Received: October 29, 2024. Revised: November 21, 2024. 608

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Accurate Bladder Cancer Classification and Prognosis with a Hybrid Vision Transformer

Roaa Alkhalidy[1] Ebtesam AlShemmary[2]* Zhentai Lu[3]

[1]*Department of Computer Science, Faculty of Computer Science and Mathematics, University of Kufa, Kufa, Iraq*
[2]*IT Research and Development Center, University of Kufa, Kufa, Iraq*
[3]*School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China*
*Corresponding author's Email: dr.alshemmary@uokufa.edu.iq

Abstract: Bladder cancer, a common and increasingly prevalent malignant neoplasm, has become a major threat to public health. The major significance enjoyed by this combination is that it emphasizes the need to comprehend and diagnose a condition at a very young age. A number of automatic diagnostic techniques employing Artificial Intelligence technology have been deployed in the area of bladder cancer and have generally been found to be effective but the actual percentage accuracy required for real-world diagnosis has not been achieved yet. The proposed hybrid model integrates Vision Transformers and Convolutional Neural Networks to enhance the performance of medical vision tasks, such as the classification and prediction of bladder cancer. Vision Transformers excel in capturing long-range relationships through self-attention mechanisms, while Convolutional Neural Networks are more effective in extracting local features using spatial convolution filters. This combination leverages the strengths of both architectures to achieve superior results in medical image analysis. The Endoscopy dataset, and Pathological dataset were used and the hybrid model provided better accuracy for diagnosing bladder cancer than the basic CNN model and transferred learning, VGG16, Inception-v3, MobileNetV2 of 99.93%, 93.67%, 97.8%, 95.5%, 95% respectively. The experiments show that these hybrid architectures are more effective than the conventional Convolutional Neural Networks for the Bladder cancer Image classification and establish the suitability of Vision Transformer in progressing towards higher performance with the Bladder cancer Image analysis.

Keywords: Artificial intelligence, Bladder cancer, Classification, Transfer learning, Vision transformer.

1. Introduction

Bladder cancer is among the most frequently diagnosed and rapidly rising neoplasms, and it represents a considerable threat to world health care. The time of diagnosis can be considered as the critical factor related to the potential treatment efficacy as with early diagnosis appropriate treatment can be started. Diagnostic techniques with AI in automatic diagnosis methods proposed for bladder cancer diagnosis shows a positive outcome; but high accuracy is still a problem. Rule-based and statistical AI paradigms are relatively insensitive to local structures and cannot successfully identify long-range connections inside the medical images for which they are used. Bladder cancer is an important global health concern and this has remained a leading instance across the globe. As stated by the World Health Organization WHO [1]. It is estimated to be one of leading causes of cancer worldwide. However, the outcome for bladder cancer still remains unsatisfactory, mainly because of diagnostic delays [2]. Consequently, timely identification continues to present as a major driver of increased survival and lessening of the impact of bladder cancer on patients and healthcare organizations. Current approaches to early diagnosis of bladder cancer are based on the results of clinical examination and, in certain circumstances, biopsy. However, these methods have their downfall. These images are difficult to analyse with precision and speed; a fast processing demands computational methods that create the basis of artificial intelligence within the field of bladder

cancer diagnosis [3]. Many applications of artificial intelligence have shown a great deal of efficiency in some areas, and AI has had a particularly strong performance in medical imaging [4]. CNNs have always dominated the image classification area especially in the diagnosis of pathologies from medical images. However, there has been a shift of paradigm with the advances of transformer based architectures [5]. The transformer model, introduced by Vaswani et al. [6], revolutionized the field of NLP with its self-attention mechanism, enabling the modelling of long-range dependencies. This study investigates the adaptation of the transformer architecture for image classification, known as the Vision Transformer (ViT). Unlike traditional Convolutional Neural Networks (CNNs), ViT operates on sequences of image patches, thereby leveraging the transformer's inherent ability to capture intricate features across the entire image. Chapman-Sung et al. in 2020 [7], the authors' argument is that extracted features with CNN through classifying classifiers are less accurate when manually extracted from domain knowledge. Hence, the aim was to classify two challenging early stages of bladder cancer: Ta (noninvasive) and T1 (superficially invasive) that are histologically similar. Overall, 1177 scans of bladder formed totals in the dataset, of which 460 were identified as minimally invasive, and 717 as superficially invasive. The CNN classifiers achieved only 84.0% accuracy and was significantly below the other supervised machine learning classifiers developed using manually defined features. Sarkar et al. [8] in 2023, intended to build a diagnosis model of bladder cancer by using the radionics aided interpretation of CT scan. The three classification tasks performed include: Normal vs bladder cancer, NMIBC vs MIBC and PTC vs MIBC. The dataset contained 165 regions of interest (ROI), 100 normal images, and 65 cancer images. Firstly, it was a retrospective study; secondly, it was conducted in a single-centre only. Cross validation was performed 10 time The results of tests for the LDA classifier on characteristics based on Xception Net for differentiation of having or not having cancer were 86,07%.

Liu D, Wang S, Wang J, et al in 2020 [9] used CT data from 75 patients with bladder for classification and staging of this disease by using a ResNet based model and applied super-Resolution to improve medical pictures. This model developed has a sensitivity rate of 94.74% and was achieved after the data received from 76 people with bladder cancer had been reviewed in order to compare it with the preoperative pathological diagnosis.

Zhang, G. et al [10] in 2021, in this study 183 patients that made up the dataset utilized were divided into three sets: There are 110 participants for training, 73 for internal validation and 75 for external testing. The researchers developed a completely novel convolutional network known as FGP-Net to incorporate DFL and Dense Blocks. Specification of the FGP-Net on the internal dataset resulted in an AUC (Area Under the Receiver Operating Characteristic Curve) of 0.861 and accuracy at 0.795. The AUC and the accuracy of the FGP-Net on the external data base were equal to 0.791 and 0.747 respectively. These results pinpoint the efficiency of the FGP-Net in the classification and its performance on the external dataset.

Chen et al. [11], introduced TransUnet, a model incorporating both transformers and Unet for medical image segmentation. Their investigation demonstrated that transformers function as robust encoders for medical image segmentation tasks. The amalgamation with Unet enhances finer details by recovering localized spatial information, leading to superior performance compared to various competing methods across diverse medical applications. The strengths of the proposed method can be summarized as follows: the hybrid model, which combines Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs), represents a significant advancement in bladder cancer classification. Vision Transformers excel at capturing long-range dependencies through self-attention mechanisms, making them particularly effective for medical images, where dispersed pathologies may exist across various regions. Conversely, CNNs are highly efficient at identifying local features using spatial convolution filters. By integrating these two architectures, the hybrid model successfully overcomes the limitations inherent in each individual approach, resulting in enhanced performance. Specifically, this hybrid model addresses the challenges faced by traditional convolution-based networks, such as difficulty in capturing global contextual information, as well as the feature mismatch issues observed in transfer learning techniques applied to medical images.

The experimental results indicate a marked improvement in accuracy, with the hybrid model outperforming established models such as VGG16, Inception-v3, and MobileNetV2, achieving an exceptional 99.93% accuracy in bladder cancer diagnosis. Consequently, these results indicate the promise of the proposed method in the development of the state of the art in bladder cancer image analysis.

The organization of the paper is as follows: the introduction brought out the subject of bladder cancer

around the world and the difficulties of diagnosing the disease and the subsequent literature review section reviews some of the general traditional AI techniques and their drawbacks. The next topic to be presented will be the introduction of the hybrid model alongside a discussion of how the features of the ViTs and the CNNs are integrated. The features notations, data sets employed in this study, the proposed model design and the experimental details will be captured under the methodology section. The performance analysis as well as comparison with traditional models will be presented in the result section. At the end, the conclusion will briefly discuss the main findings of the current study and possible research avenue for the future.

2. Literature review

The application of deep learning solutions in bladder cancer analysis has extended its contributions in the diagnosis, detection, classification as well as staging. Investigators have used different deep learning approaches, such as CNNs, ResNet, U-Net, and Transformer networks to perform diagnostic analysis of histopathology, endoscopy, and cystoscopy image modalities.

