scientific reports

Check for updates

OPEN

GBCHV an advanced deep learning anatomy aware model for accurate classification of gallbladder cancer utilizing ultrasound images

Md. Zahid Hasan[1]✉, Md. Awlad Hossen Rony[1], Sadia Sultana Chowa[1], Md. Rahad Islam Bhuiyan[1] & Ahmed A. Moustafa[2,3]

This study introduces a novel deep learning approach aimed at accurately classifying Gallbladder Cancer (GBC) into benign, malignant, and normal categories using ultrasound images from the challenging GBC USG (GBCU) dataset. The proposed methodology enhances image quality and specifies gallbladder wall boundaries by employing sophisticated image processing techniques like median filtering and contrast-limited adaptive histogram equalization. Unlike traditional convolutional neural networks, which struggle with complex spatial patterns, the proposed transformer-based model, GBC Horizontal-Vertical Transformer (GBCHV), incorporates a GBCHV-Trans block with self-attention mechanisms. In order to make the model anatomy-aware, the square-shaped input patches of the transformer are transformed into horizontal and vertical strips to obtain distinctive spatial relationships within gallbladder tissues. The novelty of this model lies in its anatomy-aware mechanism, which employs horizontal-vertical strip transformations to depict spatial relationships and complex anatomical features of the gallbladder more accurately. The proposed model achieved an overall diagnostic accuracy of 96.21% by performing an ablation study. A performance comparison between the proposed model and seven transfer learning models is further conducted, where the proposed model consistently outperformed the transfer learning models, showcasing its superior accuracy and robustness. Moreover, the decision-making process of the proposed model is further explained visually through the utilization of Gradient-weighted Class Activation Mapping (Grad-CAM). With the integration of advanced deep learning and image processing techniques, the GBCHV-Trans model offers a promising solution for precise and early-stage classification of GBC, surpassing conventional methods with superior accuracy and diagnostic efficacy.

Keywords Gallbladder cancer, Ultrasound images, Anatomy-aware model, Vision transformer, Horizontal and vertical strips

Gallbladder Cancer (GBC) presents a formidable clinical challenge globally, characterized by its high mortality rates and often late-stage diagnosis[1,2]. GBC is the most prevalent cancer of the biliary system and the fifth most common gastrointestinal cancer[3–5]. Global statistics from 2020 indicate that there was a diagnosis of approximately 115,949 cases of GBC, leading to 84,695 deaths[6]. With a five-year survival rate of approximately 13% for late stages, the prognosis of GBC is quite difficult. However, early diagnosis is significant for superior outcomes since it might enable radical resection, raising the survival percentage to 53%[3,4,7].

Recent advancements in imaging such as computed tomography (CT), magnetic resonance imaging (MRI), histopathology, among others, may provide superior visualization of gallbladder wall characteristics and adjacent tissues, offering crucial insights into the differentiation between benign and malignant lesions[8,9]. However, transabdominal ultrasound remains the primary imaging modality for assessing gallbladder diseases due to its cost-effectiveness, non-invasive and painless process[10]. Diagnosing GBC by manually observing ultrasound images is both time-consuming and labor-intensive for clinical experts. Additionally, the high volume of patients requiring daily evaluation increases the risk of misdiagnosis or overlooked abnormalities, leading to poor patient

[1]Health Informatics Research Laboratory (HIRL), Department of Computer Science and Engineering, Daffodil International University, Dhaka 1341, Bangladesh. [2]School of Psychology, Faculty of Society and Design, Bond University, Gold Coast (City), QLD, Australia. [3]Department of Human Anatomy and Physiology, The Faculty of Health Sciences, University of Johannesburg, Johannesburg, South Africa. ✉email: zahid.cse@diu.edu.bd

outcomes. To mitigate these issues, it is essential to implement automated deep learning techniques, which can significantly assist radiologists.

Artificial intelligence (AI), specifically deep learning (DL), have shown promise in revolutionizing medical imaging by enhancing diagnostic accuracy and efficiency[11–15]. Utilizing DLs to assist medical professionals with the diagnostic process is an emerging research subject. Researchers have extensively explored DL applications for classifying GBC through comprehensive analysis of ultrasound (US) image features[5,6,16,17]. However, these studies face several limitations, such as using small, low-resolution ultrasound images and lacking sufficient discussions on model interpretability, computational resources, and comparisons with other state-of-the-art models, which significantly impact the models' generalizability, reliability, and clinical trust.

The objective of this study is to present a novel DL-based approach for accurate classification of GBC into benign, malignant, and normal classes. For this classification we have utilized the GBC USG (GBCU) dataset. This dataset presents significant challenges such as low image quality, misaligned views, shadow interference, bias from spurious textures, and difficulty in detecting malignant cases. To address these challenges, our current methodology in this study focuses on enhancing ultrasonography image clarity and accurately defining the shape and boundaries of the GB wall. Advanced image enhancement techniques including median filtering to reduce speckle noise and Contrast Limited Adaptive Histogram Equalization (CLAHE) are employed. Traditional convolutional neural networks (CNNs) often struggle to capture these nuanced spatial patterns, effectively. In contrast, this study introduces a transformer-based architecture named GBC Horizontal-Vertical Transformer (GBCHV) and a transformer block named GBCHV-Trans block with a self-attention mechanism. The proposed anatomy-aware model for diagnosing GBC using ultrasound images focuses on leveraging the distinctive spatial relationships observed between malignant and benign gallbladder tissues. This approach enhances the model's capacity to discern and utilize spatial correlations among visual elements within the gallbladder wall, which are crucial for accurate GBC classification. By transforming square-shaped visual elements into horizontal and vertical strips and integrating anatomical prior knowledge, this model, for the first time, not only improves diagnostic accuracy but also provides insights into the spatial dynamics' indicative of GBC. The assessment of the proposed model's potential for overfitting includes the analysis of accuracy and loss curves, as well as the ROC curve. Several performance metrics and statistical analysis are utilized to evaluate the model outcome. Furthermore, a comparative analysis is carried out between the performance of the proposed model and seven transfer learning models. The major contributions of this study are outlined below:

- An image processing step involving a median filter and CLAHE is performed to mitigate the influence of low-quality or redundant US images.
- A method to transform square-shaped visual elements into horizontal and vertical strips, enhancing the model's ability to accurately distinguish between benign, malignant, and normal gallbladder images is introduced.
- The integration of GBCHV-Trans blocks with a convolutional stem for early visual processing enhances the overall performance of the proposed model through early visual processing.
- The model's classification efficiency is explained through Gradient-weighted Class Activation Mapping (Grad-CAM).

This approach surpasses conventional CNN approaches, showing promise for enhancing diagnostic accuracy in clinical ultrasound settings. It represents a significant advancement in medical imaging and AI applications in oncology, potentially improving patient outcomes by enabling earlier diagnosis and intervention.

Literature review

The diagnosis of GBC is still challenging because of the disease's complicated anatomical presentation and early symptomatic traits. Many object detection and deep learning techniques have been proposed by researchers to identify and identify different types of cancer, including brain, liver, breast, and lung cancers. This section reviews a number of studies that particularly used various methodologies for the diagnosis and classification of anomalies related to the gallbladder. These techniques improve the diagnosis precision of GBC by utilizing the distinct anatomical characteristics seen in ultrasound images.

Machine learning and deep learning approaches

Medical image analysis has gone through advancement through the use of machine learning and, more recently, deep learning approaches[18]. Gallstones, xanthogranulomatous cholecystitis, calcified or porcelain gallbladder, cholelithiasis with typhoid carriers, gallbladder adenoma, red meat consumption, and tobacco use are associated with an increased risk of GBC[19]. Soumen Basu et al.[20] used GBCNet to tackle the challenges in the identification of GBC. By initially identifying the GB (rather than the cancer), GBCNet extracts the regions of interest (ROIs). It then applies a novel multi-scale, second-order pooling architecture that is specifically designed to classify GBC. It was found that GBCNet can achieve 91% accuracy, 95% specificity, and 92.1% precision. Interpretable representation is one of the new and robust methods in medical image analysis. In another study Soumen Basu et al.[5] proposed a novel framework for deep neural networks to develop an interpretable representation for the analysis of medical images. This architecture first creates a global attention mechanism for the area of interest, and then it uses local attention to learn deep feature embeddings in the bag of words style. A modern transformer architecture is used to merge the global and local feature maps to detect GBC from ultrasonography (USG) images with excellent accuracy. Their proposed method RadFormer achieved an effective accuracy of 92.1% and superior compared to human radiologists. There are several proposed architectures to correctly classify the GBC but before classifying ROI segmentation is the basic term of any medical image analysis work. With this motive, Tao Chen et al.[21] developed a principal components analysis (PCA) and AdaBoost technique for segmenting ultrasound images, which can be used to build a computer-aided diagnosis system that can distinguish between

cancerous and non-cancerous gallbladder polyps. The gallbladder region was accurately delineated with a high accuracy of 95% using the proposed segmentation method. In contrast to 69.05%, 67.86%, and 70.17% with convolutional neural networks, the accuracy, sensitivity, and specificity of the proposed computer-aided diagnosis method based on the segmented images are 87.54%, 86.52%, and 89.40%. With a substantially faster diagnostic speed (0.02 s vs. 3 s), the diagnosis result is also significantly greater than the results of sonologists using their human eyes (86.22%, 85.19%, and 89.18% on average across four sonologists).

Anatomy-aware models in medical imaging

In a number of medical imaging applications, the integration of anatomical knowledge into machine learning algorithms has produced promising outcomes. The CheXRelNet by Gaurang Karwande et al.[22] is a neural model that is capable of tracking the changes in two chest X-rays (CXRs) over time due to disease utilizing the chest ImaGenome dataset. The dataset includes nine pathological condition of lungs including lung opacity, pleural effusion, atelectasis, enlarged cardiac silhouette, pulmonary edema/hazy opacity, pneumothorax, consolidation, fluid overload/heart failure, and pneumonia. In order to precisely forecast disease change for a pair of CXRs, CheXRelNet integrates both local and global visual features, makes use of anatomical information from both within and between images and discovers relationships between anatomical region properties. The Chest ImaGenome dataset experimental results demonstrate improved downstream performance over baseline performance. In order to produce a large variety of shape-dependent masks, Yousef Yeganeh et al.[23] proposed an unsupervised guided masking technique based on a commercially available illustrating model and a super pixel over-segmentation algorithm. The results of their experiments on the reconstruction of abdominal MR images demonstrate the superiority of their proposed masking strategy over conventional techniques that use datasets of irregularly shaped or square-shaped masks. The results show that in segments PSNR, SSIM, LPIPS of segmented images are 19.38, 0.77, and 0.255 respectively at the same image. Zeyu Fu et al.[24] developed a method using anatomy-aware contrastive learning (AWCL), which adds anatomy knowledge to enhance positive/negative pair sampling in a contrastive learning fashion, to enhance visual representations of medical images. In another study, the Anatomy-XNet by Uday Kamal et al.[25] proposed an attention-based thoracic illness classification network that is aware of anatomy and prioritizes spatial features based on previously determined anatomy regions. The pre-trained DenseNet-121 serves as the foundation network for the proposed Anatomy-XNet, which has two related structured modules. Experimental results demonstrate that the method they propose establishes a new state-of-the-art benchmark with AUC scores of 85.78%, 92.07%, and 84.04% on three large-scale CXR datasets that are available to the public: MIMIC-CXR, Stanford CheXpert, and NIH.

