

RIA-ViT: A Residual Inception Atrous Vision Transformer for Detection of Gallbladder Cancer from CT Images

Abhishek Sehrawat¹, Anita Gupta¹, Vishnupriya² and Varun P. Gopi³

¹Chitkara School of Health Sciences, Chitkara University, Punjab, India

²Department of Electrical and Electronics Engineering, National Institute of Technology Tiruchirappalli, Tamil Nadu, India

³Department of Electronics and Communication Engineering, National Institute of Technology Tiruchirappalli, Tamil Nadu, India

Article history

Received: 23-05-2026

Revised: 20-07-2026

Accepted: 20-08-2026

Corresponding Author:

Varun P. Gopi

Department of Electronics and

Communication Engineering,

National Institute of

Technology Tiruchirappalli,

Tamil Nadu, India

Email: varun@nitt.edu

Abstract: Gallbladder Cancer (GBC) is a rare but highly aggressive malignancy of the biliary system. In early-stage gallbladder cancer, there are neither overt symptoms nor specific signs. In most patients, it is diagnosed late, resulting in poor outcomes. Even when imaging is available, diagnosing GBC remains difficult because its features overlap with those of non-neoplastic gallbladder pathologies. Furthermore, the overlap between GBC features and those of benign gallbladder conditions may require expert radiological interpretation, which is sometimes unavailable in developing countries. Therefore, the purpose of this study is to design and develop a CT-based framework for automated gallbladder cancer detection that is clinically interpretable, enabling early, consistent, and reliable detection to mitigate diagnostic variability and assist clinicians in making informed decisions. The framework incorporates a hybrid deep learning model, the Residual Inception Atrous Vision Transformer (RIA-ViT), which uses convolutional neural networks to extract local features and vision transformers to contextualize global features. The innovative design incorporates inception residual blocks for multi-scale learning, atrous convolution for enlarging the receptive field, and separable convolution for efficient design. The features are embedded and processed using transformer encoders. The proposed framework was evaluated on a well-curated dataset of CT images of GBC and normal cases, achieving an average classification accuracy of 99.56%. The model accurately separated GBC CT images from normal gallbladder CT images. The proposed framework offers reliable diagnostic outcomes with interpretable, trustworthy visual evidence. The RIA-ViT framework is a major advancement towards the development of automated systems for GBC from CT images.

Keywords: Gallbladder Cancer, Convolutional Neural Network, Vision Transformer, Residual Inception Network, CT Imaging

Introduction

Gallbladder Cancer (GBC) is the most prevalent malignancy of the biliary tract and ranks among the most aggressive gastrointestinal cancers, with its poor survival outcomes (Halaseh et al., 2022). Early symptoms are often non-specific (abdominal discomfort, nausea, weight loss, jaundice) and may mimic benign gallbladder disease, delaying diagnosis (Misra et al., 2003). GBC, at its very early stages, can either be asymptomatic or present with symptoms of benign gallbladder disease. This further compounds the problem of diagnostic delay (Dutta et al., 2016). Clinically, the greatest obstacle is not only

recognizing GBC but also distinguishing it from the normal gallbladder contour. More importantly, GBC must be differentiated from several benign conditions, including inflammatory gallbladder wall thickening, xanthogranulomatous cholecystitis, and gallbladder polyps (He et al., 2022). The differences are of great therapeutic importance; benign lesions only require conservative treatment or simple cholecystectomy, while confirmed malignancy mandates radical resection with lymphadenectomy and then oncologic follow-up. In such regions, Daly has documented that underdeveloped, low- and middle-income regions lack high-quality imaging and sophisticated radiologist diagnosis, worsening the

diagnostic deficit (Kumar et al., 2024). India, as a single country, accounts for almost 10% of the GBC cases worldwide, with a much higher prevalence in the upper and northeastern parts of the country (Kumar et al., 2024). A representative CT image of gallbladder cancer is shown in Figure 1. The dismal survival rates that continue to be the case are largely explained by late-stage diagnosis, lack of effective nationwide screening programs, and underdeveloped diagnostic imaging. While its use is widespread in the diagnosis and staging of GBC, the utility of CT scans is highly dependent on specialist interpretation. Inattention to overlapping pathological features in lesions could lead to surgical procedures being authorized for benign conditions or malignancies being missed entirely. Recent progress in Artificial Intelligence (AI) and Deep Learning (DL) has reignited interest in radiology, as these methods can reveal imaging artifacts that might be missed by humans (Litjens et al., 2017). Neural network frameworks involving Convolutional Neural Networks (CNN) and Vision Transformers (ViTs) have demonstrated the ability to describe local texture components at different scales in the spatial context of medical images, and their CNN counterparts (Dosovitskiy et al., 2021; Chen et al., 2021). The integration of AI technology remains a challenge, as such technology must be clinically validated and aligned with the requisite compliance standards. To this end, the present study aims to demonstrate the Vision Transformer with Residual Inception Atrous Convolutional Neural Network (RIA-CNN) and its combined agile architecture as a proof-of-concept for automated GBC detection in CT imaging scans.

Our current dataset comprises CT imaging scans of normal and malignant gallbladders. In the near future, we intend to expand the dataset to include benign lesions, thereby facilitating a more nuanced differential diagnosis. The salient contributions of this research are as follows:

- Proposed a new architecture for classifying GBC that integrates Vision Transformer and RIA-CNN
- The proposed architecture extracts multiple scales of local features and long-distance contextual information
- The proposed model produces an impressive slice-level classification accuracy of 99.56%
- The distinct feature separability is visualized using a t-SNE plot
- The model's interpretability is verified using Grad-CAM visualizations

AI-driven Computer-Aided Diagnosis (CAD), particularly Deep Learning (DL), has been applied to the assessment of GBC Using Ultrasound (USG), Computed Tomography (CT), and Magnetic Resonance Imaging (MRI). Although the literature is relatively limited, several studies provide benchmarks for automated

detection and characterization. USG has dominated early DL work due to its routine use and accessibility. The GBCU dataset enabled standardized comparisons. Basu et al. (2023) proposed RadFormer and reported 0.92 accuracy on GBCU Basu et al. (2023). GBCNet combined tailored training with multi-scale pooling and achieved sensitivity/specificity comparable to expert radiologists (Basu et al., 2022). The GBCHV framework introduced anatomy-aware modules for normal/benign/malignant classification (Hasan et al., 2025). Despite progress, ultrasound CAD remains sensitive to operator dependence and image quality. CT provides a richer anatomic context for detection and staging. Fujita et al. (2022) reported 98.5% accuracy (AUC 0.9985) using a residual CNN. Zhao et al. (2024) employed a 3D DenseNet for volumetric CT feature learning, and Chang et al. (2023) used GAN-based processing for low-dose CT, achieving 91.0% average precision. Gupta et al. (2025) evaluated MedSAM and TransUNet on contrast-enhanced CT and reported a mean Dice score of 0.776 and an IoU of 0.653 for MedSAM on external datasets. CT-based DL has also been applied to differential diagnosis and outcome prediction. Zhang et al. (2023) proposed a DL nomogram to distinguish XGC from GBC using Contrast-Enhanced (CE) CT. Xiang et al. (2024) classified malignant versus benign gallbladder lesions using CE-CT. Jin et al. (2025) combined CT radiomics with DL, using nnU-Net and DeepSurv for preoperative survival prediction. However, many CT studies remain limited by cohort size, protocol heterogeneity, and incomplete external validation. Transformer-based models have improved medical imaging by enhancing global context modeling. Gated axial attention improved classification on datasets such as BrainUS and MoNuSeg (Valanarasu et al., 2021).