In (2021), Qiu et al., applied a deep learning algorithm for classification of subtypes of bladder cancer based on histopathological images. Made use of ResNet50 model for distinguishing between multiple forms of cancers that are characterized by histopathological patterns. The extraction of features was done using some of its layers and was further trained using transfer learning to achieve ideal characteristic of images of bladder cancer. Contributed a set of 1389 histopathological images where each image was labeled for at least three subtypes of bladder cancer. None of the preprocessing steps was left unlabeled, and normalization and augmentation were performed to improve the model. The proposed methodology was validated by getting a maximum of 88.7% accuracy of bladder cancer subtypes classification. This could help pathologists to well diagnose and sort out various types of cancer at high level of accuracy [12]. In (2020), Zhu et al further proposed a deep learning model for detecting bladder cancer from cystoscopy images. Used the cystoscopy image specialized custom CNN model. The architecture was aimed at identifying features that where associated with bladder tumours. Image pre-processing included contrast stretching and noise filtering. Consisted of cystoscopy picture captured during normal endoscopic examinations. This training set contain samples of tissues of cancer and normal status. Reaching a sensitivity of 93.5 % and specificity of 89.2 % the model proved itself good at recognising cars in automated manner [13].

In (2019), Wang et al. perform a method of segmenting tumour regions in endoscopic image and classifies for detection of bladder cancer. Implemented segmentations that used a U-Net model and a CNN for classification. This architecture was selected for its potential to segment the medical images effectively. These were comprised of endoscopic images with regions of tumour outlined. For the purpose of the evaluation the dataset was further divided into training, validation, and test sets. The measures of segmentation accuracy were 85%, and classification accuracy were 87%. It also demonstrated that the method could assist clinicians in differentiating malignant tissues during endoscopic examination [14].

Song et al. in (2022) develop an advanced pathway for performance improvement in the classification of bladder cancer through the use of the AHBIC fusing pathological and endoscopic images deep learning model. To handle multimodal data, the CNNs were combined with Transformer architectures through a model they established. The Transformer part learned contextual interactions in the image input, while the CNN part detected informative features. Combined both pathological and endoscopic images from a particular patient. The use of multiple data modes proved to yield better representation of data and, therefore, better classification across the various scenarios. They labelled a combined dataset that when used in training the model attained ninety-one-point two percent accuracy was higher than the accuracy of models trained on either kind of data [15]. Liu et al., in their work in (2020) have used deep learning to distinguish between malignant and benign bladder tumours. In the present study, transfer learning was utilised where the pre-trained InceptionV3 model was fine-tuned on the bladder tumour image dataset. This approach it made use of the model in its capacity as a general image processor thus efficient in processing data in the different image forms. Contains histopathological images of bladder tissues where the images are ‘Benign’ or ‘Malignant’. Pre-processing of data was done to increase the size of the dataset and minimize over fitting of the models. Using transfer learning, the developed model reached a classification accuracy of 90.5% thus proving that transfer learning is a useful technique for histopathological image analysis [16].

In (2021), Matsumoto et al. proposed a method of using deep learning in order to compare and detect early bladder cancer in optical endoscopy images.

Used a deep CNN in order to detect pre-malignant changes in mucosa in endoscopic pictures of the gastrointestinal tract. The above model architecture was designed to minimize the reduction of certain easy patterns decreasing the possibility of performing well when faced with complicated patterns within the frames of endoscopic videos. Contained endosomal images in which patients that were normally diagnosed with bladder cancer underwent normal screening test. In image pre-processing, efforts were made to control noise and also scale the images to equal pixel dimensions. The model had a sensitivity of 92.3% and specificity of 88.7% to diagnose early-stage bladder cancer [17]. In 2022, Jiang et al. Integrated pathological and endoscopic imaging data to predict bladder cancer stages, utilizing a multi-view CNN technique to combine features from both datasets.

Table 1. Summary of the deep learning techniques in bladder cancer classification

Author / Ref.	Dataset Type	Model	Objective	Result	Strengths	Limitations
Qiu et al. (2021) [12]	Histopathological images	ResNet50 with transfer learning	Classify bladder cancer subtypes	Accuracy of 88.7%	High accuracy with transfer learning; aids pathologists in subtype classification	Limited to histopathological images only; subtype-specific performance unreported
Zhu et al. (2020) [13]	Cystoscopy images	Custom CNN	Detect bladder cancer in cystoscopy images	Sensitivity of 93.5%, specificity of 89.2%	Effective for clinical detection with high sensitivity and specificity	Custom model may lack generalizability; focus limited to cystoscopy data
Wang et al. (2019) [14]	Endoscopic images	U-Net for segmentation, CNN for classification	Segment and classify tumor regions	Segmentation accuracy of 85%, classification accuracy of 87%	Provides accurate segmentation; assists clinicians during endoscopy	Limited to endoscopic imaging; separate segmentation and classification stages
Song et al. (2022) [15]	Pathological and endoscopic images	Hybrid CNN-Transformer model	Improve classification with multimodal data	Accuracy of 91.2% with combined data	High accuracy leveraging multimodal data; better representation	Complexity of model integration; requires both data types from each patient
Liu et al. (2020) [16]	Histopathological images	InceptionV3 with transfer learning	Differentiate malignant from benign tumors	Classification accuracy of 90.5%	Effective transfer learning on histopathology; robust performance	Focus limited to binary classification (malignant vs. benign); less nuanced
Matsumoto et al. (2021) [17]	Endoscopic images	Deep CNN	Detect early-stage bladder cancer	Sensitivity of 92.3%, specificity of 88.7%	High sensitivity for early-stage detection; useful in routine screening	Limited to early-stage cancers; possible overfitting due to image preprocessing
Jiang et al. (2022) [18]	Endoscopic and pathological images	Multi-view CNN	Predict bladder cancer stage	Staging accuracy of 89.8%	Joint learning from multimodal data; valuable for staging	Dependent on availability of both imaging types; may not generalize across institutions

The design improved the model's capacity to forecast cancer stage by enabling collaborative learning from both sources of input. included matching pathology and endoscopic pictures of patients with bladder cancer, together with staging data as ground truth. demonstrated the benefit of using multi-view data for clinical decision support by achieving an 89.8% staging accuracy [18].

Table 1 shows various deep learning techniques applied to bladder cancer classification focusing on the approach, methodology, and results across various datasets and clinical applications. These studies highlight the effectiveness of deep learning techniques in bladder cancer imaging, revealing that hybrid models and transfer learning improve classification and detection accuracy. However, challenges include limited generalizability and the need for multimodal data, which may restrict clinical applicability.

3. Transformers

Transformers were originally developed for natural language processing (NLP) tasks, especially machine translation, and have since proven highly effective across multiple fields, including NLP, image generation, and bioinformatics. This neural network architecture is based on an encoder-decoder framework, where each component consists of several identical transformer blocks. The encoder processes input data into a series of encodings, which are then used by the decoder, enhanced with contextual information, to generate output sequences. Each transformer block incorporates a multi-head attention layer for capturing relationships across different parts of the input, a feed-forward neural network, and optimization layers such as shortcut connections and layer normalization. [19].

Transformers employ an encoder and decoder structure with stacked encoder and decoder layers. The encoder layer has two sub-layers: self-attention and position-wise feed-forward [20]. The decoder layer has three sub-layers: self-attention, encoder-decoder, attention, and position-wise feed-forward as shown in Fig. 1 [21].

3.1 Encoder and decoder stack

The encoder is composed of a stack of N = 6 identical layers. Two sub-layers make up each layer. A multi-head self-attention mechanism is the first, and a straightforward feed-forward network with position-wise full connectivity is the second. implemented layer normalization after a residual connection was made around each of the two sub-layers [22]. The stack of N = 6 identical layers also makes up the decoder. Each encoder layer has two sub-layers, and the decoder adds a third sub-layer to carry out multi-head attention over the encoder stack's output. Like the encoder [20-22].