Transformer approaches in computer vision

This study employed a transformer-based model for GBC classification. Transformer models in computer vision have significantly impacted many fields, including healthcare, by improving the accuracy and efficiency of image analysis. The use of such model in this study aims to enhance the reliability and effectiveness of GBC classification through advanced image processing techniques. Onat Dalmaz et al.[26] present ResViT, a unique generative adversarial method for medical image generation. This approach leverages the contextual sensitivity of vision transformers, the precision of convolutional operators, and the realism provided by adversarial learning. The results show that ResViT outperforms alternative CNN- and transformer-based techniques in terms of both quantitative measures and qualitative observations. Junyu Chen et al.[27] presented ViT-V-Net, a bridge that enables volumetric medical image registration by connecting ViT and ConvNet. The experimental findings shown here show that the suggested design outperforms a number of highly effective registration techniques and results show that ViT-V-Net achieves 0.726 ± 0.130 dice score. Qianying Liu et al.[28] demonstrates how to create a precise and small Transformer network for MISS called CS-Unet. This network uses convolutions to improve the spatial and local modeling capabilities of Transformers in a hierarchical fashion via multiple stages of construction. They cleverly constructed Convolutional Swin Transformer (CST) block, which combines convolutions with Multi-Head Self-Attention and Feed-Forward Networks to provide intrinsic localized spatial context and inductive biases, is primarily responsible for this.

While existing methods have shown promising results in terms of accuracy, there are still gaps in generalizability, anatomical integration, and interpretability. A major gap is that, although some recent studies have explored anatomy-aware models, they often lack the ability to visualize the model's interpretability in detecting abnormal areas. Additionally, in most of the studies, the validation of their proposed models is not demonstrated, which fails to explain the concerns about overfitting or underfitting issues. By addressing these challenges and combining the strengths of modern architectures, such as transformers and anatomy-aware models, we focus on providing more reliable and clinically applicable solutions for GBC detection.

Methodology and implementation

Dataset description

The GBC USG (GBCU) dataset was used in this analysis. It was obtained from PGIMER, a tertiary care referral hospital in Northern India, and was initially produced by Basu et al.[20]. There are three categories in the GBCU dataset: normal, malignant, and benign. A total of 1,255 images are included in it, of which 432 are normal, 558 benign, and 265 malignant. These images have different sizes; their widths range from 801 to 1,556 pixels, and their heights range from 564 to 947 pixels. Bounding boxes have been used to annotate the dataset by expert radiologists. Three subsets of the GBCU dataset have been identified: testing (125 images), validation (126 images), and training (1,005 images). The ROIs in each image, which include the entire gallbladder and the liver parenchyma neighborhood, have been annotated by experienced radiologists, and the ground-truth labels are biopsy-confirmed. Samples from the three classes in the GBCU dataset are shown in Fig. 1.

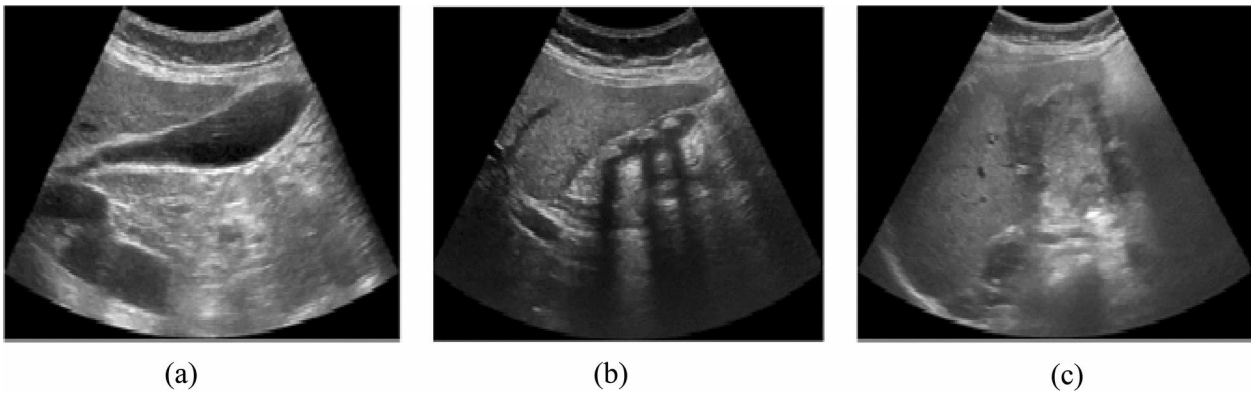

Fig. 1. Illustration of the dataset distinguished between three classes (**a**) Normal, (**b**) Benign, (**c**) Malignant.

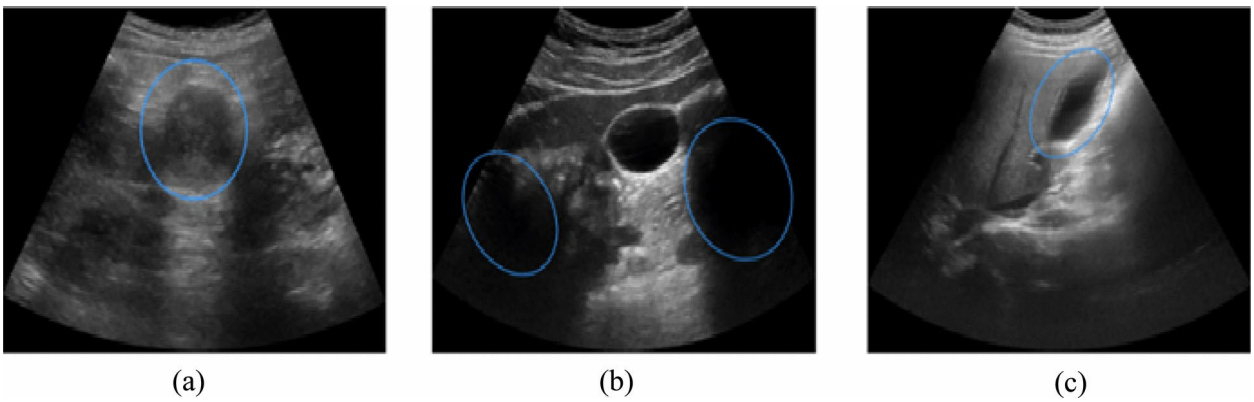

Fig. 2. Navigating few complexities with the dataset (**a**) Malignant GB boundary is not visible, (**b**) Shadow in USG image, (**c**) GB view is not aligned.

Challenges related to analyzing the GBCU dataset
Working with the GBCU dataset presents several challenges. Therefore, the key challenges in GB cancer detection are as follows:

1) Low Image Quality: Noise and other sensor artifacts frequently cause low-quality ultrasound (USG) images, which makes it challenging to differentiate gallbladder areas with accuracy.
2) Misaligned Views: Consistent image analysis is complicated by the images' frequent mismatch caused by the portable nature of the ultrasonography sensor.
3) Shadow Interference: Shadows often have a similar appearance to gallbladder in USG images, leading to misclassification and affecting detection accuracy.
4) Bias from Spurious Textures: Training object detectors for GBC detection can result in poor accuracy due to learning from noise-induced distorted textures and adjacent organ tissues, rather than focusing on the appearance or boundaries of the GB wall.
5) Difficulty in Malignant Case Detection: Malignant masses are difficult to identify because they frequently lack recognizable gallbladder boundaries or forms. This makes diagnosing malignant gallbladder instances difficult.

Fig. 2 presents a few challenges with the dataset's images.

Proposed methodology

In ultrasound images, anatomical structures indicative of GBC displays distinct spatial relationships between malignant and benign gallbladders. Leveraging this prior knowledge, we propose an anatomy-aware model for fully automatic GBC diagnosis.

The methodology, illustrated in Fig. 3, begins with preprocessing steps detailed in Section "Ablation study", preparing the dataset for the proposed model's application. Section 4.2.1 introduces the core concept of anatomy-aware formulation, while Sect. 4.2.2 outlines the GBCHV-Trans stage derived from this concept. Sections 4.2.3 defines the overall network architecture of our model, and Section "Discussion and limitation" provides implementation specifics.

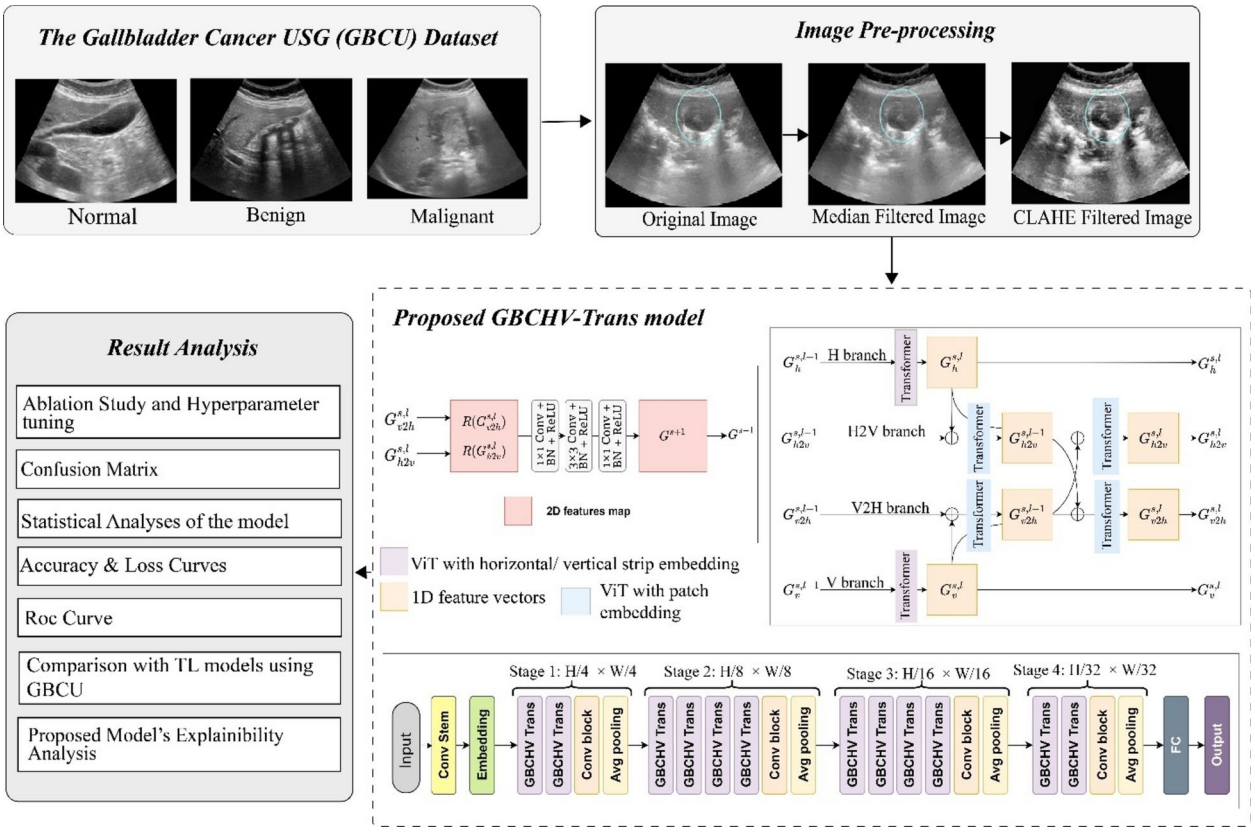

Fig. 3. Proposed methodology of this study.

