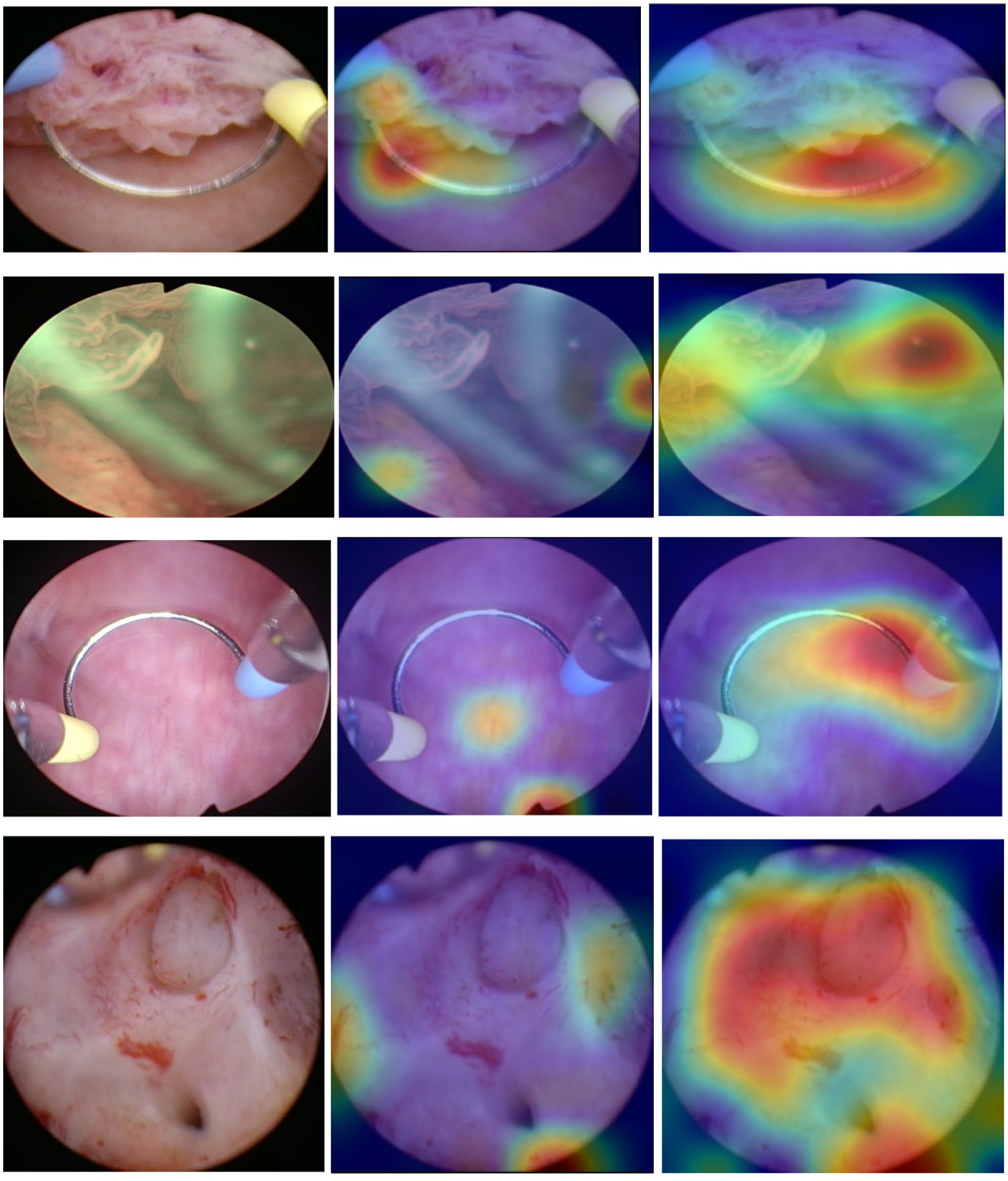


Fig. 6. (**a**) Original Image (**b**) Grad-cam results without attention transformation module and (**c**) Grad-cam results with attention transformation module.

Model	GFlops	Parameters (M)
ResNet-50	4.1	23.51
InceptioV3	5.3	23.85
MobileNetV3	0.22	5.47
YOLOV9	22.5	25.30
ViT	48	86
CellViT	57	21.7
Proposed	3.8	20.63

Table 12. Parameters and flops comparison with SOTA methods.

Data availability

The data of the present study can be downloaded from the URL: https://zenodo.org/records/7741476.