Fig. 1: Representative CT scan image of gallbladder cancer

Furthermore, TransUNet demonstrated that hybrid CNN encoders with transformer modules can enhance segmentation by jointly modeling local texture and long-range dependencies (Chen et al., 2021). Related hepatobiliary studies further support the use of imaging-based decision support for DL. Pang et al. trained YOLOv3 on 220,000 CT images for gallstone classification (Pang et al., 2019). Jeong et al. reported an AUC up to 0.92 for USG-based gallbladder polyp differentiation (Jeong et al., 2020). Jalali et al. proposed ResBCDU-Net for biliary-structure segmentation (Jalali et al., 2021), and Sugimoto et al. combined imaging and biomarkers for distal bile duct obstruction diagnosis (AUC 0.908) (Sugimoto et al., 2023). Despite the remarkable success of transformer-based architectures in medical image analysis, hybrid CNN-Transformer frameworks specifically designed for CT-based Gallbladder Cancer (GBC) detection remain largely underexplored. Existing studies have primarily focused on either conventional convolutional neural networks or standalone transformer architectures, with limited attention to developing lightweight hybrid models that effectively capture both local morphological features and long-range contextual dependencies while remaining computationally efficient. This research gap motivates the development of the proposed Residual Inception Atrous Vision Transformer (RIA-ViT) framework for automated GBC classification.

While many recent studies have incorporated Convolutional Neural Networks (CNNs) and transformers for medical image analysis, most existing hybrid architectures incorporate CNNs with transformer encoders and do not consider broader issues of representation at different scales, receptive fields, and computational efficiency in one integrated architecture. Our RIA-ViT goes beyond traditional architectures by incorporating depthwise separable convolutions, residual inception, and deep convolutions with atrous prior to transformer encoding. Depthwise separable convolutions reduce the number of parameters, residual inception modules capture multi-scale features, and atrous convolutions increase the receptive field without adding parameters. The transformer encoder then captures long-range contextual dependencies to add further information to the feature representation. We believe that our design simultaneously achieves efficient multi-scale feature representation and global contextual understanding in a lightweight and low-cost architecture that is amenable to clinical deployment.

Methods

The proposed model presents a new hybrid deep learning framework, RIA-ViT, that aims to combine the differential strengths of CNNs and ViTs. The

convolutional backbone of the network utilizes Inception-Residual modules to extract features at multiple spatial scales in conjunction with atrous (dilated) convolution, which expands the receptive field without compromising spatial resolution. To decrease processing complexity and improve computational efficiency, depthwise separable convolutions, which reduce the number of computations per layer, are implemented. With these aspects, the network can accurately capture the local morphological characteristics relevant to GBC, such as wall irregularities, polypoid structures, and mural thickening. After extracting feature maps, the system creates sequence embeddings and sends them through multiple Transformer encoder layers. These layers feature multi-head attention, which helps learn long-range spatial and global interrelations, both contextually and globally, at the organ and perilesional levels. The final output of the model is produced by a fully connected classification layer that follows a Global Average Pooling (GAP) operation, from which a diagnostic prediction is made. The complete RIA-ViT model architecture, along with the data flow, is illustrated in Fig. 2.

Unlike existing hybrid CNN-Transformer architectures, such as Swin Transformer and MedViT, the proposed RIA-ViT framework integrates depthwise separable convolutions, residual inception modules, atrous convolutions, and transformer attention within a unified feature-learning pipeline. These components perform complementary roles, where depthwise separable convolutions improve parameter efficiency, residual inception modules capture multi-scale morphological characteristics, atrous convolutions enlarge the receptive field without increasing parameter complexity, and transformer attention models long-range contextual dependencies. This coordinated integration enables more discriminative feature representation while maintaining substantially lower computational complexity, thereby providing an improved balance between diagnostic performance and computational efficiency for gallbladder cancer classification.

Initial Convolutional Block

The initial block consists of separable convolution, batch normalization, ReLU activation, and max pooling. Separable convolution improves efficiency by decomposing a standard convolution into depthwise and pointwise operations (Howard et al., 2017; Chollet, 2017).

Depthwise convolution applies a spatial kernel to each channel independently:

$$X_{depthwise}^{(k)}(i, j) = \sum_{m=0}^{k_h-1} \sum_{n=0}^{k_w-1} W_{depthwise}^{(k)}(m, n) \cdot X_{input}^{(k)}(i + m, j + n) \quad (1)$$

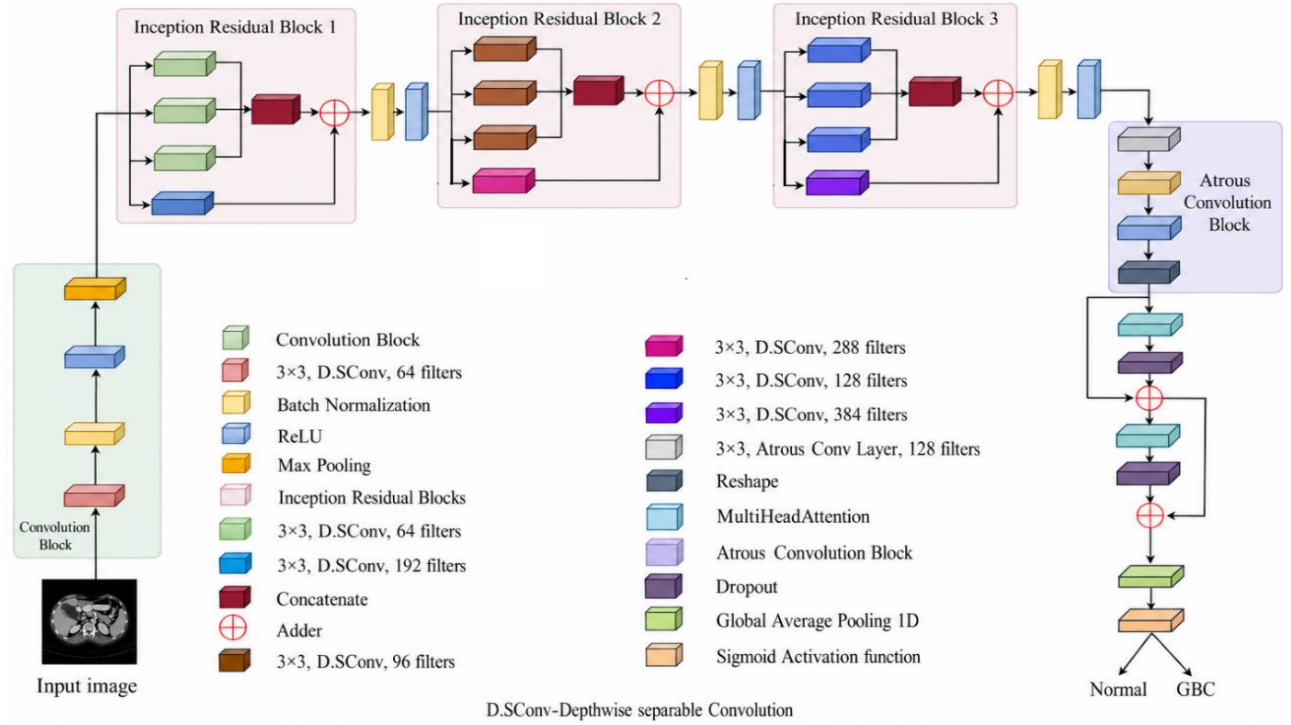


Fig. 2: Proposed model architecture

Where k indexes the input channel, $k_h \times k_w$ is the kernel size, $W_{depthwise}^{(k)}$ are the filter weights, and $X_{input}^{(k)}$ is the k -th channel of the input.

Pointwise convolution then aggregates features across channels:

$$X_{separable}(i, j, c') = \sum_{k=0}^{C-1} W_{pountwise}^{(c',k)} \cdot X_{depthwise}^{(k)}(i, j) \quad (2)$$

Where C is the total number of input channels and $W_{pountwise}^{(c',k)}$ are the 1×1 convolution weights mapping channel k to output channel c' .

The Inception Residual Block

The Inception Residual Block is an original part of a deep learning architecture to perform multi-scale convolutions with a residual connection. To ensure multiple-scale features are captured and to reduce the training of complex deep networks, the design solution proposes multi-scale feature extraction and gradient flow improvement.

Multi Scale Convolutions

In the same convolutional layer, multiple filters can work on different scales. A convolutional network layer can thus distinguish details at multiple scales in the input

to the layer. Large kernels capture large contextual and structural segments. Smaller kernels capture fine details and more local segments. Therefore, multiple convolutional layers operating at different scales are a good way to represent and understand a very complex region in an image. Thus, in the RIA-ViT model, the residual inception module uses parallel convolutional branches (usually 1×1 , 3×3 , and 5×5). Each of these branches converts the input to a feature map, and the parallel branches' outputs are combined to get a multi-scale representation of the input. The convolutional weights, given a tensor input X , can be expressed as:

$$F_{1 \times 1} = \text{Conv}_{1 \times 1}(X)$$

$$F_{3 \times 3} = \text{Conv}_{3 \times 3}(X)$$

$$F_{5 \times 5} = \text{Conv}_{5 \times 5}(X)$$

Here, $\text{Conv}_{k \times k}$ represents a convolutional operation using a kernel of size $k \times k$, and $F_{k \times k}$ denotes the corresponding feature map generated by that operation.