3.2 Attention

The mapping of a query and a collection of key-value pairs to an output, where the query, keys, values, and output are all vectors, is known as an attention function. The results are calculated as a weighted sum of the values, with each value's weight determined by the query's compatibility function with the relevant key [22].

3.2.1. Scaled dot-product attention

Scaled Dot-Product Attention is a fundamental mechanism in the Transformer architecture. It allows a model to focus on different parts of the input sequence, effectively capturing dependencies and relationships within the data. The core idea behind this mechanism is to compute the attention between different elements in the input data using a mathematical formulation based on dot products. Attention mechanism, three key components, queries (Q), keys (K), and values (V) are used. These are derived from the input sequence. Each query vector is dot-multiplied with all key vectors to measure the similarity between them. The dot product measures how much focus each key should get from the query [21].

3.2.2. Position-wise feed-forward networks

Position-wise Feed-Forward Networks (FFNs) are essential components of both the encoder and decoder layers. Unlike the attention mechanisms, which model interactions between tokens in a sequence, FFNs apply transformations independently to each token in a sequence.

The FFN is responsible for introducing non-linearity into the model and contributes to the representational power of the Transformer architecture. FFNs consists of two fully connected layers with a nonlinear activation function ReLU, each token in the input sequence is processed individually and independently by the same FFN. This position-wise processing ensures that each token is transformed based on its own features without any interaction with other tokens, and the final output as follows [22, 24]:

$$\mathrm{FFN}(x) = \max(0\,.xW1 + b1)W2 + b2 \quad (1)$$

The output of the point-wise FFN contributes to the enriched representation of each position in the ViT, enabling the model to learn and represent intricate features in the input image sequence. While the linear transformations are the same across different positions, they use different parameters from layer to layer [20, 24].

4. Materials and methods

4.1 Vision transformer

In the field of computer vision, the Vision Transformer (ViT) architecture was first introduced by Dosovitskiy et al., ViT have been widely adopted for various tasks, including image classification, video classification, semantic segmentation, and object detection, video object segmentation, and 3D object detection [25]. Unlike traditional convolutional neural networks (CNNs), ViT leverage a self-attention mechanism, enabling them to learn intricate relationships between different parts of an image by processing all image segments simultaneously. This global context understanding is particularly beneficial for tasks that require capturing long-range dependencies.

ViT offer several advantages over CNNs. They do not rely on image-specific biases, as they divide images into positional embedding patches, which are then processed by the transformer encoder to capture both local and global image features. ViT have garnered significant interest in the medical imaging community, where they have been applied to multiple medical imaging tasks different image segments during learning, providing interpretability and improving performance across a range of vision tasks [28].

The architecture of ViT includes residual connections after each block, facilitating the flow of information through the network without needing to pass through non-linear activations, and the MLP layer implements the classification head for image classification [21, 26].

The self-attention estimates the significance of one item with others, explicitly modelling the interactions among them for structured prediction, updating each component via global information aggregation from the entire input sequence as shown in Fig. 2. Consider a sequence of n items with d embedding dimension i.e., $X = \{x_1, x_2, \cdots, x_n\} \in R^{n \times d}$ then the aim is to capture all the interaction, encoding each entity in terms of the global contextual information by three learnable weight matrices, including Keys ($W^K \in R^{n \times d_k}$), Queries ($W^Q \in R^{n \times d_q}$) and Values $(W^V \in R^{n \times d_v})$,

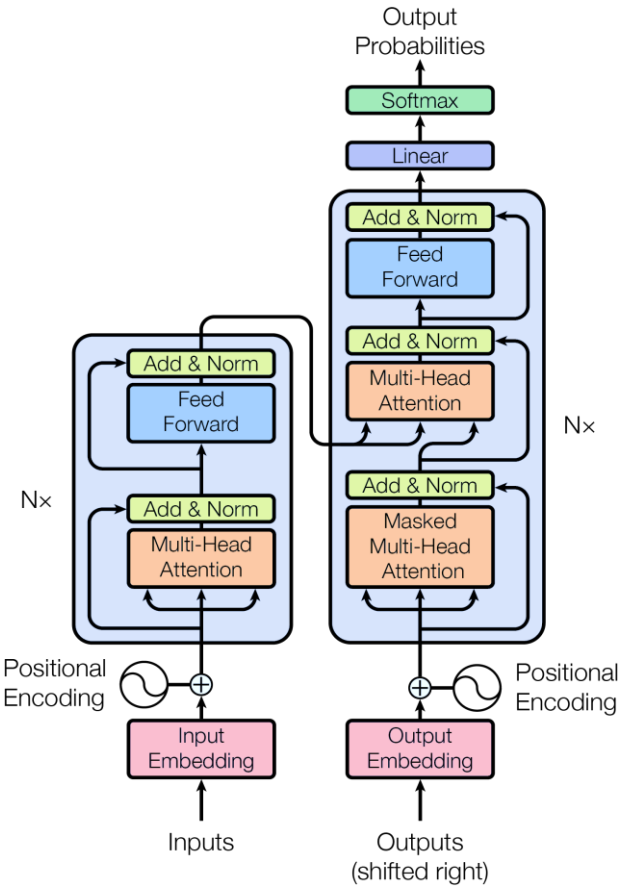

Figure. 1 Structure of the original transformer [21]

then projecting X on the mentioned matrices to obtain $K = XW^K$, $Q = XW^Q$, and $V = XW^V$, compute the matrix of $\text{Attention}(Q\,.K\,.V)$ as follows [22, 29]:

$$S_a = softmax\left(\frac{Q\,K^T}{\sqrt{d_k}}\right).v \tag{2}$$

Where $S_a \in R^{n \times d_v}$ is the self-attention layer's output achieved by computing the dotproduct of the query with all keys for a given item; furthermore, softmax is applied to get the normalized attention scores where individual items become the weighted sum of all items. It is to be noted that the attention scores provide weights [27-29].

The multi-head approach allows the model to focus on different aspects of the input sequence, effectively capturing both fine-grained and coarse-grained features. This enhanced representational capacity is a critical factor in the success of ViTs across a wide range of computer vision tasks. The integral components of transformers are Self-attention and multi-head self-attention that can be mathematically expressed as follows. WQ, WK, WV are the learned weight matrices for the query (Q), key (K), and value (V) transformations. The multi-head

attention is composed of multiple self-attention modules to capture multiple complex relationships between various items in a sequence as Eq.3, where each modules learns the weight matrices W_i^Q, W_i^K, W_i^V . At the end of multi-head attention, the h self attention modules are concatenated $[S_{a0}, S_{a1}, \cdots, Sa_{(h-1)}] \in R^{n \times h \cdot d_v}$ and then projected onto a $W \in R^{h \cdot d_v \times d}$ weight matrix [21].

$$\text{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Concat}(\text{head}_1, \text{head}_2, \cdots, \text{head}_h)\boldsymbol{W^o} \quad (3)$$

Where, W^o Represents the learnable parameters, $head_i = Attention(Q\,W_i^Q, K\,W_i^K, V\,W_i^V), i = 1,2,\ldots,h$ $Q\,W_i^Q, K\,W_i^K, V\,W_i^V$ The projections are parameter matrices.

The self-attention is applied to the same features and then concatenated. Similarly, to other sequence transduction models, learned embedding to convert the input tokens and output tokens to vectors of dimension d (model), also use the usual learned linear transformation and soft max function to convert the decoder output to predicted next-token probabilities. In this model, share the same weight matrix between the two embedding layers and the pre-soft max linear transformation, similar to [22].

4.2 Convolutional neural network

A convolutional neural network (CNN) is a type of deep learning network inspired by artificial neural networks. CNNs are typically structured as a series of stages composed of different layers. Essentially, a CNN is a multi-layer network consisting of five main layers: the input layer, convolutional layer, pooling layer, fully connected layer, and output layer. The convolutional layer contains multiple feature maps, which are generated by convolving the convolution kernel from the previous layer [30].

4.3 Transfer learning

The machine learning paradigm known as transfer learning employs a model developed for a specific task as the foundational framework for a model addressing an alternative task. This methodology proves to be particularly advantageous when there exists an insufficiency of training data for the secondary task.