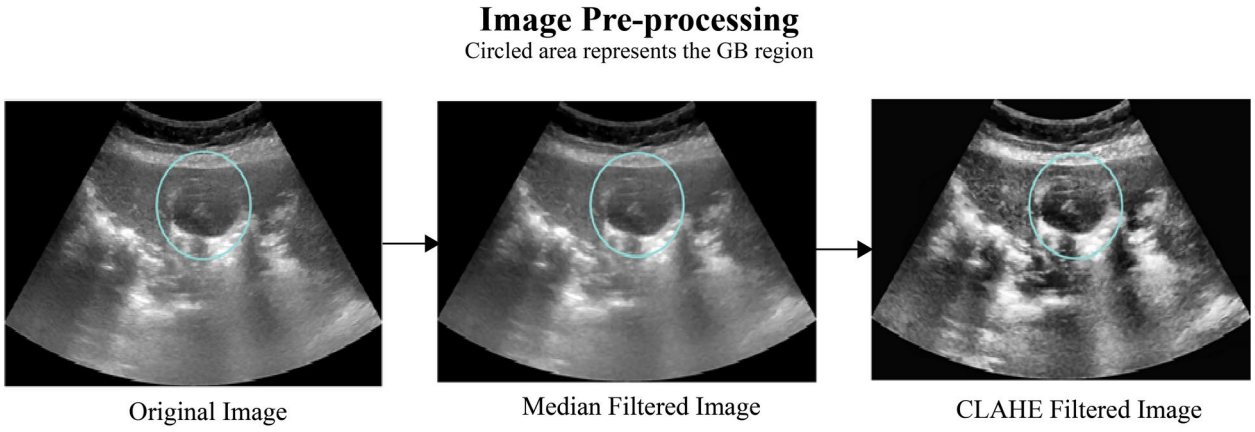

Fig. 4. Visualization of image preprocessing steps which includes median filtered and CLAHE filtered image.

Image pre-processing

Image preprocessing is a crucial step in medical imaging systems where denoising is especially essential. We have preprocessed the images in order to enhance the quality of images, making them more suitable for analysis by removing irrelevant information, standardizing pixel values, and improving overall data consistency[29]. The effective image preprocessing, employed in this study ensures the prompt learning of the proposed model, resulting in lower training times, higher accuracy, and improved generalization. This section presents the pre-processing steps and their outcomes are shown in Fig. 4.

Median filter

Median filtering is a nonlinear approach for smoothing images replacing each pixel with the median value of its neighbors[30]. This approach is very effective at reducing impulsive noise, such as salt-and-pepper noise while

maintaining edges[31]. The degree of smoothing is determined by the kernel size, more specifically, smaller kernels contain more detail, whereas bigger kernels give off a smoother image.

In the median filtering process, the central pixel in a P × P window is replaced with the median value of that window's pixels. This process is defined as follows[32]:

$$f(x,y) = median\{g(m,n)\}, where\,(m,n) \in S_{xy} \quad (1)$$

Here, the median filter defines the median gray value for each pixel within a rectangular sub-image window centered at (x, y) pixel. For a given pixel located at coordinates (x, y), the filter examines a rectangular window centered around this pixel. The size of this window is defined by a kernel, typically denoted as P × P. Within this window, the median value of all the pixels is computed. The pixel at the center of this window (m, n) is then replaced by this median value. The median filter minimizes impulsive noise, without blurring the edges, thereby enhancing the overall image quality.

CLAHE

CLAHE is an advanced method for enhancing the contrast and limiting noise amplification in digital images. It is the extended version of traditional histogram equalization such as Adaptive histogram equalization (AHE) and standard histogram equalization (HE).

CLAHE is highly effective for ultrasound images, which often contain speckle noise and low-contrast regions[33]. Farhan et al.[34] demonstrated the significance of CLAHE in enhancing USG image quality and model performance. The CLAHE procedure involves splitting an image into smaller portions, calculating histograms for each depending on gray levels, and generating contrast-limited histograms to regulate noise[35]. This approach leads to more effective image details and has been successfully employed for a variety of medical image analyses, including CT, MRI, and retinal images[33,36,37].

Model

This section details the architectural framework of the proposed GBCHV classification model, focusing on its design and components utilized for effective diagnostic accuracy.

Anatomy-aware formulation

Based on the principles of ultrasound imaging and the anatomical structure of the gallbladder, different tissues within the gallbladder produce distinct wall formations in ultrasound images. While conventional CNN models are excellent at extracting representative local features, they struggle to capture spatial relationships. To address this limitation, most current algorithms for diagnosing GBC in ultrasound images use a predefined ROI. This approach helps eliminate redundant areas and allows the CNN to focus on classifying the ROI more accurately. In contrast, the self-attention mechanism in transformers enhances the representation of spatial relationships among visual elements, as illustrated in Fig. 5a. To further leverage the intra-layer and inter-layer spatial correlations in gallbladder ultrasound imaging, we propose transforming the square-shaped visual elements into horizontal and vertical strips. This approach integrates anatomical prior knowledge into the model, as depicted in Fig. 5b.

GBCHV-trans stage

1) Vision Transformer

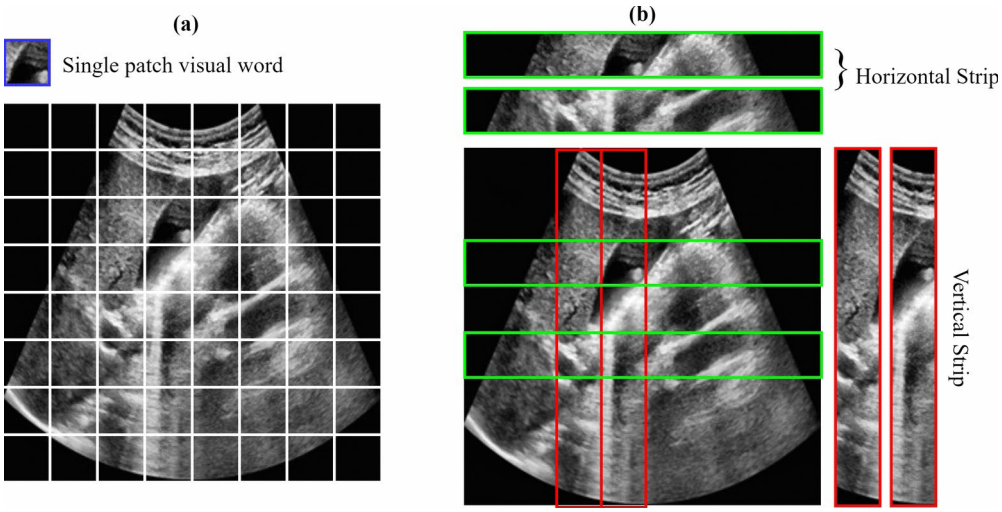

Fig. 5. Visualizes the integration of anatomical prior knowledge into the model (**a**) The visualization of single patch of a GB image; (**b**) The visualization of anatomical knowledge through horizontal and vertical strip.

The Vision Transformer (ViT) is the first model to apply a widely used technique from natural language processing to computer vision[38]. It segments the input image $X \in \mathbb{R}^{H\times W\times C}$ into patches $X_p \in \mathbb{R}^{N\times\left(p^2.c\right)}$

, treating these patches as visual words or tokens. Here, (H, W, C) denote the height, width, and channels of the original image, and (P, P, C) denote the dimensions and channels of each patch. The number of patches is denoted by N. Each visual word is then converted from a 2D image patch into a 1D vector, known as a patch embedding. The multi-head self-attention mechanism is used to create spatial correlations between these different tokens. This method is mathematically represented in the ViT model:

$$G_0 = \left[X_{class}; X_p^1 E, X_p^2 E, \ldots\ldots, X_p^N E,\right] + E_{pos}, E \in \mathbb{R}^{\left(p^2.c\right)\times D}, E_{pos} \in \mathbb{R}^{(N+1)\times D} \tag{2}$$

$$G_l\prime = MSA\left(LN\left(Z_{l-1}\right)\right) + Z_{l-1} \tag{3}$$

$$G_l = MLP\left(LN\left(Z_l\prime\right)\right) + Z_l\prime \tag{4}$$

$$y = LN\left(Z_L^0\right) \tag{5}$$

where E, E_{pos}, MSA, MLP and LN denote trainable linear projection, position embedding, multi-head self-attention module, multi-layer perceptron module and layer norm, respectively.

In our proposed GBCHV-Trans, we use several ViT blocks with the same structure without class embedding to construct the GBCHV-Trans block. Thus, we denote all the ViT blocks in the following as $Trans(\cdot)$.

2) Embedding

To incorporate anatomical prior knowledge into the transformer model, we introduce two additional embedding methods as illustrated in Fig. 6. Starting with the input image $I \in \mathbb{R}^{H\times W\times 3}$, we first use a convolutional stem to downsample the image by a factor of four, resulting in $I\prime \in \mathbb{R}^{H/4\times W/4\times 3}$ and introducing early inductive bias. Subsequently, we process patch embedding, horizontal strip embedding, and vertical strip embedding before inputting them into the model. Patch embedding divides $I\prime$ into $N \times N$ patches $X_L^{(r,c)}$, where r and c represent the row and column indices. After flattening these patches, we obtain a set of 1D vectors G_p.

$$G_p = \left\{X_p^{(r,c)} | X \in \mathbb{R}^{\frac{H}{4N}\times\frac{W}{4N}\times C}\right\}, r, c = 1, \ldots\ldots, N \tag{6}$$

Horizontal strip embedding is introduced to represent the visual words of the same anatomical layer with M strip, defined as:

$$G_h = \left\{X_h^{(r)} | X \in \mathbb{R}^{\frac{H}{4M}\times\frac{W}{4}\times C}\right\}, r = 1, \ldots\ldots, M \tag{7}$$

Vertical strip embedding is introduced to represent the visual words across anatomical layers with M strips, defined as:

$$G_v = \left\{X_v^{(c)} | X \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4M}\times C}\right\}, c = 1, \ldots\ldots, M \tag{8}$$

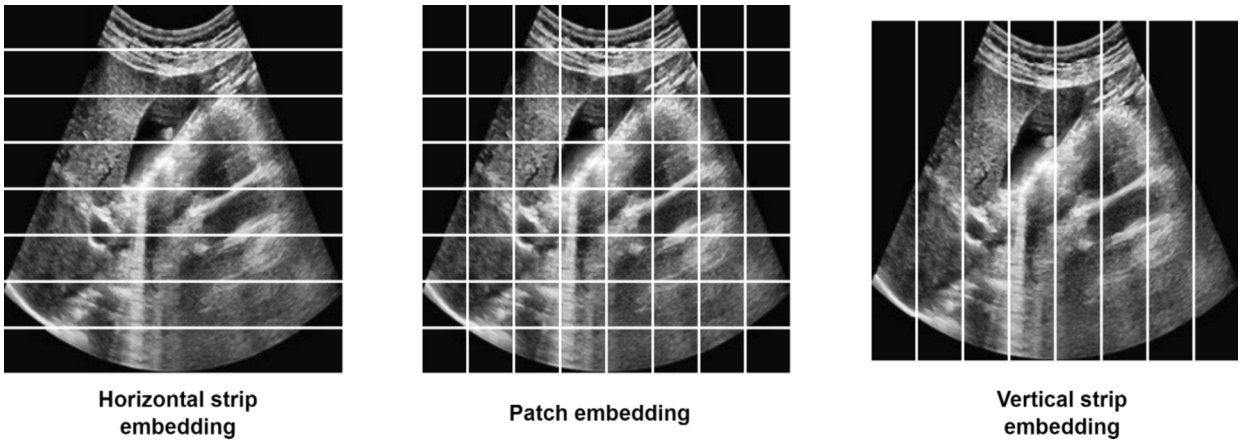

Fig. 6. Incorporating anatomical prior knowledge into the transformer model visualizing embedding methods including horizontal, patch, and vertical strips.

3) GBCHV-Trans Block

The architecture of the GBCHV-Trans block is depicted in Fig. 7. We have designed a symmetry structure with four branches in one GBCHV-Trans block, H branch (horizontal), V branch (vertical), H2V branch (horizontal to vertical) and V2H branch (vertical to horizontal).