Concatenation and Residual Addition

In the proposed RIA-ViT framework, residual connections are incorporated within the Inception module to preserve low-level morphological information while

facilitating stable gradient propagation during optimization. Learning residuals means that instead of learning entire functions directly, the neural network can learn the residual mappings, which are the differences between the output and the function's input. Learning mappings enables the network to learn functions more efficiently. These connections were constructed to enhance the network. They promote efficient gradient flow, stabilize randomly oscillating networks, and facilitate the network's learning of deep structures. Now, the multi-feature maps coming from the preceding stage are summed up as follows:

$$F_{concat} = Concat(F_{1 \times 1}, F_{3 \times 3}, F_{5 \times 5}) \quad (4)$$

To align the dimensionality for residual learning, a 1×1 convolution is applied to the input tensor X :

$$F_{shortcut} = Conv_{1 \times 1}(X) \quad (5)$$

The residual output is then obtained through elementwise addition:

$$F_{residual} = F_{concat} + F_{shortcut} \quad (6)$$

The user information and the transformed data are fused during this step, which augments the representation while preserving some spatial aspects of the input.

Normalization and Activation

Once the residuals are added, Batch normalization (BatchNorm) is performed on the feature map obtained as a result of the residuals, $F_{residual}$, in order to control and stabilize the spread of activations and boost the speed of convergence of the network. Batch normalization is a technique that normalizes a network's activations and feature statistics to ensure numerical stability and prevent deep network problems such as vanishing and exploding gradients. In deep networks, activations capture the complexity of structure and dependencies through nonlinear functions. Within the stated neural network, the simple way to obtain the final block output is defined as:

$$Y = ReLU(BatchNorm(F_{residual})) \quad (7)$$

The Inception-Residual block structure is repeated three times in the model, with each repetition consisting of three parallel 3×3 convolutional layers. The number of filters per block increases consecutively from 64 to 96 and then to 128. The model's structural design enables it to capture high-level features while retaining the spatial resolution of details.

Atrous Convolution Block

After the Inception-Residual module, an Atrous (dilated) convolution block is used to increase the receptive field without increasing the parameter count or

altering the balance between the model's computational cost and complexity. In an atrous convolution, a dilation rate is introduced, which is labeled r and indicates the spacing between the elements of the kernel:

$$Y(i, j) = \sum_{m=0}^{k_h-1} \sum_{n=0}^{k_w-1} W(m, n) \cdot X(i + r \cdot m, j + r \cdot n) \quad (8)$$

After the atrous convolution operation, the result Y is typically passed through batch normalization and an activation function, such as ReLU, to improve training stability and introduce non-linearity. The combined operation is expressed as:

$$Y_{atrous} = ReLU(BatchNorm(Y)) \quad (9)$$

Reshape Layer

After feature extraction, the 3D feature map is reshaped into $B \times T \times d$ to form a sequence suitable for transformer processing. This transformation enables the self-attention mechanism to capture long-range dependencies and contextual relationships across all tokens within the input sequence.

Multi-Head Attention (MHA) Block

Multi-Head Attention (MHA) blocks are used in conjunction with other techniques to capture long-range spatial dependencies in high-dimensional input data. In the case of high-dimensional medical images, MHA allows for simultaneous attention across various spatial locations (Vaswani et al., 2017). The architecture integrates an MHA block, as depicted in Figure 3. Within the proposed RIA-ViT framework, the MHA block models long-range contextual relationships among the extracted convolutional features. A given input sequence is represented as $X \in R^{N \times d}$, where d is the embedding dimension, N is the number of tokens, and MHA projects X into queries Q , keys K , and values V via learned linear transformations:

$$Q = XW^Q, K = XW^K, V = XW^V, W^Q, W^K, W^V \in R^{d \times d_k} \quad (10)$$

Each attention head computes a scaled dot-product attention:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (11)$$

To increase modeling capacity, MHA computes h such heads in parallel and concatenates the outputs:

$$MHA(X) = Concat(head_1, \dots, head_h)W^O, W^O \in R^{hd_k \times d} \quad (12)$$

Where $head_i = Attention(Q_i, K_i, V_i)$, W^O is the output projection weight matrix. This allows the network to learn different kinds of relationships at different places and in different subspaces. Each attention head's output is

stitched together, then sent to a dense layer to merge the multi-head context. A skip connection is added to preserve information from the input and improve gradient flow (Devlin et al., 2018).

Classification Head

The MHA block's output is passed to the GAP layer after dropout. This output is then passed to the sigmoid layer for classification.

Experimental Settings

The proposed method uses the TensorFlow backend in the Keras Python Deep Learning Library. Model evaluations are performed on a high-performance workstation equipped with an Nvidia GeForce RTX 3090 Ti GPU with 24 GB of memory. The proposed model and optimization parameters are shown in Table 1.

Before training, all CT images were resized from 512×512 pixels to 200×200 pixels using bilinear interpolation to match the network input size. Pixel intensity values were normalized to the range $[0,1]$ to facilitate stable optimization during training. No manual segmentation or handcrafted feature extraction was performed. To improve model robustness and reduce overfitting, online data augmentation was applied exclusively to the training dataset, including random horizontal flipping, random rotation ($\pm 10^\circ$), zooming (0.9-1.1), width and height translation (10%), and brightness variation ($\pm 15\%$). The validation and test datasets were not augmented to ensure an unbiased assessment of the model's generalization performance.

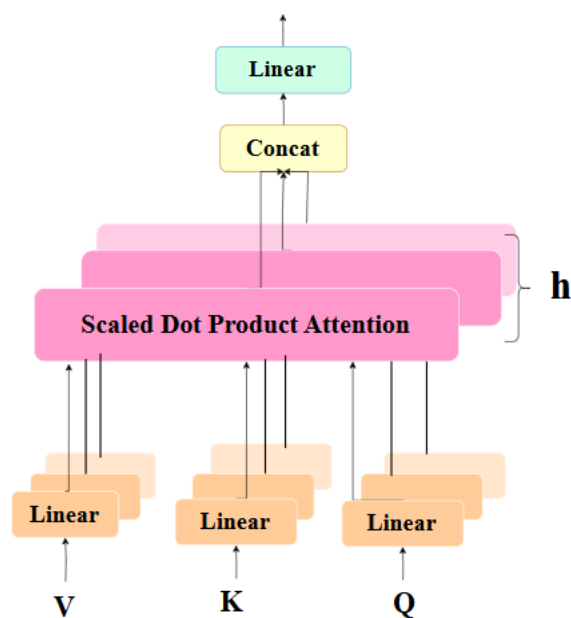


Fig. 3: Multi Head Attention (MHA) block

Table 1: Model and Optimization Parameters

Parameter	Value / Description
Input Image Size	200×200 pixels
Multi-Head Attention Heads	4
Atrous Convolution Filters	128
Atrous Convolution Dilation Rate	2
Transformer Encoder Layers	2
Weight Initialization	Glorot Uniform
L ₂ Regularization Coefficient	1×10^{-4}
Optimizer	Adam
Learning Rate	0.001
Batch Size	8
Number of Training Epochs	128
Early Stopping	Enabled

Overfitting Prevention

Several measures were adopted to minimize overfitting during model training. First, data partitioning was performed on a patient-wise basis, ensuring that CT slices from the same patient were never shared across the training, validation, and test datasets, thereby preventing information leakage. Second, L₂ regularization (1×10^{-4}), dropout, and early stopping were incorporated during optimization to improve model generalization and reduce the risk of overfitting. Third, the dataset contained balanced numbers of normal and biopsy-confirmed GBC subjects, thereby minimizing class imbalance bias during training.