Transfer learning demonstrates significant efficacy in computer vision applications, wherein the low-level features (for instance, edges and shapes) acquired by a convolutional neural network (CNN) model trained on an extensive dataset such as ImageNet can be proficiently repurposed for a diverse array of tasks including image classification, object detection, or segmentation [31]. VGG16: - A Convolutional Neural Network (CNN) architecture that was developed by researchers at the University of Oxford's Visual Geometry Group (VGG). It was introduced in 2014 and has become one of the most widely used pre-trained models for transfer learning in computer vision tasks.

VGG16: has a total of 16 weighted layers (13 convolutional layers, 5 max-pooling layers, and 3 fully connected layers) [32].

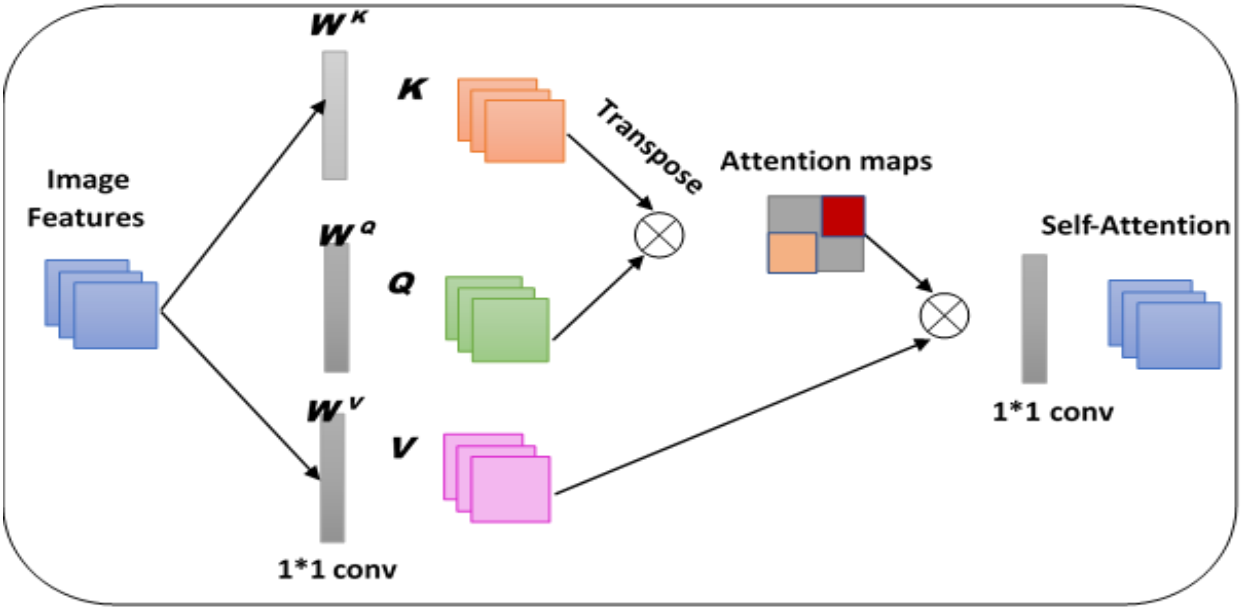

Figure. 2 Architecture of self-attention [29]

MobileNetV2: A convolutional neural network architecture developed by researchers at Google in 2018. It is an improvement over the original MobileNet architecture and is designed for efficient performance on mobile and embedded devices [33].

Inception-v3: An improved version of the original GoogLeNet (Inception-v1) architecture, developed by researchers at Google. It was introduced in 2015 and is one of the more advanced Inception-based models in the Inception family. Inception-v3 has a total of 48 layers (not including the final average pooling layer) [34].

5. The proposed hybrid model

5.1 Model's algorithm

The proposed model is to combine ViT and CNN as described in Algorithm 1, to form a new network structure, to enhance the extraction of deep local and learn global relationships between them. The Vision Transformer (ViT) and Convolutional Neural Networks (CNN) are employed in a sequential architecture, wherein the output generated by one module is directly transmitted to subsequent modules. The initial module is dedicated to the extraction of features, while the subsequent module is responsible for producing an abstract representation that is informed by the interpretation of the preceding module. Consequently, the interdependence between the CNN and ViT modules is significantly pronounced. Despite the successful implementation of vision transformers across a spectrum of visual tasks, attributable to their proficiency in capturing long-range dependencies within the input data, performance disparities persist between transformers and traditional CNNs. A principal factor contributing to this phenomenon is the transformer's inadequacy in local information extraction. In addition to the aforementioned variants of ViT that augment locality, the integration of transformers with convolutional mechanisms represents a more straightforward approach to incorporate local features into the standard transformer architecture. The initial phase of this framework involves employing Convolutional Neural Networks (CNN) to derive localized features as a substitute for unprocessed image patches; subsequently, the input sequence can be generated from the feature maps, followed by the reshaping of the feature map into patches, which are then subjected to linear projection to form embeddings. In the subsequent phase utilizing Vision Transformers (ViT), the image is divided into linear patches (tokens) that are processed through encoder blocks via linear layers to capture the global interrelationships within the images, whereby patch embedding projection is implemented on the patches extracted from the CNN feature map.

Algorithm 1:- The Hybrid Model (CNN + Vit) of BC Classification

Input: Image dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$ with images x_i and labels y_i.

Output: Predicted labels $\hat{y}_i$ for each x_i.

Begin

Step1: Import necessary libraries and modules: Import PyTorch, NumPy, Pandas, and Sklearn modules.

Step2: Load and pre-process the dataset: Load the dataset from folders,

Step3: Split the dataset into training and validation sets, and testing set.

Step4: Pre-processing:

- Normalize images x_i to size $H \times W$ and rescale pixel values.
- Optionally apply data augmentation (e.g., flip, rotate).

Step5: CNN Feature Extraction:
Extract feature maps z_i from x_i using CNN.

Step6: Define the Vision Transformer model:

- Patch Embedding: Reshape feature map z_i into patches p_i, and linearly project to embedding p_i'.
- ViT Encoding: Add positional embedding to p_i' and process through ViT using MHSA and FFN layers.
- Classification Token: Append [CLS] token to the patch embeddings, and use the [CLS] output after ViT encoding as the global image representation.
- Classification Head: Apply FFN and softmax to [CLS] to predict class probabilities $\hat{y}_i$.
- Loss Computation: Calculate cross-entropy loss $\mathcal{L}_{\text{CE}}$ between $\hat{y}_i$ and y_i. Define the optimizer (Adam) to update the model parameters.

Step7: Train the model: Train the Vision Transformer model using the training set.

- For each epoch, iterate through the training set in batches,
- Forward pass the images through the model,
- Calculate the loss
- Back propagate the gradients.

- Update the model parameters with the optimizer to minimize $\mathcal{L}_{CE}$ using Adam optimizer.

Step8: Evaluate the model on the validation set by iterating through the validation set in batches,

- Forward pass the images through the model
- Calculate the accuracy and other evaluation metrics such as SE, SP and F1 score.

Step9: Test the final model on a separate test set to evaluate its performance. For new images, repeat steps 2 to 6 for predictions $\hat{y}_{new}$.

End

Notably, as a specific instance, the patches may possess a spatial dimension of 1x1, indicating that the input sequence is derived by merely flattening the spatial dimensions of the feature map and projecting it onto the transformer dimensionality. The input embedding and positional embedding (PE) are incorporated into the feature maps as delineated in Fig.3. CNN use a convolution operation, which applies a filter or kernel to the input image to produce feature maps. This operation can be mathematically expressed as follows:

$$(F * K)(i,j) = \sum_m \sum_n I(i+m, j+n) \cdot K(m,n) \quad (4)$$

Where, I is the input image, K is the kernel, and (i,j) represents the pixel position in the output feature map.

The resulting feature maps capture local spatial features such as edges, textures, and shapes.the output of each convolutional layer is passed through an activation function (e.g., ReLU) to introduce non-linearity, allowing the network to learn complex patterns. The self- attention mechanism in ViTs, which allows the model to capture long-range dependencies across the entire image.