Let us define the features at the -th block in the s-th stage as $G^{s,l}_{\{h,v,h2v,v2h\}}$. GBCHV-Trans block takes the outputs from the previous block and generates the features for the next block, defined as:

$$\left\{G_h^{s,l}, G_v^{s,l}, G_{h2v}^{s,l}, G_{v2h}^{s,l}\right\} = f\left(G_h^{s,l-1}, G_v^{s,l-1}, G_{h2v}^{s,l-1}, G_{v2h}^{s,l-1}\right) \tag{9}$$

where f (·) denotes the GBCHV-Trans block. The inputs of four branches in the first GBCHV-Trans block (when $l = 1$) are equivalent to the features from the previous GBCHV-Trans stage G^{s-1}.

$$G^{s,l}_{\{h,v,h2v,v2h\}} = G^{s-1} \tag{10}$$

H and V branches are two auxiliary branches to extract the inter-layer and intra-layer spatial correlations, with two identical H and V branches with horizontal strip embedding and vertical strip embedding respectively.

$$G_h^{s,l} = Trans(G_h^{s,l-1}) \tag{11}$$

$$G_v^{s,l} = Trans(G_v^{s,l-1}) \tag{12}$$

The anatomy-aware spatial features $G_h^{s,l}$ and $G_v^{s,l}$ are then passed into the next two main branches (H2V and V2H). They are also regarded as the inputs of the next GBCHV-Trans block.

H2V and V2H branches are served as the main feature extraction branches which fuse the features from two auxiliary branches (H and V). For example, in the H2V branch, the horizontal features $G_h^{s,l}$ are added to the features from the previous GBCHV-Trans block $G_{h2v}^{s,l-1}$. After a transformer encoder, the vertical features $G_v^{s,l}$ are added behind. The V2H branch is the mirror of the H2V branch.

$$G_{h2v}^{s,l} = Trans(Trans\left(G_h^{s,l} + G_{h2v}^{s,l-1}\right) + G_v^{s,l}) \tag{13}$$

$$G_{v2h}^{s,l} = Trans(Trans\left(G_v^{s,l} + G_{v2h}^{s,l-1}\right) + G_h^{s,l}) \tag{14}$$

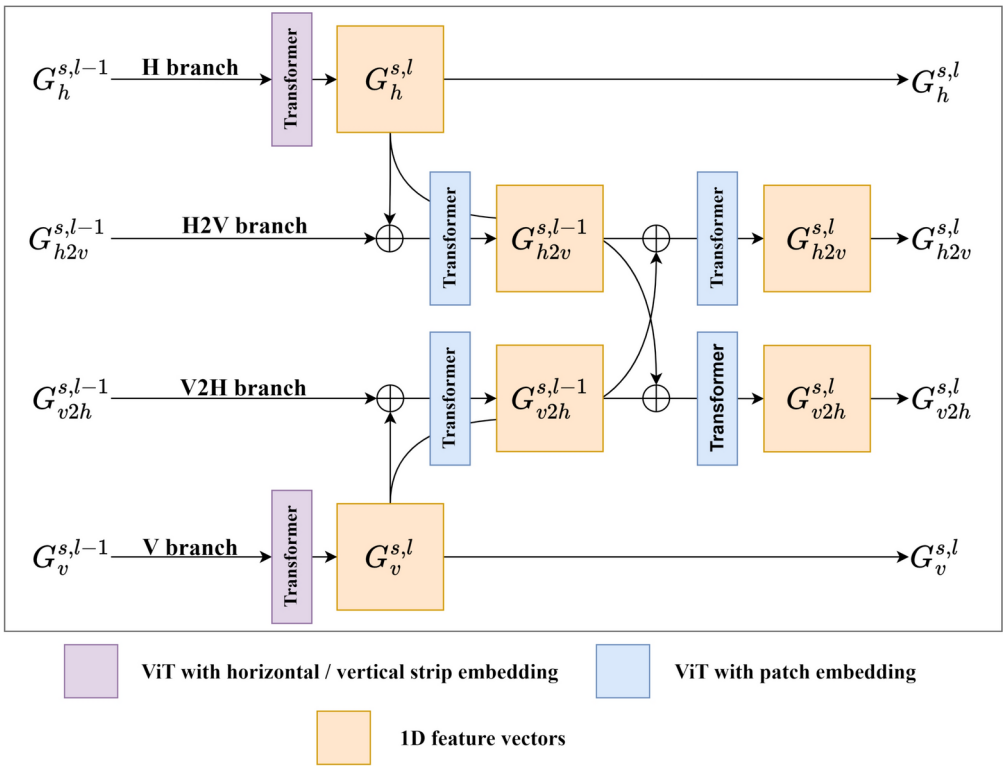

Fig. 7. The architecture of the GBCHV-Trans block including horizontal / vertical strip embedding and patch embedding.

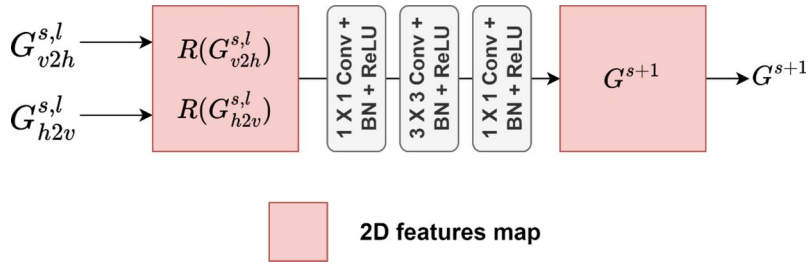

Fig. 8. The architecture of the convolutional block of the proposed GBCHV model.

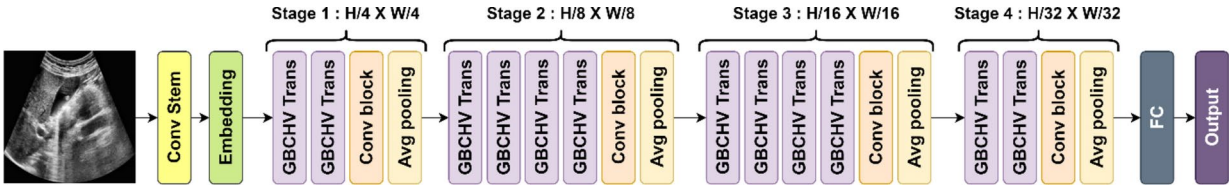

Fig. 9. The architecture of the proposed GBCHV model.

The output features $G^{s,l}_{h2v}$ and $G^{s,l}_{v2h}$ will be passed into the next block. Note that, for the last GBCHV-Trans Block in each stage, the features will be passed into a Conv Block, described as follows.

4) Convolutional Block

The transformer performs well in identifying spatial correlations and processing sequential data. However, inductive bias is absent. We have added a convolutional block after the last GBCHV-Trans block at each stage to fuse the H2V and V2H features and the inductive bias in order to take use of the transformer's and CNN's strengths.

$$G^{s+1} = conv(G^{s,l}_{h2v}, G^{s,l}_{v2h}) \tag{15}$$

As shown in Fig. 8, the Conv block takes the output 1D feature vectors $G^{s,l}_{h2v}$ and $G^{s,l}_{v2h}$ from two main transformer branches as the inputs. These two feature vectors are reshaped to 2D feature maps and concatenated together. After three convolutional layers, the Conv block outputs the 2D feature maps G^{s+1} for the Stage. There are three convolutional layers in conv block. The channel numbers of each convolutional layer in four stages are[8,8,16,16,16,32], [32, 64, 32] and [64, 128, 64] respectively.

Proposed model

Our proposed model (see Fig. 9) comprises four stage modules, each containing multiple GBCHV-Trans blocks, a Conv block, and a pooling layer. Initially, we use a convolutional stem for early visual processing. Unlike the patchy stem in the original ViT, this early convolutional stem introduces an inductive bias early on, enhancing optimization stability and overall model performance. After the convolutional stem, the input image I is reduced to dimensions $H/4 \times W/4 \times C$, where C is 4. The feature map sizes for the subsequent three stages are $H/8 \times W/8 \times 2C$, $H/16 \times W/16 \times 4C$,and $H/32 \times W/32 \times 8C$. A Conv block integrates horizontal and vertical information and links adjacent stages, ensuring that the input for each stage remains a 2D image or 2D feature map. Embedding or flattening is applied to adapt the input for the transformer. In the final stage, a fully connected layer is used for inference, with cross-entropy loss employed to optimize the model. In the proposed model, we employed the Categorical Cross-Entropy Loss (CCE) to diminish this disparity by modifying the model's weights via backpropagation. In the context of medical imaging tasks, reducing the cross-entropy loss enhances the model's accuracy and dependability in identifying anomalies, classifying disease classes, or segmenting areas of interest. The loss function may be represented as Eq. (16).

$$\mathcal{L}_{CCE} = -\sum_{i=1}^{C} y_i \log(y\prime_i) \tag{16}$$

If the model outputs probabilities for C classes, and the true label is represented as a one-hot vector y, the categorical cross-entropy loss can be calculated with Eq. (16). Here, y_i will be when the true class is i, otherwise, it will be 0, and $y\prime_i$ is the predicted probability for class i. Our GBCHV model is trained with a total of 13,24,678 parameters. The proposed model architecture is shown in Fig. 9.

Transfer learning models

To benchmark the performance of our proposed model, we evaluated it against several state-of-the-art transfer learning architectures known for their efficacy in various computer vision tasks. Here, a brief description is provided of these models to understand their architectural definitions and significance. MobileNetv3 leverages depth-wise separable convolutions and linear bottlenecks to maximize efficiency for embedded and mobile devices[39]. It presents effectively inverted residuals with linear bottlenecks and mobile-friendly activation functions like Swish to lower computing costs without sacrificing excellent performance in applications like object detection and image classification. The convolutional neural network VGG16 utilizes max-pooling layers and compact (3×3) convolution filters to reduce spatial dimensions gradually. It has 13 convolutional layers and 3 fully connected layers. VGG16 is a benchmark used to assess deeper architectures and is renowned for its robust performance in image recognition applications[40]. ResNet50, an extension of the ResNet (Residual Network) family, uses residual connections to overcome the difficulty of training extremely deep neural networks. Because of these connections, gradient flow through the network is maintained even in situations when stacked layers would reduce it, allowing for the training of models with up to 152 layers that are noticeably deeper. Specifically, ResNet50 has 50 layers total, including residual blocks connected by shortcuts[41]. InceptionV3, also referred to as GoogLeNet v3, comprises inception modules that integrate various filter sizes (1×1, 3×3, and 5×5 convolutions) into a single layer. This architecture enhances computational efficiency and facilitates feature extraction at various scales. To enhance gradient flow during training, InceptionV3 additionally incorporates auxiliary classifiers. EfficientNet-B7 balances model depth, width, and resolution to attain cutting-edge accuracy. The largest variant, EfficientNet-B7, combines computational economy and model accuracy well enough to perform a variety of tasks, including object identification and image categorization. Compound scaling is used to optimize the dimensions of the network for better performance. RetinaNet is a one-stage object detection model that combines a focus loss function and a feature pyramid network (FPN). The FPN enables multiscale feature extraction, which improves the model's capacity to identify objects of various sizes in images. To increase detection performance, the focal loss function solves class imbalance by down-weighting simple examples during training and concentrating more on challenging cases. DenseNet-264 provides feed-forward connections between each layer and all other layers. This structure of dense connection minimizes the number of parameters, promotes feature reuse across the network, and improves gradient flow during training. DenseNet-264 specifically consists of 264 layers organized into dense blocks, where each layer receives feature maps from all preceding layers as input.

Results

The results of our proposed GBCHV-Trans model for fully automated ultrasound image-based GBC diagnosis are shown in this section. The findings are divided into four subsections: Adulation Study, Performance analysis of the proposed GBCHV-Trans model, Comparison with Other Models and Explainability of the model. To ensure a thorough evaluation, each subsection provides an in-depth analysis of the model's performance from different perspectives. The results show that the GBCHV-Trans model is both robust and successful, delivering a significant advance over previous methods. The details of these findings will be covered in the section, emphasizing the improvements achieved by our proposed methodology.