Dataset Description

This study uses CT images from patients with cancer of the gallbladder (GBC) and healthy control subjects collected at MGM Medical College, Indore, Madhya Pradesh, India. We conducted this research according to the ethical standards expressed in the Declaration of Helsinki (1964) and its later amendments. For all patients, their corresponding scans were performed using a Somatom go. Top CT Scanner. For all scans, the acquisition protocol was set as follows: 100 kV, pitch of 0.6, and slice thickness of 1 mm. The system offers approximately isotropic spatial and temporal resolution of 0.33 mm and 75 ms, respectively, for the visualization of the gallbladder anatomy. The study cohort was predominantly male, and approximately 80% of the participants were around 45 years of age. All GBC cases were histopathologically confirmed, thereby providing reliable diagnostic ground-truth labels. The present study focused on binary discrimination between GBC and healthy control cases and did not attempt to grade or stage the malignancy.

A radiographer, under the guidance of a radiologist, initially annotated each CT image as either GBC or normal. The radiologist subsequently validated all annotations to ensure diagnostic accuracy and consistency. Ethical approval for the study was obtained from the Institutional Ethics Committee of Chitkara

University (Approval No. CCP/CU/IEC/25/134). Informed consent was obtained from every participant and/or their legally authorized representatives before inclusion in the study. The CT images were stored in Tagged Image File Format (TIFF) with an original resolution of pixels. Each image was independently reviewed by a radiologist to delineate the cancer and normal gallbladder regions of interest, with corresponding ground-truth masks.

Inclusion and Exclusion Criteria

Predefined eligibility criteria were applied independently to the GBC and healthy control groups to establish a clinically well-defined study cohort.

Inclusion Criteria for the GBC Group

Participants were included if they had histopathologically confirmed primary gallbladder adenocarcinoma, were aged ≥ 18 years, had pre-treatment contrast-enhanced abdominal CT images acquired during the portal venous phase, had CT images acquired using the standardized imaging protocol with adequate diagnostic quality, and had complete clinical, imaging, and histopathological records.

Exclusion Criteria for the GBC Group

Participants were excluded if they had undergone surgery, chemotherapy, radiotherapy, or other cancer-related treatment before CT acquisition; had recurrent GBC or metastatic involvement of the gallbladder originating from another primary malignancy; had CT images with severe motion artifacts, excessive noise, or incomplete anatomical coverage; had incomplete clinical, pathological, or imaging records; or had a concurrent hepatobiliary malignancy or significant anatomical abnormality that could confound image interpretation.

Inclusion Criteria for the Healthy Control Group

Participants were included if they had no clinical or radiological evidence of gallbladder malignancy, were age- and sex-matched with the GBC cohort, had contrast-enhanced abdominal CT images acquired using the same scanner and imaging protocol, exhibited a normal gallbladder appearance without focal lesions or suspicious wall thickening, and had complete demographic and imaging records.

Exclusion Criteria for the Healthy Control Group

Participants were excluded if they had a history of GBC or other hepatobiliary malignancy, had undergone previous hepatobiliary surgery, had CT images of inadequate diagnostic quality or with significant artifacts, exhibited major gallbladder abnormalities such as severe inflammatory disease, large focal lesions, or

congenital anomalies, or had incomplete demographic or imaging data.

Using the provided eligibility criteria, the final study cohort included 762 participants, with 381 healthy controls and 381 patients with biopsy-confirmed GBC. In order to minimize any class imbalance bias, the number of participants was balanced between the two groups. This was done prior to developing the model. The CT slices of each participant were isolated and then assigned to the training, validation, or test sets in order to prevent information leakage. The CT slices were divided so that 80 percent of the CT slices were allocated to the training set, 10 percent to the validation set, and the remaining 10 percent to the test set. More information regarding the participants and the distribution of CT slices is included in Table 2. The patient-wise partitioning of the data ensured that the training set, validation set, and test set were independent from one another and avoided classification performance being over-boosted by classifying anatomically related slices of the same patient.

Although dataset partitioning was performed at the patient level, the proposed RIA-ViT framework was trained and quantitatively evaluated using individual CT slices. Each CT slice contains distinct anatomical and pathological information that contributes to discriminative feature learning. Slice-wise training therefore enables the network to capture fine-grained morphological variations that may not be consistently visible across all slices of an individual examination while making effective use of the available slice-level annotations. From a clinical perspective, however, radiological diagnosis is ultimately established at the patient level by integrating information across the complete CT examination rather than relying on an individual CT slice. Therefore, the performance reported in the present study represents slice-level classification under a strictly patient-wise independent data-partitioning protocol and should not be interpreted as patient-level diagnostic performance. Future studies will investigate patient-level diagnosis by aggregating slice-level predictions using majority voting, probability averaging, or other clinically relevant fusion strategies to more closely reflect routine radiological practice.

Evaluation Metrics

Basic classification metrics, including accuracy, recall, F1-score, Cohen's kappa, and the area under the Receiver Operating Characteristic curve (AUC-ROC), were used to evaluate the model. These metrics bring together detail and a comprehensive understanding of the model and, thus, its performance.

Table 2: Dataset distribution

Subset (Participants)	GbC	Normal
Train (305)	18,300 slices	18,300 slices
Validation (38)	2,280 slices	2,280 slices
Test (38)	2,280 slices	2,280 slices

Statistical Analysis

To quantify the statistical uncertainty associated with the classification performance of the proposed RIA-ViT framework, 95% Confidence Intervals (CIs) were estimated using non-parametric bootstrap resampling of the test-set predictions. The test samples were resampled with replacement for 1,000 bootstrap iterations, with the original test-set size maintained in each iteration. For each bootstrap sample, the performance metrics, including accuracy, precision, recall, F1-score, AUC, and Cohen's kappa, were recalculated to obtain an empirical bootstrap distribution for each metric. The 2.5th and 97.5th percentiles of the resulting bootstrap distributions were used as the lower and upper bounds, respectively, of the corresponding 95% CIs. The resulting performance estimates and their 95% CIs are reported in Table 5 in the Results section.

Results

According to Table 3, RIA-VT performed well when tested against slice-wise test data. Of the 2280 GBC slices, only 20 were classified as normal when they should have

been classified as abnormal. All other 2260 normal slices were classified correctly. Additionally, there were no false negatives for normal-class slices, proving the model's reliability for classifying normal tissue (healthy tissue). The classification accuracy of the entire framework was 99.56%. The framework demonstrated the classification of CT slices as either cancerous or non-cancerous with adequate generalization and robustness. According to Table 4, both classes achieved a precision, recall, and F1-score of 1.0, indicating highly efficient classification of both classes with an ideal balance of sensitivity and specificity. Also, the model was able to achieve an AUC of 1.0000, meaning it was able to successfully discriminate GBC from normal, with an optimal discrimination score. Relative to the Quadratic Kappa Score (QKS), it was able to achieve an agreement of 0.9912. In conclusion, the RIA-ViT model is able to accurately diagnose GBC a large majority of the time.

Table 3: Confusion matrix

Targeted Class	Predicted: GBC	Predicted: NORMAL
GBC	2260	20
NORMAL	0	2280

Table 4: Classification report

Model	Prec. (%)	Rec. (%)	F1 (%)	Acc. (%)	AUC	QKS
GBC	100	99.12	99.56	99.12	–	–
NORMAL	99.12	100	99.56	100	–	–
Weighted Avg	99.56	99.56	99.56	99.56	1.0000	0.9912

To gauge the statistical uncertainty related to the performance metrics reported, 95% Confidence Intervals (CIs) for the primary evaluation metrics were computed. The results are presented in Table 5. All evaluation metrics showed relatively small 95% confidence intervals, which indicates that for the RIA-ViT framework, statistical uncertainty was low and the performance was consistent. Highly stable classification was evidenced by narrower confidence intervals for accuracy, precision, recall, and F1-score, as well as a high Cohen's Kappa. In Figure 4, it is shown that both training and validation loss were tending to zero rapidly, and the accuracy was increasing with no diverging behavior. The gap between training and validation not concentrating in the center indicated that the RIA-ViT framework was not memorizing the training data and was instead learning in a generalized manner.