Received: 8 March 2025; Accepted: 15 May 2025
Published online: 23 May 2025

References

1. Kwong, J. C. C. et al. Predicting non-muscle invasive bladder cancer outcomes using artificial intelligence: a systematic review using APPRAISE-AI. *NPJ Digit. Med.* **7**(1), 1–11 (2024).
2. Jubber, I. et al. Epidemiology of bladder cancer in 2023: a systematic review of risk factors. *Eur. Urol.* **84**(2), 176–190 (2023).
3. Sehrawat, A., Gopi, V. P. & Gupta, A. A systematic review on role of deep learning in CT scan for detection of gall bladder cancer. *Arch. Comput. Methods Eng.* **31**(6), 3303–3311 (2024).
4. American Cancer Society. *Key Statistics for Bladder Cancer* (American Cancer Society, 2024). https://www.cancer.org/cancer/types/bladder-cancer/about/key-statistics.html
5. Organization, W. H. Cancer Today (International Agency for Research on Cancer, 2020). https://gco.iarc.fr/today/en.
6. Feretzakis, G. et al. Emerging trends in AI and radiomics for bladder, kidney, and prostate cancer: A critical review. *Cancers (Basel).* **16**(4), 810 (2024).
7. He, K. et al. Progress of multiparameter magnetic resonance imaging in bladder cancer: A comprehensive literature review. *Diagnostics (Basel).* **14**(4), 442 (2024).
8. Zhang, S. et al. Machine learning assisted microfluidics dual fluorescence flow cytometry for detecting bladder tumor cells based on morphological characteristic parameters. *Anal. Chim. Acta.* **1317**(342899), 342899 (2024).
9. Bazarkin, A. et al. Assessment of prostate and bladder cancer genomic biomarkers using artificial intelligence: A systematic review. *Curr. Urol. Rep.* **25**(1), 19–35 (2024).
10. Li, C., Qin, W., Hu, J., Lin, J. & Mao, Y. A machine learning computational framework develops a multiple programmed cell death index for improving clinical outcomes in bladder cancer. *Biochem. Genet.* **62**(6), 4710–4737 (2024).
11. Sun, R. et al. Preoperative CT-based deep learning radiomics model to predict lymph node metastasis and patient prognosis in bladder cancer: a two-center study. *Insights Imaging* **15**, 1 (2024).
12. Asif, S., Khan, S. U. R., Amjad, K. & Awais, M. SKINC-NET: An efficient lightweight deep learning model for multiclass skin lesion classification in dermoscopic images. *Multimedia Tools Appl.* 1–27 (2024).
13. Khan, S. U. R., Asif, S., Bilal, O. & Ali, S. Deep hybrid model for Mpox disease diagnosis from skin lesion images. *Int. J. Imaging Syst. Technol.* **34**(2), e23044 (2024).
14. Khan, S. U. R., Asif, S., Zhao, M., Zou, W. & Li, Y. Optimize brain tumor multiclass classification with manta ray foraging and improved residual block techniques. *Multimedia Syst.* **31**(1), 1–27 (2025).
15. Shahzad, I., Khan, S. U. R., Waseem, A., Abideen, Z. U. & Liu, J. Enhancing ASD classification through hybrid attention-based learning of facial features. *Signal. Image Video Process.* **18**(Suppl. 1), 475–488 (2024).
16. Jiao, P. et al. Prediction of HER2 status based on deep learning in H&E-stained histopathology images of bladder cancer. *Biomedicines* **12**(7), 1583 (2024).
17. Zhou, Y., Zheng, X., Sun, Z. & Wang, B. Analysis of bladder cancer staging prediction using deep residual neural network, radiomics, and RNA-Seq from high-definition CT images. *Biochem. Genet. (Camb.)* **2024**, 1–11 (2024).
18. Li, J., Wang, Z. & Wang, T. Machine-learning prediction of a novel diagnostic model using mitochondria-related genes for patients with bladder cancer. *Sci. Rep.* **14**(1), 1–14 (2024).
19. Lee, S. et al. Early-stage diagnosis of bladder cancer using surface-enhanced Raman spectroscopy combined with machine learning algorithms in a rat model. *Biosens. Bioelectron.* **246**(115915), 115915 (2024).
20. Luo, X., Chen, X. & Yao, Y. Integrating Multiview, multi-scale graph convolutional networks with biological prior knowledge for predicting bladder cancer response to neoadjuvant therapy. *Appl. Sci. (Basel).* **14**(2), 669 (2024).
21. Sarkar, S. et al. Performing automatic identification and staging of urothelial carcinoma in bladder cancer patients using a hybrid deep-machine learning approach. *Cancers (Basel).* **15**(6), 1673 (2023).
22. Lee, J. Y. et al. Selection of convolutional neural network model for bladder tumor classification of cystoscopy images and comparison with humans. *J. Endourol.* **38**(10), 1036–1043 (2024).
23. Yue, X., Huang, X., Xu, Z., Chen, Y. & Xu, C. Involving logical clinical knowledge into deep neural networks to improve bladder tumor segmentation. *Med. Image. Anal.* **95**(103189), 103189 (2024).
24. Khedr, O. S., Wahed, M. E., Al-Attar, A. S. R. & Abdel-Rehim, E. A. The classification of the bladder cancer based on vision Transformers (ViT). *Sci. Rep.* **13**(1), 20639 (2023).
25. Shalata, A. T. et al. Precise grading of non-muscle invasive bladder cancer with multi-scale pyramidal CNN. *Sci. Rep.* **14**(1), 1–12 (2024).
26. Fahoum, I., Naamneh, R., Silberberg, K., Hagege, R. & Hershkovitz, D. Detection of muscularis propria invasion in urothelial carcinoma using artificial intelligence. *Technol. Cancer Res. Treat.* **23** (2024).
27. Yang, R. et al. Automatic recognition of bladder tumours using deep learning technology and its clinical application. *Int. J. Med. Rob. Comput. Assist. Surg.* **17**(2), e2194 (2021).
28. Dudaie, M., Dotan, E., Barnea, I., Haifler, M. & Shaked, N. T. Detection of bladder cancer cells using quantitative interferometric label-free imaging flow cytometry. *Cytometry A.* **105**(8), 570–579 (2024).
29. Yu, J. et al. A novel predict method for muscular invasion of bladder cancer based on 3D mp-MRI feature fusion. *Phys. Med. Biol.* **69**(5), 055011 (2024).