Ablation study

To improve the performance of the GBCHV-Trans model an in-depth ablation study is carried out. Different configuration with anatomy-aware formulation models, with or without convolution stem and block, lastly the sizes of different embedding ways are analyzed. The metrics gathered from different configurations of our model are compiled in the Table 1.

The anatomy-aware formulation's ablation study shows how well the GBCHV-Trans model performs in different configurations. As anatomical knowledge is gradually added, Models A, B, and C demonstrate considerable gains in AUC, accuracy, precision, recall, specificity, and F1 score. In Model A, we eliminate both the horizontal (H) and vertical (V) branches with their respective strip embeddings, retaining only the two main branches with patch embeddings. For Model B, we remove the H branch, preserving the other three branches. Conversely, in Model C, we remove the V branch, keeping the remaining three branches intact. With an accuracy of 94.28% and an AUC of 98.77%, Model A outperforms Model B and Model C in terms of performance metrics. The full GBCHV model outperforms all other configurations with an AUC of 99.68%, accuracy of 96.21%, precision of 95.54%, recall of 96.07%, specificity of 95.89%, and an F1 score of 94.91%. This model includes significant anatomy-aware features. This emphasizes how important anatomy-aware formulation is to improving the diagnostic potential of the model.

The impact of adding convolutional layers (Conv Stem and Conv Block) to the GBCHV-Trans model is examined in the ablation study on convolution. With no convolutional layers (×, ×), the baseline model has an accuracy of 90.17% and an AUC of 92.35%. Performance is much enhanced with the addition of convolutional blocks (×, ✓), with an accuracy of 92.78% and an AUC of 95.01%. The model is also improved by adding simply the Conv Stem (✓, ×), which results in an accuracy of 93.09% and an AUC of 95.73%. The best performance is shown by the entire model, which includes both Conv Stem and Conv Block (✓, ✓). Its AUC is 99.68%, accuracy, precision, recall, and specificity are 96.21%, 95.54%, 96.07%, and 94.91%, respectively. This study emphasizes the significant advantages of using convolutional layers.

The impact of altering the patch sizes (p) and the quantity of horizontal and vertical embeddings (h&v) on the GBCHV-Trans model's performance is investigated in the ablation study on the sizes of various embedding methods. The findings demonstrate that better performance is typically achieved with smaller patch sizes and embedding dimensions. For example, the model obtains the greatest AUC of 99.68% and accuracy of 96.21% with a patch size of 2 and 2 embeddings (2, 2); the F1 scores are 94.91%, 95.54%, 96.07%, and 95.89%, respectively, for

Ablation Study – Anatomy-aware formulation						
	AUC (%)	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1 Score (%)
Model A	98.77	94.28	93.03	93.76	93.68	91.15
Model B	99.04	94.76	94.08	93.63	94.39	92.08
Model C	98.31	93.82	93.18	92.93	93.12	90.88
GBCHV	99.68	96.21	95.54	96.07	95.89	94.91

Ablation Study – Convolution							
Conv Stem	Conv Block	AUC (%)	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1 Score (%)
×	×	92.35	90.17	89.82	89.36	90.09	85.91
×	✓	95.01	92.78	91.46	91.63	92.27	88.93
✓	×	95.73	93.09	90.03	89.77	94.73	89.84
✓	✓	99.68	96.21	95.54	96.07	95.89	94.91

Ablation Study – Sizes of Different Embedding Ways							
p	$h\&v$	AUC (%)	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1 Score (%)
2	1	98.62	95.89	95.37	94.28	94.78	94.19
2	2	99.68	96.21	95.54	96.07	95.89	94.91
2	4	98.39	95.75	94.93	95.20	95.10	93.89
4	1	97.63	94.82	94.28	94.38	94.73	92.48
4	2	97.12	94.57	93.71	94.27	94.02	92.29
4	4	97.96	95.02	94.81	95.38	94.73	92.97
8	1	96.29	94.27	94.03	94.12	94.84	91.79
8	2	96.76	94.31	94.73	94.92	94.27	92.01
8	4	97.05	94.53	94.26	94.63	94.13	92.18

Table 1. Ablation Study.

precision, recall, and specificity. Performance somewhat declines with increasing patch size, as demonstrated by patch size 8 and 2 embeddings (8, 2), where the accuracy is 94.31% and the AUC is 96.76%.

Hyperparameter tuning

In this section, the hyperparameter tuning process of the proposed model is presented. With an epoch size of 100, we have conducted five individual experiment and the results are presented in Table 2.

During the hyperparameter tuning process, configurations were selected based on the highest accuracy achieved by the model. In the first study, we evaluated five different activation functions: Rectified Linear Unit (ReLU), Parametric Rectified Linear Unit (PReLU), Leaky Rectified Linear Unit (LeakyReLU), Gaussian Error Linear Unit (GELU), and Swish. Among these, the model achieved the highest performance of 94.32% with the activation function ReLU. In the subsequent study, we explored different loss functions, including Binary Crossentropy, Categorical Crossentropy, Mean Squared Error, Mean Absolute Error, and Mean Squared Logarithmic Error. Among them, Categorical Crossentropy resulted the same accuracy of 94.32%. Additionally, the model performed 95.45% with a batch size of 64. The performance of the model was further evaluated using four optimizers: Adaptive Moment Estimation (Adam), Nesterov-accelerated Adaptive Moment Estimation (Nadam), Adamx, and Root Mean Square Propagation (RMSprop). Among these, Adam achieved the highest accuracy of 95.45%. Lastly, an experiment was conducted by varying the learning rates: 0.001, 0.005, 0.0001, and 0.0005. A learning rate of 0.0001 demonstrated the highest accuracy of 96.21%. Therefore, our model's final classification accuracy of 96.21% was achieved with the configuration of activation function: ReLU, loss function: Categorical Crossentropy, optimizer: Adam, and learning rate: 0.0001.

Performance analysis of the proposed GCN model

To evaluate the performance of the GBCHV-Trans model confusion matrix, AUC (Area Under the Curve) and loss curve and others evaluation metrics are analyzed to assess the performance of the model. These assessments provide an evaluation of the diagnostic capabilities of the model, highlighting its advantages and outlining potential areas for improvement. Figure 10 shows the confusion matrix of the model.

In Fig. 10, the GBCHV-Trans model's confusion matrix shows how well it performs in dividing GBC into three categories: benign, malignant, and normal. Twenty-two cases are appropriately classified as malignant, 53 as benign, and 42 as normal by the model. Only 1 Normal instance was incorrectly classified as Benign, 2 Benign cases as Normal, and 1 Benign case as Malignant, indicating a low number of misclassifications. Furthermore, one case of malignancy is incorrectly identified as benign. These findings highlight the model's outstanding efficacy in diagnosing GBC, with very few false positives and false negatives and high reliability and precision.

The GBCHV-Trans model's performance further assessed by evaluating the loss curves and AUC. These curves provide important information about the model's training efficiency and class discrimination capabilities. The loss curve monitors the model's convergence during training, whereas the AUC curve aids in evaluating the

Study: 01 (Activation Function)			
Configuration no	**Activation function**	**Accuracy (%)**	**Findings**
01	relu	94.32	Highest accuracy
02	prelu	93.87	Accuracy dropped
03	leakyrelu	93.89	Accuracy dropped
04	gelu	93.14	Accuracy dropped
05	swish	93.6	Accuracy dropped
Study: 02 (Loss Function)			
Configuration no	**Loss function**	**Accuracy**	**Findings**
01	Binary Crossentropy	91.57	Accuracy dropped
02	Categorical Crossentropy	94.32	Highest accuracy
03	Mean Squared Error	93.07	Accuracy dropped
04	Mean absolute error	93.51	Accuracy dropped
05	Mean squared logarithmic error	93.98	Accuracy dropped
Study: 03 (Batch Size)			
Configuration no	**Batch Size**	**Accuracy (%)**	**Findings**
01	16	93.23	Accuracy dropped
02	32	94.32	Accuracy dropped
03	64	95.45	Highest accuracy
04	128	95.22	Accuracy dropped
Study: 04 (Optimizer)			
Configuration no	**Optimizer**	**Accuracy (%)**	**Findings**
01	Adam	95.45	Highest accuracy
02	Nadam	94.64	Accuracy dropped
03	Adamax	95.07	Accuracy dropped
04	RMSprop	95.22	Accuracy dropped
Study: 05 (Learning Rate)			
Configuration no	**Learning rate**	**Accuracy (%)**	**Findings**
01	0.001	95.45	Accuracy dropped
02	0.005	95.23	Accuracy dropped
03	0.0001	96.21	Highest accuracy
04	0.0005	95.64	Accuracy dropped

Table 2. Hyperparameter tuning.

model's diagnostic accuracy. A thorough analysis of these curves, emphasizing the model's functionality and training dynamics, is provided in the ensuing sections. Figure 11 shows the curves of this model.

The Receiver Operating Characteristic (ROC) curve is also presented to help explain how the model is verified during the training process. This curve provides deeper insights into the classification performance of the model by helping to visualize the trade-off between the true positive rate and false positive rate. Figure 12 shows the ROC curve of this model.

The GBCHV-Trans model performs exceptionally well in classifying gallbladder disorders into three categories: Normal, Benign, and Malignant, according to the ROC curve. A perfect AUC of 1.00 is achieved using the normal class curve, signifying perfect classification. With a high AUC of 0.96, the Benign class curve indicates very little misclassification. With an AUC of 0.99, the malignant class curve likewise exhibits almost perfect performance. Furthermore, the AUCs of 0.98 for the macro-average and micro-average ROC curves show steady and equitable performance across all classes. With high true positive rates and low false positive rates, these results emphasize the model's great diagnostic power and demonstrate its efficacy and reliability in the diagnosis of GBC.

In the next section, several performance metrics in terms of test accuracy, sensitivity, precision, specificity, NPV, FPR, FDR, FNR, F1 score and MCC are analyzed from confusion matrix. Table 3 shows the values of these metrics.

A detailed overview of the GBCHV-Trans model's diagnostic efficacy for gallbladder disorders can be found in the performance metrics table. The model performs robustly as demonstrated by excellent training accuracy of 97.72%, test accuracy of 96.21%, and validation accuracy of 95.15%. The model's ability to accurately identify positive instances and its precision in diagnosis are shown in the sensitivity (96.15%) and precision (95.99%) parameters. At 97.88% specificity, the model performs well in detecting negative cases. The F1 score of 96.06% shows that recall and precision were performed in a balanced manner. A low percentage of incorrectly positive predictions is indicated by the Negative Predictive Value (NPV) of 97.82% and the False Positive Rate (FPR)

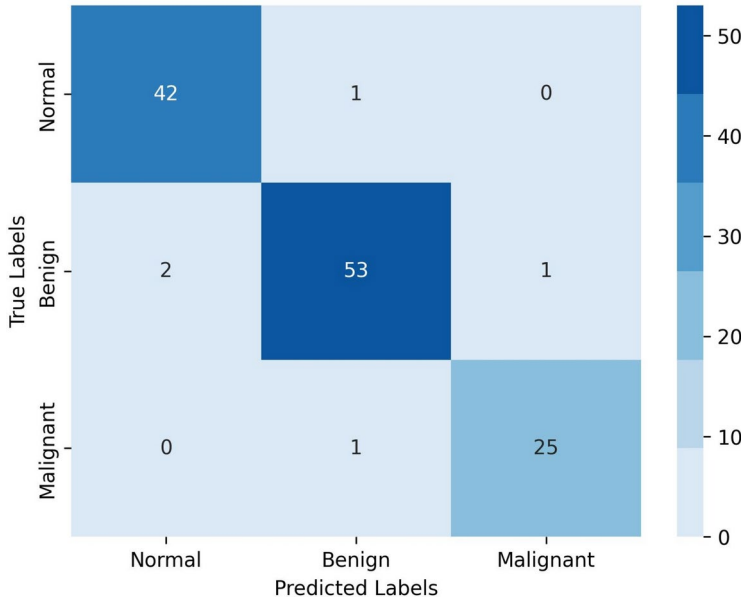

Fig. 10. Confusion Matrix of the proposed GBCHV model.