5.2 Mathematical overview and nomenclature

This section offers a comprehensive explanation of the formulas and variables used in the study. Table 2 provides clear definitions of the variables and symbols used in the mathematical formulas throughout the study. The main mathematical elements in the study involve the convolutional neural network (CNN) operations, Vision Transformer (ViT) processes, and the self-attention mechanism, each playing a crucial role in processing bladder cancer images as follows:

A. CNN Convolution Operation:

$$F(x,y) = \sum_i \sum_j I(x-i, y-j) \cdot K(i,j)$$

This operation defines the feature extraction process in the CNN module, where each image segment (patch) is processed with a kernel or filter matrix K to produce a feature map F. The kernel slides across the image I, accumulating weighted sums of pixel values.

B. Self-Attention Mechanism:

$$\text{Attention (Q, K,V)} = softmax\left(\frac{Q\,K^T}{\sqrt{d_k}}\right) \text{V}$$

The self-attention mechanism computes attention scores, allowing the model to focus on different image regions based on their importance. Here, queries (Q), keys (K), and values (V) are computed using learned weight matrices. The softmax function normalizes the attention scores.

Table 2. Nomenclature and Symbol Definitions

Symbol	Definition
F(x, y)	Feature map produced at coordinates (x, y) during convolution
I(x, y)	Input pixel value from the image or feature map at position (x, y)
K(i, j)	Convolutional kernel (filter) value at coordinates (i, j)
Q	Query matrix, derived from input tokens for computing attention
K	Key matrix, derived from input tokens for computing attention
V	Value matrix, derived from input tokens for computing attention
d_k	Dimension of key vectors, used to scale the dot product in self-attention
softmax	Softmax function to convert attention scores into probabilities
$head_i$	Output of the iii-th attention head in multi-head attention
W^O	Output weight matrix applied after concatenating multiple attention heads
x	Input vector to the feed-forward network (FFN)
W_1,W_2	Weight matrices in the feed-forward layers of transformer
b_1,b_2	Bias terms in the feed-forward network
CLS	Special token in ViT that aggregates information for final classification
output	Classification result after softmax transformation

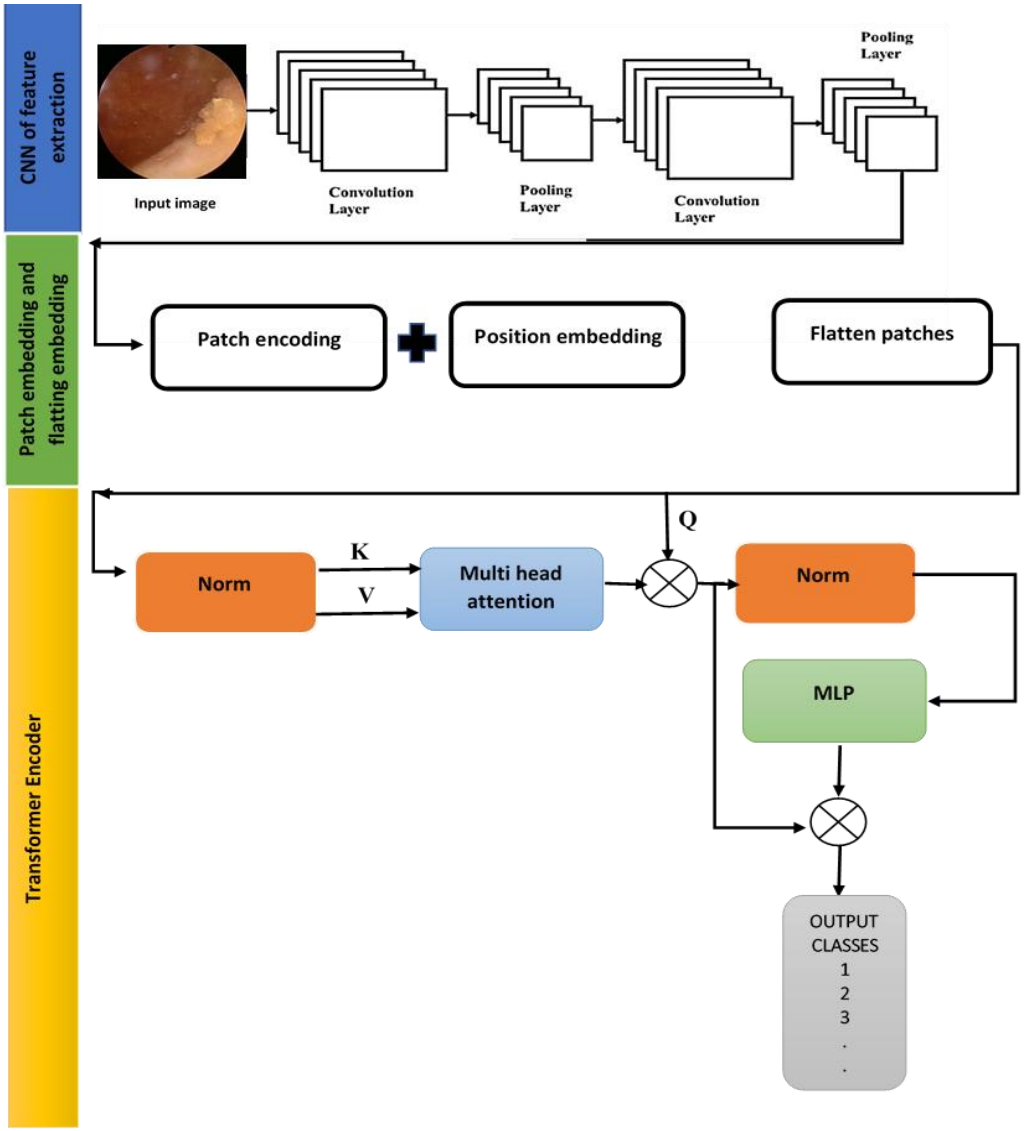

Figure. 3 Hybrid model (CNN and ViT)

C. Multi-Head Attention:

MultiHead (Q,K,V)= Concat($head_1$,…,$head_h$)W^O

Multi-head attention enables the model to capture diverse aspects of the image by applying self-attention multiple times in parallel (each as a "head") and combining results.

D. Feed-Forward Network in Transformer Layers:

$FFN(x)=ReLU(xW_1+b_1)W_2+b_2$

This layer provides additional non-linearity and capacity to learn complex features. Each token's vector is processed independently by two linear transformations with a ReLU activation in between.

E. Classification Head:

Output= softmax (FFN(CLS token))

The classification token (CLS) is passed through the final feed-forward network and softmax for classification.

6. Experimental results

6.1 The dataset

In this study, two distinct datasets were used to classification images. These datasets include a proprietary collection developed at Zagazig University in Eygpt, and a set of endoscopic images from patients undergoing clinical procedures. The comprehensive details of these datasets are as follows:

Dataset1 (Pathological):

The first dataset used in this research is a proprietary collection created by the team at Zagazig University in Egypt, authorized under the

Institutional Review Board (IRP) number 11044-22-8-2023. This dataset consists of 2,629 pathological images categorized into three classes: non-invasive malignant, invasive malignant and normal bladder mucosa, which serves as a standard for deep learning measurement [35].

Dataset2 (Endoscopic):

The second dataset comprises 1,754 endoscopic images from 23 patients undergoing Trans-Urethral Resection of Bladder Tumour (TURBT). These images are labelled based on histopathology analysis from the resected tissue. The endoscopic procedures utilize White Light Imaging (WLI) and, when available, Narrow Band Imaging (NBI). This dataset is categorized into four classes following the World Health Organization WHO and the International Society of Urological Pathology guidelines: Low-Grade Cancer (LGC), High-Grade Cancer (HGC), No Tumour Lesion (NTL) which includes cystitis and other inflammatory conditions, and Non-Suspicious Tissue (NST) [36].