Table 5: Performance evaluation of the proposed model with 95% confidence intervals

Metric	Mean	95% Confidence Interval
Accuracy	0.99564	[0.99364, 0.99737]
Precision	0.99568	[0.99372, 0.99738]
Recall	0.99564	[0.99364, 0.99737]
F1-score	0.99564	[0.99364, 0.99737]
AUC	1.00000	[1.00000, 1.00000]
Cohen's Kappa (κ)	0.99127	[0.98728, 0.99474]

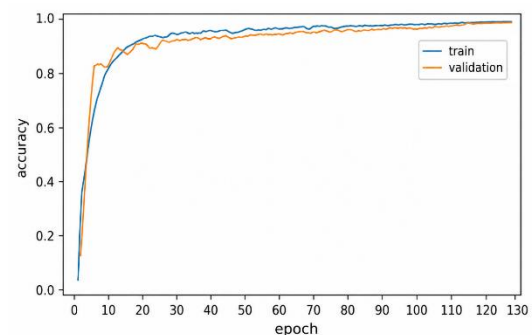


Fig. 4: Training and validation accuracy curves of the proposed RIA-ViT model over 128 epochs

Along with the regularization strategies and patient-wise evaluation, it is indicated that the reported classification performance was not high due to overfitting.

Training Convergence Analysis

The training and validation accuracy curves of the proposed RIA-ViT model over 128 epochs are presented in Fig. 4. A rapid improvement in accuracy is observed during the initial training phase, followed by gradual convergence as training progresses. After approximately 20 epochs, both curves remain consistently above 90%

and continue to improve, ultimately approaching 99% accuracy. Although minor fluctuations are observed in the validation accuracy during the early and intermediate epochs, the validation curve closely follows the training curve throughout the optimization process. The absence of a persistent gap between the training and validation accuracies indicates stable convergence and suggests that substantial overfitting was effectively controlled.

Discussion

Ablation Study

An ablation study listed in Table 6 primarily focused on the effectiveness of data streams in the components of the RIA-ViT architecture for GBC detection. The focus was on the effectiveness of the Inception Residual building blocks, atrous convolution, and MHA relative to the rest of the model. In this respect, different configuration sets were created and analyzed on the same data splits and hyperparameters. The integration of all three constituents in the proposed approach yields a substantial increase in accuracy and robustness, demonstrating the importance of each constituent in the overall performance of the GBC detection system.

Comparison With Other Standard Architectures

To gauge the effectiveness of the proposed RIA-ViT framework, a comparison was performed with several state-of-the-art CNN- and transformer-based architectures, including VGG-16, VGG-19, ResNet-50, DenseNet-121, MobileNetV2, Inception-v3, Swin Transformer, and MedViT-T. For a fair comparison, all baseline models presented in Table 7 were trained and evaluated using the same GBC dataset, identical patient-wise data partitioning, image preprocessing pipeline, training protocol, hyperparameter settings, and evaluation metrics, thereby ensuring a consistent and unbiased performance comparison.

The classification results are summarized in Table 7. Among the traditional CNN-based models, Inception-v3 demonstrated the best performance with an accuracy of 96.78%, and MobileNetV2 achieved 94.26%. For Inception-v3, the multi-scale inception modules enable it to extract different levels of image features, achieving an F1-score of 96.75%, as well as precision of 97.0% and recall of 96.50%. MobileNetV2 built an impressive AUC of 0.96 and displayed strong predictive ability. This makes it a good choice for constrained clinical situations. DenseNet121 and ResNet50 showed accuracies of 93.40% and 91.86%, respectively, and the classical VGG16 and VGG19 showed even poorer results. All of these results illustrate the inability of the traditional convolutional architectures to adequately learn the many complex and diverse structural phenomena observed in gallbladder CT images.

Among the transformer-based architectures, Swin Transformer achieved an accuracy of 97.85%, precision of 97.90%, recall of 97.70%, and an F1-score of 97.80%, demonstrating the effectiveness of hierarchical window-based self-attention in capturing long-range contextual information. MedViT-T further improved the classification performance, achieving an accuracy of 98.42%, precision of 98.50%, recall of 98.30%, and an F1-score of 98.40%. The superior performance of MedViT-T can be attributed to its hybrid convolution-transformer design, which effectively combines local feature extraction with global contextual learning. Nevertheless, under the same experimental conditions, the proposed RIA-ViT achieved the highest classification accuracy of 99.56%, together with precision, recall, and F1-score of 99.56%, while attaining an AUC of 1.00. These results indicate that the proposed integration of residual inception modules, atrous convolutions, and transformer attention enables more discriminative feature learning than either conventional CNNs or existing transformer-based architectures.

Table 6: Ablation study performance comparison

Configuration	Acc.	Prec.	Rec.	F1	AUC	QKS
Baseline	93.40	93.00	92.50	92.75	0.9600	0.9200
w/o Inception Res.	96.50	96.00	95.50	95.75	0.9800	0.9600
w/o Atrous Conv.	97.00	96.50	96.00	96.25	0.9850	0.9650
w/o MHA	98.50	98.00	97.50	97.75	0.9900	0.9800
Full Model	99.56	99.56	99.56	99.56	1.0000	0.9912

Table 7: Comparison with recently published methods and different pre-trained models

Method	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)	AUC
VGG16 Kaur and Gandhi (2019)	76.18	75.00	77.50	76.23	0.82
VGG19 Simonyan and Zisserman (2014)	78.77	77.50	79.00	78.24	0.84
Xception Chollet (2017)	87.91	88.00	87.50	87.75	0.90
ResNet50 McAllister et al. (2018)	91.86	92.00	91.50	91.75	0.93
Inception-V3 Wang et al. (2019)	96.78	97.00	96.50	96.75	0.97
DenseNet121 Jaiswal et al. (2021)	93.40	93.50	93.00	93.25	0.95
MobileNetV2 Sandler et al. (2018)	94.26	94.50	94.00	94.25	0.96
Swin Transformer Liu et al. (2021)	97.85	97.90	97.70	97.80	0.98
MedViT-Manzari et al. (2023)	98.42	98.50	98.30	98.40	0.99
Proposed Method	99.56	99.56	99.56	99.56	1.00

The integration of depthwise separable convolutions, residual inception modules, atrous convolutions, and ViT attention elegantly solves the challenges of stereotypical architectures of CNNs and Transformers, and allows RIA-ViT to reach SOTA performance. Compared with the typical hybrid CNN-Transformer architectures that use a standard CNN backbone and a Transformer encoder, RIA-ViT integrates a variety of techniques, such as parameter-efficient feature extraction, multi-scale representation, receptive field expansion, and modeling global context. Depthwise separable convolutions reduce the computational costs, and convolutional inception modules extract features from multiple contexts. Atrous convolutions enlarge receptive fields with very few additional parameters. Finally, transformer attention models contextual dependencies. The proposed RIA-ViT is suitable for actual clinical use because of its performance and competitive efficiency relative to current architectures. While the proposed RIA-ViT model achieved such promising results that classification accuracy and AUC were almost perfect (99.56%), this performance should still be regarded as internal validation and should not make strong claims of generalizability in the clinic. Several factors provide complementary explanations for the excellent classification performance of the RIA-ViT model. First, all gallbladder cancer cases were confirmed by histopathological biopsy, resulting in ground-truth labels of high certainty. Second, a patient-wise data partition confined CT slices to exactly one of the three mutually exclusive sets of training, validation, or test data, thereby preventing data leakage. Third, a consistent contrast-enhanced CT protocol ensured equal numbers of both normal subjects and biopsy-proven GBC subjects in the dataset. Finally, the proposed RIA-ViT model integrates depthwise separable convolutions, residual inception modules, atrous convolutions, and the transformer beneath an attention layer. While these factors contribute to excellent internal validation, the

robustness and clinical generalizability of the framework will only be established after further independent multi-centric external validation.