30. Jiang, Y. H. et al. Hypermethylation loci of ZNF671, IRF8, and OTX1 as potential urine-based predictive biomarkers for bladder cancer. *Diagnostics (Basel)*. **14**(5), 468 (2024).
31. Lazo, J. F. et al. Endoscopic bladder tissue classification dataset, IEEE Transactions on Biomedical Engineering, vol. 70, no. 10, pp. 2822–2833, [Online]. https://doi.org/10.5281/zenodo.7741476 (2023).
32. Amaouche, M. et al. Redefining cystoscopy with AI: Bladder cancer diagnosis using an efficient hybrid CNN-transformer model. in *IEEE International Conference on Image Processing (ICIP)* 3030–3036 (2024).
33. Kurata, Y. et al. Development of deep learning model for diagnosing muscle-invasive bladder cancer on MRI with vision transformer. *Heliyon* **10**(16), e36144 (2024).
34. Li, X. et al. MH2AFormer: an efficient multiscale hierarchical hybrid attention with a transformer for bladder wall and tumor segmentation. *IEEE J. Biomed. Health Inf.* **28**(8), 4772–4784 (2024).
35. Tao, T., Chen, Y., Shang, Y., He, J. & Hao, J. SMMF: a self-attention-based multi-parametric MRI feature fusion framework for the diagnosis of bladder cancer grading. *Front. Oncol.* **14**, 1337186 (2024).
36. Borna, M. R., Sepehri, M. M., Shadpour, P. & Khaleghi Mehr, F. Enhancing bladder cancer diagnosis through transitional cell carcinoma polyp detection and segmentation: an artificial intelligence powered deep learning solution. *Front. Artif. Intell.* **7**, 1406806 (2024).
37. Cao, K. et al. A multicenter bladder cancer MRI dataset and baseline evaluation of federated learning in clinical application. *Sci. Data*. **11**(1), 1147 (2024).
38. Hwang, W. K. et al. Artificial intelligence-based classification and segmentation of bladder cancer in cystoscope images. *Cancers (Basel)*, **17**, 1 (2024).
39. Alazwari, S. et al. Automated gall bladder cancer detection using artificial gorilla troops optimizer with transfer learning on ultrasound images. *Sci. Rep.* **14**(1), 21845 (2024).
40. 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* 770–778 (2016).
41. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2818–2826 (2016).
42. Howard, A. et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 1314–1324 (2019).
43. Yaseen, M. What is yolov9: an in-depth exploration of the internal features of the next-generation object detector. *ArXiv Preprint arXiv*: 240907813. (2024).
44. Vaswani, A. et al. Attention is all you need. *Adv. Neural. Inf. Process. Syst.* 30 (2017).
45. Horst, F. et al. Cellvit: vision Transformers for precise cell segmentation and classification. *Med. Image Anal.* **94**, 103143 (2024).

Author contributions

Conceptualization, Poonam Sharma, Dhirendra Prasad Yadav, Bhisham Sharma; Data Curation, Dhirendra Prasad Yadav, Bhisham Sharma, Deepti Thakral; Formal analysis, Poonam Sharma, Deepti Thakral; Investigation, Poonam Sharma, Bhisham Sharma, Julian L. Webber; Methodology, Poonam Sharma, Dhirendra Prasad Yadav, Bhisham Sharma; Project administration, Deepti Thakral, Julian L. Webber; Software, Poonam Sharma, Bhisham Sharma; Visualization, Deepti Thakral, Bhisham Sharma, Julian L. Webber; Writing – original draft, Poonam Sharma, Dhirendra Prasad Yadav; Writing – review & editing, Bhisham Sharma, Deepti Thakral, Julian L. Webber.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.S.

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.