6.2 Experimental setup

Code implementation was made on colab, trained all the networks in this study with cross-entropy loss and Adam optimization algorithm. In this study, divided the dataset into three distinct subsets to ensure a robust evaluation of the model's performance. Specifically, allocated 70% of the data for training, allowing the model to learn from a substantial portion of the data. The remaining 30% was further split into 10% for validation and 20% for testing. The evaluation measures used in this paper are, accuracy, precision, recall, and F1-score as follows [7]:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Sensitivity\ (Recall) = \frac{TP}{TP + FN} \tag{6}$$

$$F1\ Score = 2 * \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

Where:

TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative.

6.3 The results

In this study, the hybrid model is compared with CNN and transfer learning methods, Inception V3, VGG16, and MobileNet-V2. The performance of these models was compared through sensitivity, specificity, recall, train accuracy, validation accuracy as shown in Table 3, when using ViT alone to train dataset1. Table 4, shows the performance of the hybrid model for dataset1. Table 5 shows the performance for dataset2 by ViT alone, while Table 6 display the performance of dataset2 when using hybrid model. Table 7 compares ViT alone, MobileNet-V2, Inception-V3, VGG16, and CNN models at 50 epochs.

As shown, the hybrid model achieves superior results, with a sensitivity of 0.9438, precision of 0.9439, F1 score of 0.9437, confidence of 99.00, and accuracy of 99.93%. These metrics indicate that the hybrid model surpasses the accuracy and sensitivity of ViT alone, MobileNet-V2, Inception-V3, and VGG16.

Table 8 presents a comparative analysis of classification and segmentation accuracies achieved by previous models on Pathological and Endoscopy datasets alongside the proposed model. Table 8 includes sensitivity, specificity, and overall accuracy metrics, showcasing how various models—such as CNNs, U-Nets, GAN-based frameworks, and Vision Transformers (ViT)—have performed in past studies. This comparison underscores the effectiveness of the proposed CNN-ViT combined model, which achieves notably high accuracy rates in both datasets, thereby highlighting improvements in classification performance relative to previous models.

Table 8 provides a comparison of deep learning models used for medical image classification, demonstrating that the VIT-CNN model proposed in this study outperforms previous models. This superior performance is attributed to the integration of ConvNet and Vision Transformer architectures.

Table 3. Performance results of the endoscopy dataset using ViT.

Epoch	Train-Acc.	Train Loss	Val.-Acc.	Val. Loss	Precision	Sensitivity (Recall)	F1 Score	Confidence
10	82.62%	0.5095	67.16%	0.9653	0.5110	0.3765	0.3654	0.64
20	91.83%	0.3027	73.49%	0.5735	0.4348	0.3847	0.3885	0.80
30	91.70%	0.1977	73.60%	0.3745	0.4007	0.3333	0.3270	0.91
40	92.26%	0.1234	77.19%	0.2338	0.4881	0.3698	0.3714	0.97
50	92.28%	0.0930	76.17%	0.1763	0.4568	0.3416	0.3551	0.94

Table 4. Performance results of the hybrid model (CNN with ViT) on the endoscopy dataset.

Epoch	Train-Acc.	Train Loss	Val.-Acc.	Val. Loss	Precision	Sensitivity (Recall)	F1 Score	Confidence
10	92.23%	0.2562	92.08%	1.2865	0.9146	0.9145	0.9136	0.75
20	93.45%	0.0290	92.40%	0.3384	0.9374	0.9345	0.9354	0.93
30	92.59%	0.0160	92.50%	0.2799	0.9282	0.9259	0.9251	0.91
40	99.29%	0.0367	98.29%	0.3249	0.9441	0.9430	0.9425	0.93
50	99.91%	0.0027	99.43%	0.0067	0.9715	0.9715	0.9715	0.95

Table 5. Performance results of the Pathological dataset using ViT.

Epoch	Train-Acc.	Train Loss	Val.-Acc.	Val. Loss	Precision	Sensitivity (Recall)	F1 Score	Confidence
10	70.71%	0.5828	66.86%	2.2826	0.5041	0.5056	0.5030	0.52
20	80.43%	0.4347	75.05%	1.7024	0.5422	0.5226	0.5269	0.85
30	84.55%	0.3849	79.85%	1.5077	0.6061	0.5989	0.6018	0.74
40	89.61%	0.2664	84.38%	1.0433	0.6701	0.6215	0.6358	0.89
50	93.67%	0.1918	88.95%	0.7512	0.9033	0.9067	0.8980	0.75

Table 6. Performance results of the pathological dataset using a hybrid model (CNN combined with ViT).

Epoch	Train-Acc.	Train Loss	Val.-Acc.	Val. Loss	Precision	Sensitivity (Recall)	F1 Score	Confidence
10	89.48%	0.2850	81.95%	2.4340	0.5072	0.5000	0.4882	0.64
20	95.47%	0.0233	92.83%	1.2366	0.6591	0.6469	0.6356	0.94
30	97.41%	0.0200	94.45%	1.2006	0.7591	0.7769	0.7256	0.96
40	99.47%	0.0113	97.83%	1.2366	0.8595	0.8169	0.8056	0.97
50	99.93%	0.0083	92.56%	1.4641	0.9439	0.9438	0.9437	0.99

Table 7. Performance comparison results of different models on the Pathological dataset over 50 epochs.

Model	Train-Acc.	Train Loss	Val.-Acc.	Val. Loss	Precision	Sensitivity (Recall)	F1 Score	Time (Sec)
CNN	97.89%	0.0062	99.95%	0.0088	0.9742	0.9739	0.9740	15257
VGG16	98.33%	0.0022	96.95%	0.0338	0.9782	0.9129	0.9800	14807
MobileNet-V2	97.39%	0.0032	95.95%	0.0558	0.9672	0.9459	0.9900	14607
Inception-V3	95.81%	0.1689	96.05%	0.2039	0.9542	0.9539	0.9540	29963
ViT	93.67%	0.1918	88.95%	0.7512	0.9033	0.9067	0.8980	11209
Proposed Hybrid	**99.93%**	**0.0083**	**92.56%**	**1.4999**	**0.9439**	**0.9438**	**0.9437**	**14454**

Table 8. Comparative Accuracy Analysis of Pathological and Endoscopy Studies with Proposed Model

Dataset	Author/Year/Ref.	Model	Accuracy
Endoscopy	Zhu et al. (2020) [13]	Custom CNN	91.35%
Endoscopy	Wang et al. (2019) [14]	U-Net for segmentation, CNN for classification	86%
Endoscopy	Matsumoto et al. (2021) [17]	Deep CNN	92.5%
Endoscopy (WLI & NBI)	Lazo et al. (2023) [36]	GAN-based model	91%
The proposed model for Endoscopy		**CNN combined with ViT**	**99.91%**
Pathological	Qiu et al. (2021) [12]	ResNet50 with transfer learning	88.7%
Pathological	Liu et al. (2020) [16]	InceptionV3 with transfer learning	90.5%
Pathological	Ola et al. (2023) [35]	Vision Transformer (ViT_B32 model)	99.49%
The proposed model for Pathological		**CNN combined with ViT**	**99.93%**

The training process using a vision Transformer for BC classification real dataset would involve the following stages:

1-Dataset preparation: Begin by downloading the dataset and dividing it into training, validation, and test sets. Pre-process the images by performing tasks like resizing and enhancement images.

2- Feature extraction: using CNN of feature extractors. The last fully connected layers of the models are removed and the output of the previous

layers are used as input for new models that will be trained for the dataset.

3-Model training: Train the model using the training set and evaluate its performance using the validation set.

4-Model evaluation: Evaluate the model using the test set, employing various evaluation metrics such as accuracy, sensitivity, specificity, and F1 score.

5-Inference procedure: After the model has been trained, evaluated, and deemed satisfactory, utilize it to make predictions on new, unseen data.

It is worth noting that the vision with CNN Transformer is different from traditional CNN it can process the whole image directly.