Comparison of Computational Complexity

As given in Table 8, the proposed method strikes a good balance between computational cost and inference performance when compared to the other pretrained networks. The model is made up of only 0.96 M parameters and takes up only 0.958 GFLOPs, which significantly decreases the model's complexity and reduces the memory footprint, with a model size of just 9.20 MB. The method was also able to train the model with the shortest training time of 2.10 s/epoch and an inference latency of 14.06 ms/image. This gave the proposed method a throughput of 71.13 images/s. This is also one of the fastest throughputs of any of the compared architectures. Although the proposed method requires a comparatively higher (than some lightweight models) amount of GPU memory of 2.90 GB, based on the other metrics, the trade-offs of memory, time, speed, and throughput make the proposed method an ideal choice for real-world applications with resource constraints.

t-SNE-Based Analysis of Learned Representations

The t-SNE visualization in Figure 5 shows the distribution of learned deep feature representations for the two classes, GBC and Normal. GBC samples are predominantly and clearly clustered on one side of the embedding space, while normal samples are on the other, with strong class separability. While there is a small region of overlap near the class boundary, the separation suggests that the model has learned features that are discriminatively effective in differentiating diseased from normal cases. This qualitative proof of separation strongly corroborates the quantitative results, demonstrating that the feature space constructed is indeed of high quality.

Table 8: Comparison of the computational cost of the proposed model with pretrained networks

Model	Params (M)	GFLOPs	GPU Memory (GB)	MW (MB)	Training Time/Epochs (s)	Inference Time (ms/Image)	Throughput (Images/s)
VGG16 Kaur and Gandhi (2019)	14.71	15.50	1.62	57.38	12.29	21.02	60.69
VGG19 Simonyan and Zisserman (2014)	20.02	19.60	1.76	77.63	15.48	24.33	69.76
Xception Chollet (2017)	20.86	11.00	1.60	87.42	11.26	61.79	16.18
ResNet50 McAllister et al. (2018)	23.59	3.80	9.23	226.92	10.84	68.47	14.60
Inception-v3 Wang et al. (2019)	21.80	7.35	1.60	87.42	10.69	120.87	8.27
DenseNet121 Jaiswal et al. (2021)	7.03	2.88	1.60	32.00	15.85	149.71	6.68
MobileNetV2 Sandler et al. (2018)	3.40	0.30	1.40	14.00	6.11	60.45	16.54
Swin Transformer Liu et al. (2021)	28	4.50	10.90	160.00	10.01	66.95	14.93
MedViT-T Manzari et al. (2023)	10	1.30	6.00	30.00	8.90	62.46	16.00
Proposed Method	0.96	0.958	2.90	9.20	2.10	14.06	71.13

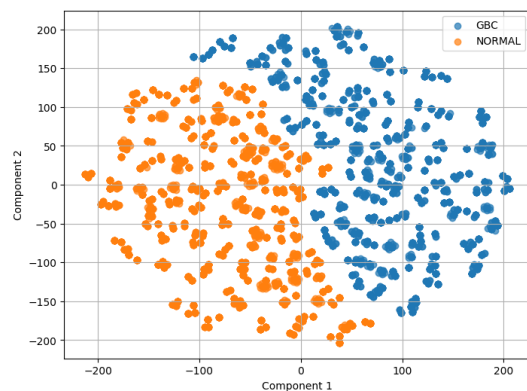


Fig. 5: t-SNE visualization of feature embeddings

Grad-CAM-Based Visual Interpretation of the Model

Figure 6 presents the Grad-CAM visualizations for both normal and GBC images. In the first row, subfigures (a) and (b) correspond to a normal image and its Grad-CAM activation, respectively. Subfigures (c) and (d) depict a GBC and its corresponding Grad-CAM. In normal cases, Grad-CAM visualizations exhibit diffuse attention over normal anatomical structures, while in GBC cases, there is a high degree of spatial concentration

within specific regions that demonstrate anomalous textures. The difference in attention allocation demonstrates that the model can effectively allocate attention to the relevant areas for GBC classification.

Statistical Significance Analysis

A non-parametric permutation test was employed to determine whether the observed classification performance of the proposed RIA-ViT framework was significantly greater than that expected under the null hypothesis of no association between the CT images and their corresponding class labels. The permutation test was selected because it does not require assumptions regarding the underlying distribution of classification performance and is therefore appropriate for evaluating machine-learning predictions when conventional parametric assumptions may not hold. Furthermore, the balanced binary-class design of the present study enables the construction of an empirical null distribution by repeatedly randomizing the class labels while maintaining the original class distribution.

Specifically, 10,000 random permutations were performed. For each permutation, the class labels were randomly shuffled, and the classification accuracy was recalculated to construct an empirical null distribution representing the performance expected under random label assignment.

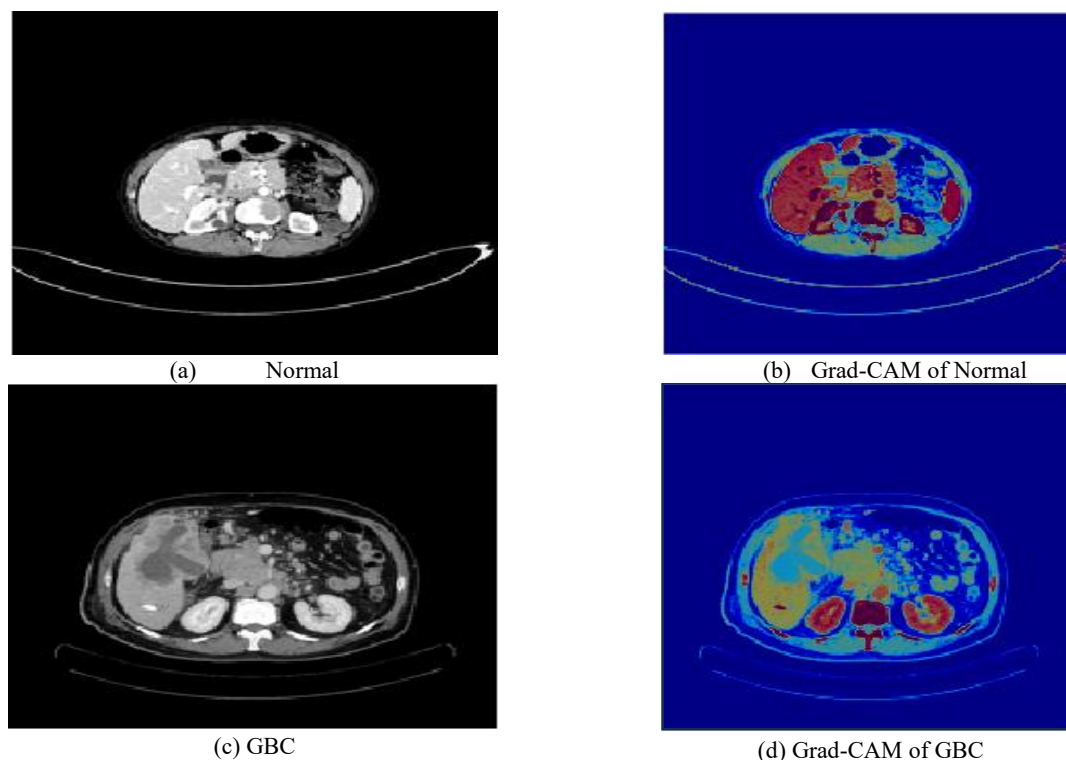


Fig. 6: Grad-CAM visualizations for normal and GBC Cases

The observed RIA-ViT classification accuracy of 99.56% was then compared with this null distribution. The resulting permutation-test p-value was 9.999×10^{-5} ($p < 0.001$), indicating that the observed classification performance was highly unlikely to have occurred under random label assignment. These findings provide statistical evidence that the predictive performance of the proposed framework is significantly greater than that expected by chance.

Clinical Workflow Integration

The designed RIA-ViT system is a computer-aided clinical decision support system, not an autonomous diagnostic system. After a contrast-enhanced CT scan is performed, individual slices of the gallbladder can be analyzed by an RIA-ViT model for predicting gallbladder carcinoma and for explaining the predictions by Grad-CAM. These outputs will help radiologists focus on the most important scans, improve diagnostic accuracy, and reduce time spent on examinations, as well as assist in decision-making, especially in hospitals lacking hepatobiliary imaging expertise. The proposed framework would integrate into a typical clinical workflow after CT image reconstruction. The framework would automatically analyze contrast-enhanced CT images and deliver malignancy probability scores with accompanying Grad-CAM explanations. The framework would produce these results with no additional work from the user. The radiologist would then evaluate the model's predictions, along with other components of the CT examination, the patient's case history and clinical data, lab data, among others. This would enable the radiologist to make a clinical diagnosis while the framework would provide support for the radiologist's clinical decision-making, ultimately enhancing the radiologist's diagnostic confidence.