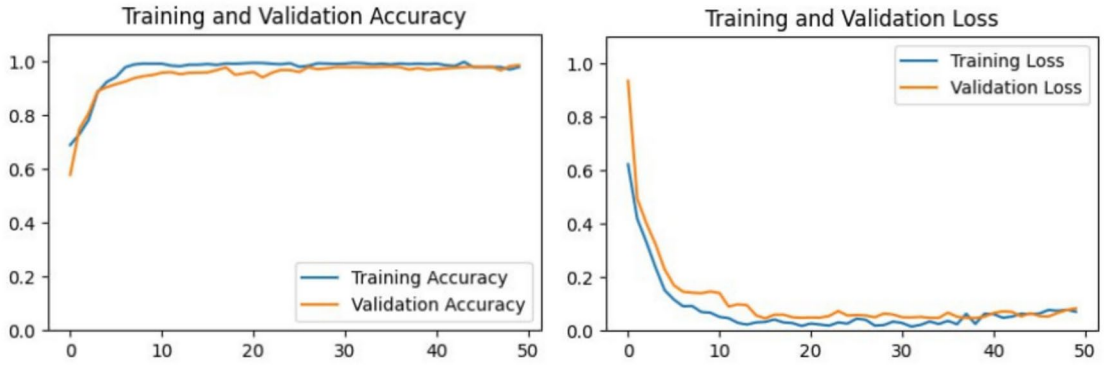

Fig. 11. Accuracy and loss curves of our proposed GBCHV-Trans model.

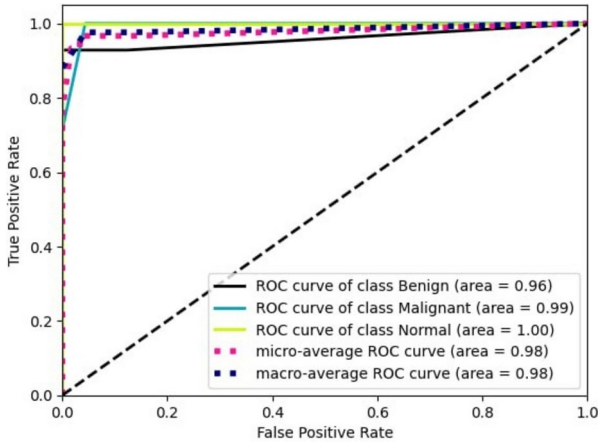

Fig. 12. ROC curve of our proposed GBCHV-Trans model.

Performance metrics	Results (%)	Performance metrics	Results (%)
Training accuracy	97.72	NPV	97.82
Test accuracy	96.21	FPR	18.83
Validation accuracy	95.15	FDR	4.06
Sensitivity	96.15	FNR	3.07
Precision	95.99	F1 Score	96.06
Specificity	97.88	MCC	93.92

Table 3. Evaluation metrics performance.

Models	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1 Score (%)
MobileNetv3	60.35	52.02	54.88	61.22	53.41
VGG16	62.89	55.21	56.96	67.12	56.07
ResNet50	75.24	68.75	69.84	78.84	69.29
InceptionV3	77.10	69.77	73.77	79.26	71.71
EfficientNet-B7	68.30	60.69	62.41	72.33	61.54
RetinaNet	72.21	65.68	66.67	76.02	66.17
DenseNet-264	69.63	60.44	61.8	74.56	61.11
GBCHV	96.21	95.54	96.07	95.89	94.91

Table 4. Comparison of Performance in Transfer Learning Models Considering Accuracy, Precision, Recall and F1 Score.

and False Discovery Rate (FDR), which are 18.83% and 4.06%, respectively. 3.07% is the False Negative Rate (FNR), which indicates a low percentage of missed diagnoses. The model's overall accuracy and reliability are highlighted by the Matthews Correlation Coefficient (MCC) of 93.92%, which shows a significant correlation between the predicted and actual classifications.

Comparison with other models

The performance parameters for each of the transfer learning models—Accuracy, Precision, Recall, Specificity, and F1 Score—are compared in the following Table 4. This comparison shows how well the GBCHV-Trans model is in the task of diagnosing GBC when compared to other well-known models such as MobileNetv3, VGG16, ResNet50, InceptionV3, EfficientNet-B7, RetinaNet, and DenseNet-264.

The GBCHV-Trans model performs significantly better than other transfer learning models, as the performance metrics table makes evident. With an accuracy of 96.21%, the GBCHV-Trans model outperforms the second-best model, InceptionV3, by a significant margin. The GBCHV-Trans model has the highest precision (95.54%), demonstrating greater capacity to identify positive cases. The GBCHV-Trans model's 96.07% recall rate demonstrates how well it captures true positive cases. The model's high F1 Score of 94.91% highlights its exceptional performance in accurately identifying negative situations, with a specificity of 95.89%. This comprehensive comparison highlights the GBCHV-Trans model's robustness and dependability in the detection of GBC.

K-fold cross validation

K-fold cross-validation represents a resampling methodology utilized to evaluate the performance of a model by partitioning the dataset into k equivalent segments, referred to as 'folds.' In this process, the model is trained utilizing $k-1$ folds while testing it on the remaining fold. This procedure is iteratively executed k times to ensure that each fold is utilized as a test set exactly once. The outcomes are averaged to yield a more robust estimation of performance. This approach mitigates the chances of overfitting and guarantees that the model's performance does not hinge on any singular subset of data. The selection of k is contingent upon the size of the dataset and the necessity to maintain an appropriate equilibrium between bias and variance in the assessment of the model. Figure 13 represents the k-fold evaluation for our proposed methodology.

To assess the efficacy and resilience of our GBCHV-Trans model, we utilized k-fold cross-validation as a methodical evaluation approach. The accompanying graph presents the accuracy outcomes across various k-fold configurations, specifically 3, 5, 7, 9, and 11 folds. The peak accuracy recorded was 96.13% with the implementation of ninefold cross-validation, followed by an accuracy of 95.89% attained with 11-fold cross-validation. Conversely, lower k values, such as 3 and 5 folds, resulted in marginally diminished accuracy, indicating that an increase in k values is associated with a more stable and precise model evaluation. This observed pattern implies that the model's training and validation in a broader array of subsets enhances its ability to generalize, thereby facilitating an overall enhancement in performance.

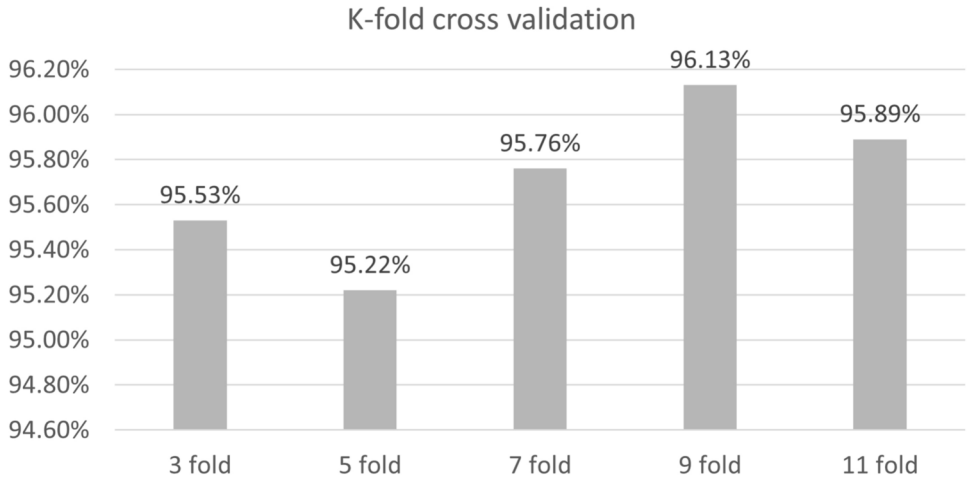

Fig. 13. K-fold cross validation.

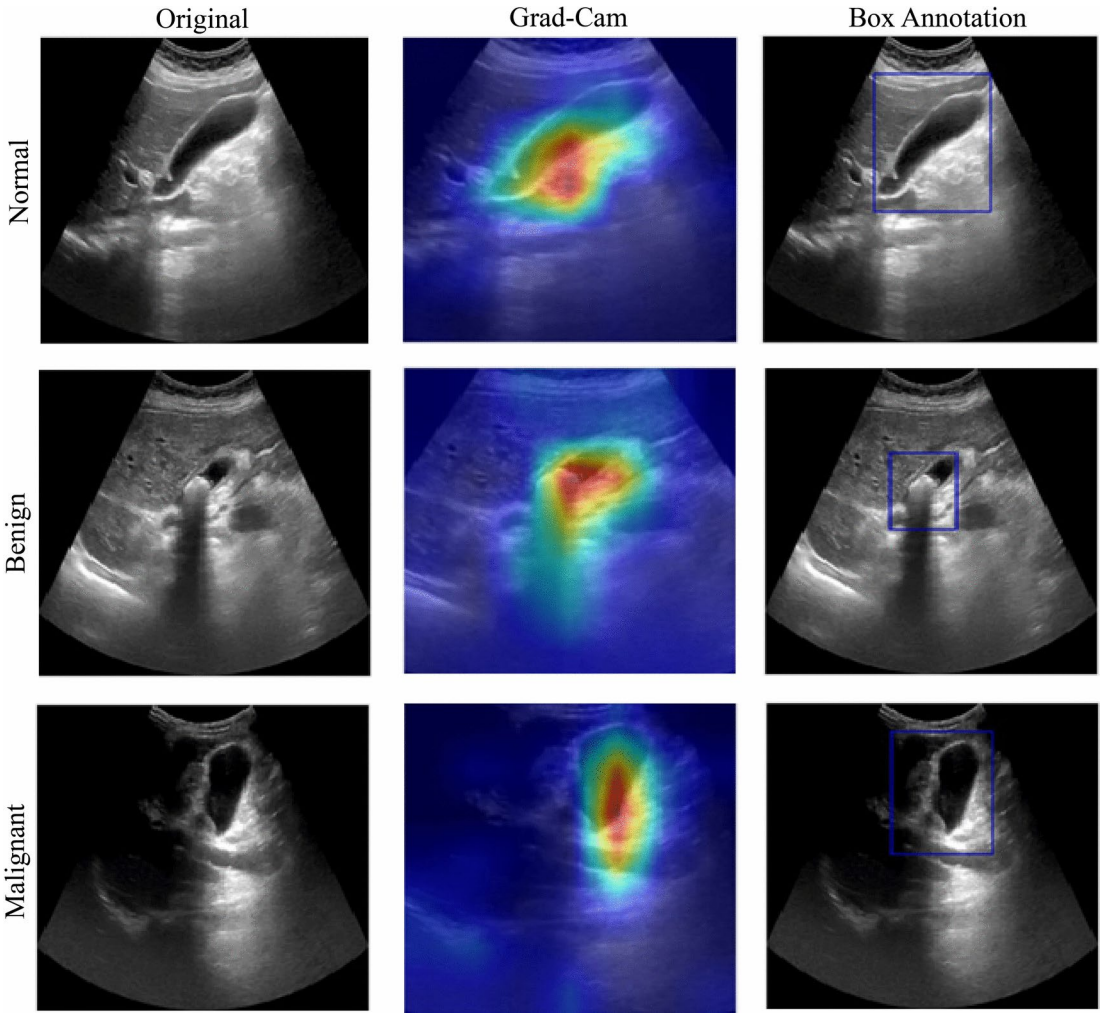

Fig. 14. The illustration of the GBCHV model's explainability utilizing Grad-Cam and Box Annotation.