Table 8 provides a detailed comparison of various deep learning models, specifically analyzing their performance on sensitivity, specificity, and overall accuracy across pathological and endoscopy datasets. Each model's architecture plays a pivotal role in its performance:

- CNNs: Known for their strong feature extraction capabilities through convolutional layers, CNNs excel in capturing spatial hierarchies in images. However, they may struggle with complex, high-dimensional data unless they are very deep, which can lead to overfitting.
- U-Nets: Typically used for segmentation tasks, U-Nets combine encoder and decoder structures to precisely localize and classify image pixels. Their strength lies in detailed segmentation, but they can be computationally expensive and less effective for classification tasks compared to CNNs.
- GAN-based Frameworks: GANs are exceptional in generating synthetic data and addressing data imbalance. However, their application in classification tasks can be challenging due to instability during training and the requirement for extensive data to produce high-quality synthetic samples.
- Vision Transformers (ViT): ViTs leverage self-attention mechanisms to process image patches, capturing long-range dependencies effectively. They are particularly powerful for high-resolution images and diverse datasets but may require large amounts of data and computational power for training.

The proposed CNN-ViT hybrid model combines the strengths of CNNs and Vision Transformers, resulting in superior performance metrics. This model particularly excels in scenarios involving complex, high-resolution medical images where both local and global feature extraction is crucial.

- Pathological Dataset: The CNN-ViT model achieves an accuracy of 99.91%, outperforming other models by effectively capturing intricate details and global patterns in pathological images. The CNN component efficiently extracts local features, while the ViT captures broader contextual information, enhancing overall classification accuracy and sensitivity.
- Endoscopy Dataset: With an accuracy of 99.93%, the hybrid model demonstrates its robustness in handling diverse imaging techniques like White Light Imaging (WLI) and Narrow Band Imaging (NBI). The combination of CNN and ViT ensures precise detection and classification of various tissue types and abnormalities.

The findings have significant implications for clinical practice:

- Improved Diagnostic Accuracy: Higher accuracy and sensitivity in detecting bladder cancer can lead to more reliable diagnoses, reducing the likelihood of false negatives and ensuring timely treatment.
- Enhanced Patient Outcomes: Accurate and early detection of cancerous lesions facilitates prompt medical intervention, potentially improving patient survival rates and quality of life.
- Data Utilization: The ability to leverage high-resolution images and diverse datasets enhances the diagnostic capabilities of medical imaging systems, paving the way for more advanced and reliable AI-assisted diagnostic tools.

To provide in depth evaluation of the proposed CNN-ViT hybrid model, Fig. 4 shows the confusion matrix that highlights the model's performance across different classes.

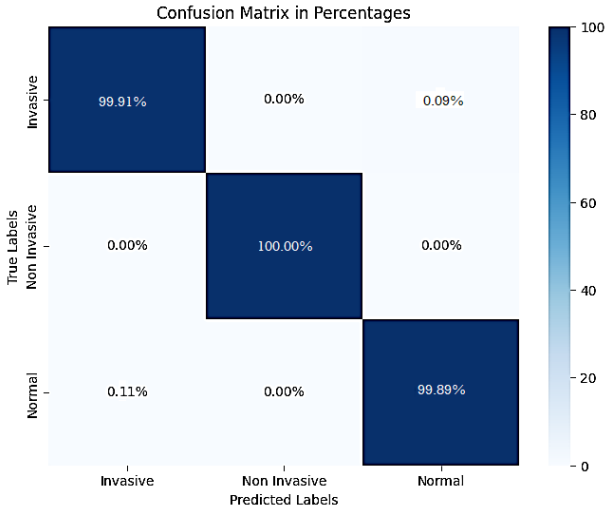

Figure. 4 The confusion matrix of the proposed model for Pathological

The confusion matrix offers a detailed analysis of true positives, true negatives, false positives, and false negatives, allowing for a more granular understanding of the model's accuracy and effectiveness. This additional analysis underscores the robustness of the proposed approach in accurately classifying bladder cancer images, further supporting the superior performance metrics discussed in the study.

The experimental results underscore the efficacy of the proposed CNN-ViT hybrid model, which outperforms existing models in terms of accuracy and sensitivity. This model's ability to integrate local and global feature extraction techniques is particularly advantageous for complex medical imaging tasks, making it a valuable tool for enhancing diagnostic accuracy in clinical settings.

7. Conclusions and future work

In the domain of global health, the rising incidence of bladder cancer necessitates the development of innovative diagnostic methodologies. This investigation examines the efficacy of hybrid Vision Transformers (ViT) in the classification of bladder cancer images. Benchmark evaluations indicate that the pre-trained hybrid ViT model attained an accuracy rate of 99.7% in multi-class classification, surpassing the ViT model by 2.3% and exceeding the performance of conventional models such as VGG16, Inception v3, MobileNetV2, and CNN, which recorded accuracy rates of 97.2%, 96.5%, and 98.33%, respectively. These results highlight the potential of Transformer-based architectures, specifically the hybrid (CNN integrated with ViT) model, in enhancing the precision of bladder cancer image analysis. The experimental outcomes illustrate the promising capabilities of the Vision Transformer in extending the frontiers of bladder cancer image analysis and advancing the contemporary standards in this scientific domain.

While this research substantiates the efficacy and promise of transformer-based classification for bladder cancer, subsequent studies could concentrate on investigating the feasibility of real-world integration, refining transfer learning methodologies, and examining multimodal strategies in light of the availability of a previously compiled multimodal bladder cancer dataset. Furthermore, it is imperative to assess the reliability, interpretability, and trustworthiness of transformer-based models in the context of bladder cancer diagnosis. Addressing these facets in future inquiries could significantly enhance the broader applicability of transformer-based architectures, thereby promoting progress in global healthcare. As delineated, the proposed model demonstrates a detection accuracy of 99.91% and a classification accuracy of 99.93% when compared to alternative methodologies. However, among more conventional models, including custom CNNs and transfer learning techniques (ResNet50 and InceptionV3), accuracy rates fluctuate between 88.7% and 93.5%, whereas alternative methodologies such as U-Net coupled with CNNs and GAN-derived models exhibit a maximum accuracy of up to 92%. The extraordinary performance of the proposed method can be attributed to the synergistic integration of aspect-based spatial features derived from CNN and global contextual insights from the enhanced ViT. The comparative analysis distinctly indicates that the proposed combined model, which incorporates both CNN and ViT, represents the most effective approach for further enhancing the accuracy of medical image analysis.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Author 1: Conceptualization, analysis, methodology, and coding, project administration, data curation, written original draft.

Author 2: Responsible for the supervision, visualization, conceptualization, review, writing review, and editing.

Author 3: Supervision, validation, formal analysis, review, writing review, and editing.

References

[1] World Health Organization. "Cancer: Bladder Cancer", *World Health Organization*, n.d. Available:https://www.who.int/cancer/prevention/diagnosis-screening/bladder-cancer

[2] R. L. Siegel, K. D. Miller, and A. Jemal. "Cancer Statistics, 2020", *CA: A Cancer Journal for Clinicians*, Vol. 70, No. 1, pp. 7-30, 2020, doi: 10.3322/caac.21590.

[3] M. Burger, J. W. Catto, G. Dalbagni, H. B. Grossman, H. Herr, P. Karakiewicz, ... & S. F. Shariat. "Epidemiology and Risk Factors of Urothelial Bladder Cancer", *European Urology*, Vol. 63, No. 2, pp. 234-241, 2013, doi: 10.1016/j.eururo.2012.07.033.

[4] M. Babjuk, M. Burger, E. Comperat, P. Gontero, A. H. Mostafid, J. Palou, ... & M. Roupret. "European Association of Urology Guidelines

on Non–Muscle-Invasive Bladder Cancer (TaT1 and Carcinoma in Situ) – 2019 Update", *European Urology*, Vol. 76, No. 5, pp. 639-657, 2019, doi: 10.1016/j.eururo.2019.08.016.

[5] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, ... & J. A. van der Laak. "A Survey on Deep Learning in Medical Image Analysis", *Medical Image Analysis*, Vol. 42, pp. 60-88, 2017, doi: 10.1016/j.media.2017.07.005.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is All You Need", In: *Proc. of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, pp. 6000–6010, 2017, doi: 10.48550/arXiv.1706.03762.