Although this framework provides a system for making malignancy predictions, it should not autonomously make a clinical diagnosis or replace radiologists. The framework should be used in combination with other components of the radiologist's practice (e.g., radiology assessment, case history, physical exam, lab work, and pathology). Additionally, the preliminary, single-center study of the framework only considered normal and biopsy-confirmed cases of gallbladder cancer. As a result, the framework is not prepared for clinical use, and further multicenter validation is needed against diverse CT scanners, imaging protocols, and patient cohorts. Future studies will position the framework within PACS, conduct prospective reader-assist studies, and use human-in-the-loop frameworks to define the framework's potential clinical value.

Failure Case Analysis

The RIA-ViT framework had very good overall classification performance, but it misclassified some CT

slices. Failure cases were analyzed. The majority of incorrect classifications occurred in slices that had walls with subtle thickening, poorly defined lesions, or low conspicuity, as well as in slices that had characteristics that were similar to normal variations. In some cases, classification was made challenging due to partial volume effects and the response of the soft tissues. Some tissue types appeared very similar to normal and to pathological and malignant tissues. Some of these types of slices may contain very little information that can enable them to be classified. It is most likely that the remaining classification errors are due to variations in the imaging and the types of tissue that have very little contrast in them and that are nearly impossible to classify clearly. Future studies will try to enhance classification in these challenging cases by combining multiple CT slices and contextual clinical information, and by adding multimodal clinical data, larger datasets, and more instances of diverse pathologies.

Limitations and Future Work

Although the proposed RIA-ViT framework demonstrated strong classification performance, several limitations should be acknowledged. The study utilized a balanced dataset comprising 762 participants (381 normal and 381 biopsy-confirmed Gallbladder Cancer (GBC) cases) acquired using a standardized CT imaging protocol. A patient-wise data partitioning strategy was employed to prevent data leakage, while biopsy-confirmed diagnoses and expert radiologist-validated annotations provided reliable ground-truth labels, thereby supporting robust internal validation of the proposed framework. However, all CT images were collected from a single tertiary medical center using a single CT scanner, using a standardized contrast-enhanced CT imaging protocol. Consequently, institution-specific factors, including scanner characteristics, acquisition protocols, patient demographics, and disease prevalence, may introduce dataset-specific bias and limit the generalizability of the proposed framework. In addition, the use of a carefully curated and balanced dataset may have reduced inter-scan variability and partially contributed to the excellent classification performance observed in this study. Owing to ethical approval requirements, patient privacy considerations, and institutional data-sharing restrictions, no independent external or multi-center validation cohort was available during the present investigation. Consequently, the proposed RIA-ViT framework was evaluated only through a strictly patient-wise internal validation protocol, and the reported results should therefore be interpreted as internal validation rather than as evidence of universal clinical performance.

Furthermore, although patient-wise data partitioning was employed throughout the experimental protocol to

eliminate information leakage, the quantitative evaluation was performed at the slice level because the available annotations and prediction pipeline were organized at the slice level. Consequently, patient-level diagnostic performance metrics were not evaluated in the present study. Future work will aggregate slice-level predictions using clinically relevant fusion strategies, such as majority voting or probability averaging, to generate patient-level diagnoses and evaluate their clinical applicability in routine radiological practice. Independent validation using heterogeneous multi-center datasets acquired from different hospitals, CT scanner vendors, imaging protocols, and geographically diverse patient populations remains essential to comprehensively establish the robustness, reproducibility, and clinical generalizability of the proposed framework before routine clinical deployment.

This study was designed as a proof-of-concept framework for binary classification between normal gallbladder tissue and GBC, as sufficiently annotated benign gallbladder datasets were not available at the time of this study. While this framework enables a critical assessment of the developed methodology in the specified clinical context, it also does not tackle the more clinically relevant scenario of separating GBC from the various benign gallbladder pathologies such as xanthogranulomatous cholecystitis, adenomyomatosis, inflammatory gallbladder wall thickening, and gallbladder polyps, which typically present with overlapping CT appearance. Thus, the reported metrics should be considered within the context of binary classification and should not be extrapolated to other aspects of gallbladder disease. The primary goal of future research will be to validate the RIA-ViT framework using large, independent datasets from multiple centers, hospitals, vendors of CT scanners, different imaging strategies, and patient populations from different geographical regions. The framework will be extended to other benign diseases of the gallbladder and will be used for multi-class, clinically valuable differential diagnosis. In addition, other prospective clinical studies and independent external validations will be performed to assess the clinical applicability and generalizability of the framework.

Conclusion

This study proposes the first automated system to detect GBC in CT scan images. To achieve this, it combines a Vision Transformer with a Residual Inception Atrous Convolutional Neural Network (RIACNN). The proposed model uses the integrated feature extraction capabilities of the Transformer-based framework to capture both local and long-range features. The CT images of patients with cancer enabled the development of a model that captures and analyzes subtle changes in structure and texture, thereby improving the accuracy of cancer identification using inception, residual, and atrous

convolutions. With an accuracy of 99.56%, the model achieved the best results to date. The model scored almost 100% for precision, recall, and F1-score, and an AUC score of 1.0. Unlike other models, the new model achieved high accuracy with only 962,652 trainable parameters, making it compact and lightweight. The model also improves efficiency, using only 0.00413 GB of GPU memory and 3.67 MB of disk space to save the trained weights. The new model also delivers high performance at a low computational cost of only 2.40 BFLOPs. From the t-SNE analysis, the new model captured the required features to discriminate between normal and GBC cases. The Grad-CAM demonstrated that the new model captured key clinical areas in GBC cases, thereby enhancing model interpretability and providing a significant aid to clinical decision-making. The RIA-ViT framework is scalable, robust, and computationally efficient, thus forming a foundation for the construction of diagnostic computer-aided medical imaging systems. Future research in this area has several options. For instance, the architecture can be used to design a multi-class disease identification system with class-specific attention modules to identify various pathological states in the gastrointestinal and hepatobiliary systems.

Acknowledgment

The authors would like to thank MGM Medical College, Indore, Madhya Pradesh, India, for providing the contrast-enhanced CT image dataset used in this study. The authors also acknowledge the radiologists and clinical staff for their assistance in image annotation, validation, and clinical support throughout this research.

Funding Information

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Author's Contributions

Abhishek Sehrawat and Vishnupriya: Conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, and writing original draft.

Anita Gupta and Varun P. Gopi: Project administration, resources, supervision, writing, review, and edited.

Ethics

This study was conducted in accordance with the ethical standards of the institutional research committee and the 1964 Declaration of Helsinki and its later amendments. The experimental protocols were reviewed and approved by the Institutional Ethics Committee of

Chitkara University (Approval No. CCP/CU/IEC/25/134).

Consent to Participate

Informed consent was obtained from all participants and/or their legal guardians prior to data collection. Consent to publish. The authors affirm that human research participants provided informed consent for publication of the images in Figure 1.

Competing Interests

The authors declare no relevant financial or non-financial interests.

Availability of Data

The datasets generated and analyzed during the current study are not publicly available due to ethical restrictions and patient privacy considerations, but are available from the corresponding author upon reasonable request and subject to institutional approval.