Explainability of the model

Additionally, Explainable AI, such as Grad-Cam, is implemented to backtrack the model's decision-making process. This allows us to visualize and understand which specific features or portions of data are crucial for classifying cancers. Figure 14 shows the visualization of this process.

This illustration shows how to see and analyze the classification decisions produced by a model on ultrasound images using Grad-CAM (Gradient-weighted Class Activation Mapping). The three rows represent the three distinct classifications: benign, malignant, and normal. The original ultrasound image is shown in the first column of each row, followed by the Grad-CAM heatmap superimposed on the original image in the second column, and the original image with box annotations indicating the regions of interest in the third column. The regions that the model thinks most crucial for making decisions are displayed in the Grad-CAM heatmaps. While the heatmaps for the benign and malignant images highlight areas with noticeable features related to benign or malignant characteristics, the heatmap for the normal image concentrates on a region without substantial abnormalities. The box annotations help with the interpretation and validation of the model's classifications by providing a clear visual confirmation of these regions. The utilization of explainable AI improves the comprehension of the model's behavior by ensuring that the crucial areas impacting the judgments are readily apparent and logical.

Impact of attention layer through explainable AI

In our proposed model, the integration of an attention layer is fundamental in augmenting performance, as it enables the network to concentrate on the most pertinent regions of the input data. The attention mechanism assigns varying weights to distinct segments of the feature map dynamically, thereby allowing the model to prioritize essential patterns while diminishing the significance of less informative areas. This is especially critical in medical imaging applications, where subtle disparities in the data may signify important pathological alterations. By capturing long-range dependencies and contextual relationships throughout the image, the attention layer enhances the model's proficiency in detecting intricate details, resulting in more precise and interpretable outcomes. Additionally, the attention layer mitigates the propensity for overfitting by directing the network to acquire essential features, thereby rendering the model more robust and generalizable across a variety of datasets. We evaluated the effect of the attention layer by not adding it to the model and performed the Grad-CAM visualization. This illustration is shown in Fig. 15.

The attention layer assigns a weight to each element of the input image, representing how important it is to the prediction outcome. The weights are then used to integrate the information extracted from the image to create a weighted sum, which serves as input to the next network layer. This helps to improve the performance by allowing the model to focus on the most crucial aspects of the image[42]. The mechanism is represented by Eq. (17)

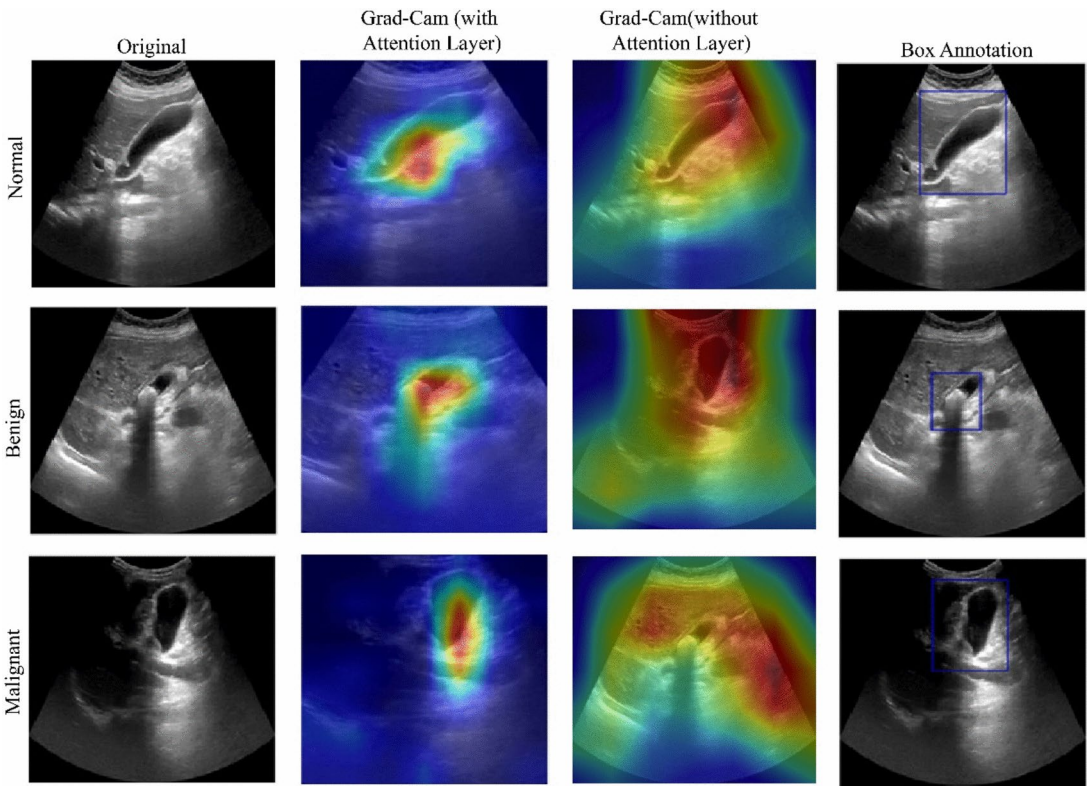

Fig. 15. The illustration of the effectiveness of the attention layer.

Articles	Dataset	Classification types	Apply image pre-processing techniques	Models	Accuracy
Lian et al.[43]	Ultrasound images from Gansu Provincial Hospital in Gansu, China (total = 60 patients)	1. Gallbladder 2. Gallstones	1. Otsu algorithm 2. Anisotropic diffusion algorithm	PA-PCNN: parameter-adaptive pulse-coupled neural network	86.01% (for the gallbladder) and 79.81% (for gallstones) (with mean similarity percentage of contour metrics)
Chen et al.[21]	Renji Hospital, China Total = 224 patients (73 samples of polypoid neoplasms, and 151 samples of cholesterol polyps)	1. polypoid neoplasms 2. cholesterol polyps	1. Contrast normalization 2. Resizing	Principal components analysis (PCA) and AdaBoost algorithm	87.54% (with segmentation) 73.03% (without segmentation
Basu et al.[20]	GBCU	1. Normal 2. Benign 3. Malignant	Visual acuity-based training curriculum to increase the sharpness	GBCNet: Object Detecting and multi-scale, second-order pooling-based classifier	91% (multi-class classification) 95.9% (binary classification)
Basu et al.[5]	GBCU	1. Normal 2. Benign 3. Malignant	N/A	RadFormer: Transformers with global–local attention	92.1%
Basu et al.[44]	GBUSV (Gallbladder USG videos) 32 malignant and 32 non-malignant videos containing a total of 12,251 and 3,549 frames	1. Malignant 2. Non-malignant	N/A	Unsupervised contrastive learning	92.1%
Jeong et al.[45]	Seoul National University Hospital (Total 535 patients)	1. nonneoplastic polyps 2. neoplastic polyps	Normalization	DL-DSS: -Deep learning-based decision support system (Inceptiov3 based)	Accuracy 85.7% And AUC 92%
Jang et al.[46]	Preoperative EUS dataset	Case1: 1. GB polys 2. Gallstones Case2: 1. Nonneoplastic 2. Neoplastic Case3: 1. Adenocarcinoma 2. Adenomatous	1. ROI in the form of a box (20 mm × 20 mm) containing lesions was extracted from the EUS image 2.Resize	EUS-AI (ResNet50	Accuracy 96.7% (GB polys vs. gallstones) Accuracy 89.8% (Nonneoplastic vs. Neoplastic) Accuracy 82.1% (Adenocarcinoma vs Adenomatous)
OURS	GBCU	1. Normal 2. Benign 3. Malignant	1. Median Filter 2. CLAHE	GBCHVS: GBC horizontal vertical stirp-based transformer	96.21% (GBCHVS proposed model)

Table 5. Performance comparison with prior studies.

$$f_{sa} = \gamma t \left(\left(\sum_{k=1}^{K} softmax(W_k * t) \right) \right) \tag{17}$$

Here k is the number of weights, and the aggregated weights produce an attention map, which is represented by α. α is multiplied by the feature value t. This is scaled by γ (learnable scaler) and the final output is the concatenation of these values, f_{sa}.

Comparison with prior studies

In this section, we offer an analysis to clarify our methodology considering relevant prior studies. To provide readers with a thorough knowledge of our proposed method's effectiveness, Table 5 includes information on each study, including the authors, dataset, classification type, image-preprocessing techniques, proposed models, and their accuracies.

Our model, GBCHVS, has demonstrated superior performance in classifying gallbladder conditions compared to existing researches. It achieves an accuracy of 96.21%, surpassing the results reported by Lian et al.[1] (86.01% for gallbladder and 79.81% for gallstones) and Chen et al.[2] (87.54% with segmentation). While Basu et al.[3] and Basu et al.[4] achieved high accuracies of 91% and 92.1%, respectively, for multi-class classification, our model maintains a lead in overall accuracy. Additionally, Basu et al.[5] achieved 92.1% using unsupervised contrastive learning, and Jeong et al.[6] achieved 85.7% accuracy and 92% AUC with DL-DSS. Jang et al.[7] reported high accuracy for specific classifications (96.7% for GB polys vs. gallstones), but our model provides a more comprehensive classification with higher overall accuracy. Despite these achievements, further validation with larger real-time datasets is necessary for our model.

Discussion and limitation

This study introduces a novel approach for classifying gallbladder conditions into benign, malignant, and normal categories, comprising advanced image processing and a transformer-based architecture. Two image enhancement techniques, including median filtering and CLAHE, were integrated to address the challenges posed by low-quality ultrasound images and shadow interferences, which further improved the input clarity. The GBCHV-Trans block, introduced in the model, integrates anatomical awareness and spatial relationships,

which enhanced the model's performance in discerning subtle differences in gallbladder tissue characteristics. This method outperforms traditional CNNs by transforming square visual elements into horizontal and vertical strips, allowing for more precise spatial feature extraction. Further, the comparative analysis with prior studies highlights the superiority of the proposed model over transfer learning techniques. While this framework shows significant promise for advancing diagnostic accuracy in ultrasound imaging, we aim to explore its efficiency and reliability on larger and more diverse datasets collected from hospitals in the future. Additionally, we aspire to include more classes of gallbladder cancer to facilitate progression analysis. Another area of focus is enhancing the model's performance on low-quality images, a common challenge in ultrasound imaging. In conclusion, the proposed model lays a solid foundation for improving gallbladder cancer diagnosis and paves the way for further advancements in clinical ultrasound applications.

Conclusion

This work presents an exceptional method for detecting GBC with the GBCU USG private dataset. Using sophisticated preprocessing methods and the incorporation of a transformer-based GBCHV model, we made significant progress in the classification task. The approach overcomes obstacles like image noise and unpredictability in ultrasonic imaging to discriminate between benign, malignant, and normal states with effectiveness. This methodology not only improves the accuracy of the diagnosis of GBC, but also provides important insights on the location aspects of the disease. Our research demonstrates the groundbreaking possibilities of integrating deep learning with medical imaging, providing doctors with an effective tool for GBC patients' early diagnosis and treatment. The GBCHV model outperforms conventional CNN methods in terms of classification accuracy, showing 96.21% overall diagnostic accuracy. Further validation of its efficacy and robustness is provided by comparative evaluations using transfer learning models. Moreover, our method has the potential to guide specific treatment plans, improve clinical results, and ultimately improve patient care in the management of GBC.

Data availability

The GBC USG (GBCU) dataset is available upon request to Dr. Pankaj Gupta and Dr. Chetan Arora. Details are provided at https://gbc-iitd.github.io/data/gbcu.