[7] D. H. Chapman-Sung, J. Chiang, S. Zhang, R. Liu, and D. Shen. "Convolutional Neural Network-Based Decision Support System for Bladder Cancer Staging in CT Urography: Decision Threshold Estimation and Validation", *Medical Imaging 2020: Computer-Aided Diagnosis*, 2020, doi: 10.1117/12.2551309.

[8] S. Sarkar, D. Kumari, V. Singh, and S. Sharma. "Performing Automatic Identification and Staging of Urothelial Carcinoma in Bladder Cancer Patients Using a Hybrid Deep-Machine Learning Approach", *Cancers (Basel)*, Vol. 15, pp. 1-15, 2023, doi: 10.3390/cancers15061673.

[9] D. Liu, S. Wang, and J. Wang. "The Effect of CT High-Resolution Imaging Diagnosis Based on Deep Residual Network on the Pathology of Bladder Cancer Classification and Staging", *Computers in Biology and Medicine*, 106635, 2022, doi: 10.1016/j.cmpb.2022.106635.

[10] G. Zhang, L. Yang, L. Zhang, and H. Sun. "Deep Learning on Enhanced CT Images Can Predict the Muscular Invasiveness of Bladder Cancer", *Frontiers in Oncology*, 2021, doi: 10.3389/fonc.2021.654685.

[11] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, and Y. Zhou. "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation", *arXiv preprint arXiv:2102.04306*, 2021, doi: 10.48550/arXiv.2102.04306.

[12] S. Qiu, et al. "Deep Learning-Based Classification of Bladder Cancer Subtypes Using Histopathological Images", *Scientific Reports*, Vol. 11, p. 8723, 2021, doi: 10.1038/s41598-021-87629-0.

[13] Q. Zhu, et al. "Automated Bladder Cancer Detection Using Cystoscopy Images with a Deep Learning Approach", *Computers in Biology and Medicine*, Vol. 126, p. 104025, 2020, doi: 10.1016/j.compbiomed.2020.104025.

[14] Z. Wang, et al. "Endoscopic Image Analysis for Bladder Cancer Detection Using a U-Net Based Approach", *Medical Image Analysis*, Vol. 55, pp. 78-87, 2019, doi: 10.1016/j.media.2019.05.001.

[15] J. Song, et al. "A Hybrid CNN-Transformer Model for Bladder Cancer Classification Using Multimodal Data from Pathological and Endoscopic Images", *Journal of Biomedical Informatics*, Vol. 131, p. 104047, 2022, doi: 10.1016/j.jbi.2022.104047.

[16] J. Liu, et al. "Transfer Learning for the Differentiation of Malignant and Benign Bladder Tumors in Histopathological Images", *Journal of Medical Imaging*, Vol. 7, No. 4, p. 045501, 2020. doi: 10.1117/1.JMI.7.4.045501.

[17] R. Matsumoto, et al. "Early Detection of Bladder Cancer Using a Deep Convolutional Neural Network on Endoscopic Images", *European Urology Open Science*, Vol. 24, p. e1168, 2021, doi: 10.1016/j.euros.2021.07.072.

[18] L. Jiang, et al. "Multi-View Convolutional Neural Networks for Bladder Cancer Stage Prediction Using Endoscopic and Pathological Images", *IEEE Journal of Biomedical and Health Informatics*, Vol. 26, No. 3, pp. 756-765, 2022, doi: 10.1109/JBHI.2022.3141512.

[19] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang. "Transformer in Transformer", *Advances in Neural Information Processing Systems*, Vol. 34, pp. 15908-15919, 2021.

[20] A. Dosovitskiy. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", *arXiv preprint arXiv:2010.11929*, 2020.

[21] A. K. Sharma and N. K. Verma. "A Novel Vision Transformer with Residual in Self-Attention for Biomedical Image Classification", *arXiv preprint arXiv:2306.01594*, 2023.

[22] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. "Training Data-Efficient Image Transformers & Distillation Through Attention", In: *Proc. of International Conference on Machine Learning*, pp. 10347-10357, 2021.

[23] W. Ullah, K. Javed, M. A. Khan, F. Y. Alghayadh, M. W. Bhatt, I. S. Al Naimi, and I. Ofori. "Efficient Identification and Classification of Apple Leaf Diseases Using Lightweight Vision Transformer (ViT)", *Discover Sustainability*, Vol. 5, No. 1, p. 116, 2024.

[24] B. Song, D. R. Kc, R. Y. Yang, S. Li, C. Zhang, and R. Liang. "Classification of Mobile-Based Oral Cancer Images Using the Vision Transformer and the Swin Transformer", *Cancers*, Vol. 16, No. 5, p. 987, 2024.

[25] S. Fayou, H. C. Ngo, Z. Meng, and Y. W. Sek. "Loop and Distillation: Attention Weights Fusion Transformer for Fine-Grained Representation", *IET Computer Vision*, Vol. 17, No. 4, pp. 473-482, 2023.

[26] D. Govindasamy. "Evaluating the Performance of Vision Transformer Architecture for Deepfake Image Classification", *University Dublin for the degree of M. Sc. in Computer Science*, 2022.

[27] Y. Hu, Y. Cheng, A. Lu, Z. Cao, D. Wei, J. Liu, and Z. Li. "LF-ViT: Reducing Spatial Redundancy in Vision Transformer for Efficient Image Recognition", In: *Proc. of the AAAI Conference on Artificial Intelligence*, Vol. 38, No. 3, pp. 2274-2284, March 2024.

[28] A. Halder, S. Gharami, P. Sadhu, P. K. Singh, M. Woźniak, and M. F. Ijaz. "Implementing Vision Transformer for Classifying 2D Biomedical Images", *Scientific Reports*, Vol. 14, No. 1, p. 12567, 2024.

[29] M. Tahir and S. Anwar. "Transformers in Pedestrian Image Retrieval and Person Re-Identification in a Multi-Camera Surveillance System", *Applied Sciences*, Vol. 11, No. 19, p. 9197, 2021.

[30] C. S. Velmahos, M. Badgeley, and Y. C. Lo. "Using Deep Learning to Identify Bladder Cancers with FGFR-Activating Mutations from Histology Images", *Cancer Medicine*, Vol. 10, No. 14, pp. 4805-4813, 2021.

[31] J. Llamas, P. M. Lerones, R. Medina, E. Zalama, and J. Gómez-García-Bermejo. "Classification of Architectural Heritage Images Using Deep Learning Techniques", *Applied Sciences*, Vol. 7, No. 10, p. 992, 2017.

[32] A. Krishnaswamy Rangarajan and R. Purushothaman. "Disease Classification in Eggplant Using Pre-Trained VGG16 and MSVM", *Scientific Reports*, Vol. 10, No. 1, p. 2322, 2020.

[33] W. Wang, Y. Li, T. Zou, X. Wang, J. You, and Y. Luo. "A Novel Image Classification Approach via Dense-MobileNet Models", *Mobile Information Systems*, 2020.

[34] C. Wang, D. Chen, L. Hao, X. Liu, Y. Zeng, J. Chen, and G. Zhang. "Pulmonary Image Classification Based on Inception-V3 Transfer Learning Model", *IEEE Access*, Vol. 7, pp. 146533-146541, 2019.

[35] O. S. Khedr, M. E. Wahed, A. S. R. Al-Attar, and E. A. Abdel-Rehim. "The Classification of Bladder Cancer Based on Vision Transformers (ViT)", *Scientific Reports*, Vol. 13, No. 1, p. 20639, 2023. doi: 10.1038/s41598-023-47731-1

[36] J. F. Lazo, B. Rosa, M. Catellani, M. Fontana, F. A. Mistretta, G. Musi, O. de Cobelli, M. de Mathelin, and E. De Momi. "Endoscopic Bladder Tissue Classification Dataset [Data Set]", *IEEE Transactions on Biomedical Engineering (TBME)*, Vol. 70, No. 10, pp. 2822-2833, 2023.