References

- Basu, S., Gupta, M., Rana, P., Gupta, P., & Arora, C. (2022). Surpassing the Human Accuracy: Detecting Gallbladder Cancer from USG Images with Curriculum Learning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20854–20864. <https://doi.org/10.1109/cvpr52688.2022.02022>
- Basu, S., Gupta, M., Rana, P., Gupta, P., & Arora, C. (2023). RadFormer: Transformers with global-local attention for interpretable and accurate Gallbladder Cancer detection. *Medical Image Analysis*, 83, 102676. <https://doi.org/10.1016/j.media.2022.102676>
- Chang, Y., Wu, Q., Chi, L., & Huo, H. (2023). CT manifestations of gallbladder carcinoma based on neural network. *Neural Computing and Applications*, 35(3), 2039–2044. <https://doi.org/10.1007/s00521-022-06973-4>
- Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M. P., Zhang, S., Xing, L., Lu, L., Yuille, A., & Zhou, Y. (2021). TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021*, 402–412. <https://doi.org/10.48550/arXiv.2102.04306>
- Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1800–1807). <https://doi.org/10.1109/cvpr.2017.195>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv / NAACL-HLT*, 4171–4186. <https://doi.org/10.48550/arXiv.1810.04805>
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR. 9th International Conference on Learning Representations (ICLR 2021), Virtual Event / Vienna, Austria*. <https://doi.org/10.48550/arXiv.2010.11929>
- Dutta, U., Bush, N., Kalsi, D., Popli, P., & Kapoor, V. K. (2016). Epidemiology of gallbladder cancer in india. In *Indian Journal of Gastroenterology* (Vol. 35, Issue 1, pp. 1–7). <https://doi.org/10.21037/cco.2019.08.03>
- Fujita, H., Wakiya, T., Ishido, K., Kimura, N., Nagase, H., Kanda, T., Matsuzaka, M., Sasaki, Y., & Hakamada, K. (2022). Differential diagnoses of gallbladder tumors using CT-based deep learning. *Annals of Gastroenterological Surgery*, 6(6), 823–832. <https://doi.org/10.1002/ags3.12589>
- Gupta, A. (2025). Deep learning-based segmentation of gallbladder cancer lesions on contrast-enhanced ct: a multicenter evaluation of 2d and 3d models. In *European Journal of Radiology* (Vol. 173, Issue 111051, pp. 4615–4624).
- Halaseh, S. A., Halaseh, S., & Shakman, R. (2022). A review of the etiology and epidemiology of gallbladder cancer: what you need to know. In *Cureus* (Vol. 14, Issue 8, p. e28260). <https://doi.org/10.7759/cureus.28260>
- Hasan, M. Z., Rony, M. A. H., Chowda, S. S., Bhuiyan, M. R. I., & Moustafa, A. A. (2025). GBCHV an advanced deep learning anatomy aware model for accurate classification of gallbladder cancer utilizing ultrasound images. *Scientific Reports*, 15(1), 7120. <https://doi.org/10.1038/s41598-025-89232-5>
- He, W.-W., Zhu, J.-G., Pylypenko, D., Liu, F., Wang, M., Wu, Y.-F., Tian, J., & Li, H.-G. (2022). Differentiating benign from malignant gallbladder wall thickening in non-contrast MRI imaging: Preliminary study of a combined diagnostic indicator. *Medicine*, 101(40), e30861. <https://doi.org/10.1097/md.00000000000030861>
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. In *arXiv preprint* (p. 04861). <https://doi.org/10.48550/arXiv.1704.04861>
- Jaiswal, A., Gianchandani, N., Singh, D., Kumar, V., & Kaur, M. (2021). Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning. *Journal of Biomolecular Structure and Dynamics*, 39(15), 5682–5689. <https://doi.org/10.1080/07391102.2020.1788642>

- Jalali, S., Fateh, M., Rezvani, M., Abolghasemi, V., & Anisi, M. H. (2021). Resbcd-net: A novel deep learning architecture for biliary structure segmentation. In *Biomedical Signal Processing and Control* (Vol. 70, Issue 103003, p. 268).
<https://doi.org/10.3390/s21010268>
- Jeong, Y., Kim, J. H., Chae, H.-D., Park, S.-J., Bae, J. S., Joo, I., & Han, J. K. (2020). Deep learning-based decision support system for the diagnosis of neoplastic gallbladder polyps on ultrasonography: Preliminary results. *Scientific Reports*, 10(1), 7700.
<https://doi.org/10.1038/s41598-020-64205-y>
- Jin, H., Chen, C., Zhang, D., Yang, M., Wang, Q., Cai, Z., Si, S., Geng, Z., & Li, Q. (2025). Preoperative clinical radiomics and deep learning model based on ct for prognosis prediction in gallbladder cancer. In *BMC Cancer* (Vol. 25, Issue 1, p. 13711).
- Kaur, T., & Gandhi, T. K. (2019). Automated Brain Image Classification Based on VGG-16 and Transfer Learning. 2019 *International Conference on Information Technology (ICIT)*, 94–98.
<https://doi.org/10.1109/icit48102.2019.00023>
- Kumar, A., Sarangi, Y., Gupta, A., & Sharma, A. (2024). Gallbladder cancer: Progress in the Indian subcontinent. *World Journal of Clinical Oncology*, 15(6), 695–716.
<https://doi.org/10.5306/wjco.v15.i6.695>
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88.
<https://doi.org/10.1016/j.media.2017.07.005>
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021 *IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.
<https://doi.org/10.1109/iccv48922.2021.00986>
- Manzari, O. N., Ahmadabadi, H., Kashiani, H., Shokouhi, S. B., & Ayatollahi, A. (2023). MedViT: A robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine*, 157, 106791.
<https://doi.org/10.1016/j.compbiomed.2023.106791>
- McAllister, P., Zheng, H., Bond, R., & Moorhead, A. (2018). Combining deep residual neural network features with supervised machine learning algorithms to classify diverse food image datasets. *Computers in Biology and Medicine*, 95, 217–233.
<https://doi.org/10.1016/j.compbiomed.2018.02.008>
- Misra, S., Chaturvedi, A., Misra, N. C., & Sharma, I. D. (2003). Carcinoma of the gallbladder. *The Lancet Oncology*, 4(3), 167–176.
[https://doi.org/10.1016/s1470-2045\(03\)01021-0](https://doi.org/10.1016/s1470-2045(03)01021-0)
- Pang, S., Ding, T., Qiao, S., Meng, F., Wang, S., Li, P., & Wang, X. (2019). A novel YOLOv3-arch model for identifying cholelithiasis and classifying gallstones on CT images. *PLOS ONE*, 14(6), e0217647.
<https://doi.org/10.1371/journal.pone.0217647>
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. 2018 *IEEE/CVF Conference*, 4510–4520.
<https://doi.org/10.1109/cvpr.2018.00474>
- Simonyan, K., & Zisserman, A. (2014). Vgg19: A deep convolutional network for large-scale image recognition. In *International Conference on Learning Representations (ICLR) / arXiv* (p. 1556).
- Sugimoto, Y., Kurita, Y., Kuwahara, T., Satou, M., Meguro, K., Hosono, K., Kubota, K., Hara, K., & Nakajima, A. (2023). Diagnosing malignant distal bile duct obstruction using artificial intelligence based on clinical biomarkers. *Scientific Reports*, 13(1), 3262.
<https://doi.org/10.1038/s41598-023-28058-5>
- Valanarasu, J. M. J., Oza, P., Hacıhaliloglu, I., & Patel, V. M. (2021). Medical Transformer: Gated Axial-Attention for Medical Image Segmentation. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, 36–46.
https://doi.org/10.1007/978-3-030-87193-2_4
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 6000–6010). <https://doi.org/10.48550/arXiv.1706.03762>
- Wang, C., Chen, D., Hao, L., Liu, X., Zeng, Y., Chen, J., & Zhang, G. (2019). Pulmonary Image Classification Based on Inception-v3 Transfer Learning Model. *IEEE Access*, 7, 146533–146541.
<https://doi.org/10.1109/access.2019.2946000>
- Xiang, F., Meng, Q.-T., Deng, J.-J., Wang, J., Liang, X.-Y., Liu, X.-Y., & Yan, S. (2024). Deep learning-based differentiation of gallbladder cancer and benign gallbladder lesions on contrast-enhanced ct. In *HPB (Oxford)* (Vol. 26, Issue 2, pp. 157–166).
- Zhang, W., Wang, Q., Liang, K., Lin, H., Wu, D., Han, Y., Yu, H., Du, K., Zhang, H., Hong, J., Zhong, X., Zhou, L., Shi, Y., Wu, J., Pang, T., Yu, J., & Cao, L. (2023). Deep learning nomogram using contrast-enhanced ct for differentiation of xanthogranulomatous cholecystitis from gallbladder cancer. *Computers in Biology and Medicine*, 161, 106939.
- Zhao, X., Liu, B., Lin, Q., Huang, J., & Yu, L. (2024). Effective Combination of 3D-DenseNet’s Artificial Intelligence Technology and Gallbladder Cancer Diagnosis Model. *Frontiers in Computing and Intelligent Systems*, 6(3), 81–84.
<https://doi.org/10.54097/imkyfave>