Received: 16 October 2024; Accepted: 4 February 2025
Published online: 28 February 2025

References

1. Kang, H. *et al.*, "EUS-guided FNA and biopsy for cytohistologic diagnosis of gallbladder cancer: a multicenter retrospective study." [Online]. Available: www.giejournal.org
2. Obaid, A. M., Turki, A., Bellaaj, H., and Ksantini, M. Diagnosis of Gallbladder Disease Using Artificial Intelligence: A Comparative Study, 2024, *Springer Science and Business Media B.V.* https://doi.org/10.1007/s44196-024-00431-w.
3. Xiang, F. et al. A deep learning model based on contrast-enhanced computed tomography for differential diagnosis of gallbladder carcinoma. *Hepatobiliary Pancreat. Dis. Int.* **23**(4), 376–384. https://doi.org/10.1016/j.hbpd.2023.04.001 (2024).
4. Adam, K. M., Abdelrahim, E. Y., Doush, W. M. & Abdelaziz, M. S. Clinical presentation and management modalities of gallbladder cancer in Sudan: A single-center study. *JGH Open* **7**(5), 365–371. https://doi.org/10.1002/jgh3.12906 (2023).
5. Basu, S., Gupta, M., Rana, P., Gupta, P. & Arora, C. RadFormer: Transformers with global–local attention for interpretable and accurate Gallbladder Cancer detection. *Med. Image Anal.* https://doi.org/10.1016/j.media.2022.102676 (2023).
6. Dadjouy, S. & Sajedi, H. Artificial intelligence applications in the diagnosis of gallbladder neoplasms through ultrasound: A review. *Biomed. Signal Process Control* https://doi.org/10.1016/j.bspc.2024.106149 (2024).
7. Yin, Y. et al. The value of deep learning in gallbladder lesion characterization. *Diagnostics* https://doi.org/10.3390/diagnostics13040704 (2023).
8. Kalage, D. et al. Contrast enhanced CT versus MRI for accurate diagnosis of wall-thickening type gallbladder cancer. *J. Clin. Exp. Hepatol.* https://doi.org/10.1016/j.jceh.2024.101397 (2024).
9. Zhang, Y.-D. et al., "COVID-19 classification using chest X-ray images based on fusion-assisted deep Bayesian optimization and Grad-CAM visualization."
10. Takahashi, K. et al. Recent advances in endoscopic ultrasound for gallbladder disease diagnosis. *Diagnostics* https://doi.org/10.3390/diagnostics14040374 (2024).
11. Barragán-Montero, A. et al. "Artificial intelligence and machine learning for medical imaging: A technology review. *Physica Medica* https://doi.org/10.1016/j.ejmp.2021.04.016 (2021).
12. van der Velden, B. H. M., Kuijf, H. J., Gilhuijs, K. G. A. & Viergever, M. A. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med. Image Anal.* https://doi.org/10.1016/j.media.2022.102470 (2022).
13. Ijaz, M. F. & Woźniak, M. Editorial: Recent advances in deep learning and medical imaging for cancer treatment. *Cancers* https://doi.org/10.3390/cancers16040700 (2024).
14. Iqbal, A., Sharif, M., Khan, M. A., Nisar, W. & Alhaisoni, M. FF-UNet: a U-shaped deep convolutional neural network for multimodal biomedical image segmentation. *Cognit. Comput.* **14**(4), 1287–1302. https://doi.org/10.1007/s12559-022-10038-y (2022).
15. Jabeen, K. et al. BC2NetRF: Breast cancer classification from mammogram images using enhanced deep learning features and equilibrium-jaya controlled regula falsi-based features selection. *Diagnostics* https://doi.org/10.3390/diagnostics13071238 (2023).
16. Obaid, A. M. et al. Detection of gallbladder disease types using deep learning: An informative medical method. *Diagnostics* https://doi.org/10.3390/diagnostics13101744 (2023).
17. Xia Yuan, H. et al. Differential diagnosis of gallbladder neoplastic polyps and cholesterol polyps with radiomics of dual modal ultrasound: a pilot study. *BMC Med. Imaging* https://doi.org/10.1186/s12880-023-00982-y (2023).
18. Rauf, F. et al. Artificial intelligence assisted common maternal fetal planes prediction from ultrasound images based on information fusion of customized convolutional neural networks. *Front. Med. (Lausanne)* https://doi.org/10.3389/fmed.2024.1486995 (2024).
19. Andrén-Sandberg, Å. Diagnosis and management of gallbladder cancer. *North Am. J. Med. Sci.* https://doi.org/10.4103/1947-2714.98586 (2012).
20. Basu, S., Gupta, M., Rana, P., Gupta, P., and Arora, C., Surpassing the Human Accuracy: Detecting Gallbladder Cancer from USG Images with Curriculum Learning. [Online]. Available: https://gbc-iitd.github.io/gbcnet.

21. Chen, T. et al. Computer-aided diagnosis of gallbladder polyps based on high resolution ultrasonography. *Comput. Methods Programs Biomed.* https://doi.org/10.1016/j.cmpb.2019.105118 (2020).
22. Karwande, G., Mbakawe, A., Wu, J. T., Celi, L. A., Moradi, M., and Lourentzou, I. CheXRelNet: An Anatomy-Aware Model for Tracking Longitudinal Relationships between Chest X-Rays, 2022, [Online]. Available: http://arxiv.org/abs/2208.03873
23. Yeganeh, Y., Farshad, A., and Navab, N. Shape-aware masking for inpainting in medical imaging, 2022, [Online]. Available: http://arxiv.org/abs/2207.05787
24. Fu, Z., Jiao, J., Yasrab, R., Drukker, L., Papageorghiou, A. T., and Noble, J. A. Anatomy-Aware Contrastive Representation Learning for Fetal Ultrasound, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Science and Business Media Deutschland GmbH, 2023, pp. 422–436. https://doi.org/10.1007/978-3-031-25066-8_23.
25. Kamal, U., Zunaed, M., Nizam, N. B. & Hasan, T. Anatomy-XNet: an anatomy aware convolutional neural network for thoracic disease classification in chest X-rays. *IEEE J. Biomed. Health Inform.* https://doi.org/10.1109/JBHI.2022.3199594 (2021).
26. Dalmaz, O., Yurt, M. & Cukur, T. ResViT: Residual vision transformers for multimodal medical image synthesis. *IEEE Trans. Med. Imaging* **41**(10), 2598–2614. https://doi.org/10.1109/TMI.2022.3167808 (2022).
27. Chen, J., He, Y., Frey, E. C., Li, Y., and Du, Y. ViT-V-Net: Vision Transformer for Unsupervised Volumetric Medical Image Registration, 2021, [Online]. Available: http://arxiv.org/abs/2104.06468
28. Liu, Q., Kaul, C., Wang, J., Anagnostopoulos, C., Murray-Smith, R., and Deligianni, F., Optimizing Vision Transformers for Medical Image Segmentation, in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2023. https://doi.org/10.1109/ICASSP49357.2023.10096379.
29. Mohd Sagheer, S. V. and George, S. N. A review on medical image denoising algorithms, 2020, *Elsevier Ltd*. https://doi.org/10.1016/j.bspc.2020.102036.
30. Juneja, M., Minhas, J. S., Singla, N., Kaur, R. & Jindal, P. Denoising techniques for cephalometric x-ray images: A comprehensive review. *Multimed Tools Appl.* **83**(17), 49953–49991. https://doi.org/10.1007/s11042-023-17495-z (2024).
31. Habib, M. et al. Convolved feature vector based adaptive fuzzy filter for image de-noising. *Appl. Sci. (Switzerland)* https://doi.org/10.3390/app13084861 (2023).
32. Ghosh, P. et al. SkinNet-16: A deep learning approach to identify benign and malignant skin lesions. *Front. Oncol.* https://doi.org/10.3389/fonc.2022.931141 (2022).
33. Singh, P., Mukundan, R. & De Ryke, R. Feature enhancement in medical ultrasound videos using contrast-limited adaptive histogram equalization. *J. Digit. Imaging* **33**(1), 273–285. https://doi.org/10.1007/s10278-019-00211-5 (2020).
34. Sadik, F., Dastider, A. G. & Fattah, S. A. SpecMEn-DL: Spectral mask enhancement with deep learning models to predict COVID-19 from lung ultrasound videos. *Health Inf. Sci. Syst.* https://doi.org/10.1007/s13755-021-00154-8 (2021).
35. dos Santos, J. C. M. et al. Fundus image quality enhancement for blood vessel detection via a neural network using CLAHE and Wiener filter. *Res. Biomed. Eng.* **36**(2), 107–119. https://doi.org/10.1007/s42600-020-00046-y (2020).
36. Hajeb Mohammad Alipour, S., Houshyari, M. & Mostaar, A. A novel algorithm for PET and MRI fusion based on digital curvelet transform via extracting lesions on both images. *Electron Phys.* **9**(7), 4872–4879. https://doi.org/10.19082/4872 (2017).
37. Siracusano, G. et al. Pipeline for advanced contrast enhancement (Pace) of chest x-ray in evaluating covid-19 patients by combining bidimensional empirical mode decomposition and contrast limited adaptive histogram equalization (clahe). *Sustainability (Switzerland)* **12**(20), 1–18. https://doi.org/10.3390/su12208573 (2020).
38. Dosovitskiy, A. et al., An image is worth 16X16 words: Transformers for image recognition at Scale." [Online]. Available: https://github.com/
39. Shamrat, F. J. M. et al. High-precision multiclass classification of lung disease through customized MobileNetV2 from chest X-ray images. *Comput. Biol. Med.* https://doi.org/10.1016/j.compbiomed.2023.106646 (2023).
40. Montaha, S. et al. BreastNet18: A high accuracy fine-tuned VGG16 model evaluated using ablation study for diagnosing breast cancer from enhanced mammography images. *Biology (Basel)* https://doi.org/10.3390/biology10121347 (2021).
41. Elpeltagy, M. & Sallam, H. Automatic prediction of COVID– 19 from chest images using modified ResNet50. *Multimed Tools Appl.* **80**(17), 26451–26463. https://doi.org/10.1007/s11042-021-10783-6 (2021).
42. Islam Bhuiyan, M. R. et al. Deep learning-based analysis of COVID-19 X-ray images: Incorporating clinical significance and assessing misinterpretation. *Digit. Health* https://doi.org/10.1177/20552076231215915 (2023).
43. Lian, J. et al. Automatic gallbladder and gallstone regions segmentation in ultrasound image. *Int. J. Comput. Assist. Radiol. Surg.* **12**(4), 553–568. https://doi.org/10.1007/s11548-016-1515-z (2017).
44. Basu, S., Singla, S., Gupta, M., Rana, P., Gupta, P., and Arora, C. Unsupervised contrastive learning of image representations from ultrasound videos with hard negative mining, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds., Cham: Springer Nature Switzerland, (2022), pp. 423–433.
45. Jeong, Y. et al. Deep learning-based decision support system for the diagnosis of neoplastic gallbladder polyps on ultrasonography: Preliminary results. *Sci. Rep.* https://doi.org/10.1038/s41598-020-64205-y (2020).
46. Jang, S. I. et al. Diagnostic performance of endoscopic ultrasound-artificial intelligence using deep learning analysis of gallbladder polypoid lesions. *J. Gastroenterol. Hepatol. (Australia)* **36**(12), 3548–3555. https://doi.org/10.1111/jgh.15673 (2021).

Acknowledgements

Not applicable.

Author contributions

Z.H. (First Author) (Corresponding Author): Project administration, Conceptualization, Supervision, Data curation. A.H.R.: Conceptualization, Methodology, Software. S.S.C Formal analysis, Visualization, Writing—original draft and editing. R.H.B: Investigation, Writing—original draft and editing. A.M.: Writing— review and editing.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.Z.H.

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

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