SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I3P110

Volume 13 Issue 3, 129-141, March 2026
© 2026 Seventh Sense Research Group®

Original Article

Hybrid Deep Learning for CT-Based Liver Lesion Classification

Trupti M. Kodinariya[1], Nikhil Gondaliya[2]

[1]*Gujarat Technological University, Ahmedabad, Gujarat, India.*
[2]*Department of Information Technology, G H Patel College of Engineering and Technology, V. V. Nagar, Gujarat, India.*

[1]*Corresponding Author : nikhilgondaliya@gcet.ac.in*

Received: 09 December 2025 Revised: 10 January 2026 Accepted: 11 February 2026 Published: 23 March 2026

Abstract *- Traditional visual evaluation is subjective and differs among observers, making an accurate identification of liver lesions on abdominal CT scans crucial for the prompt planning of treatment. In order to classify liver lesions as benign or malignant, this article presents a segmentation-free hybrid deep learning method that uses the LiTS-2017 dataset. The method involves combining systematic pre-processing with parallel feature extraction using Xception and Swin Transformer networks for synergistic local-global representations. Various fusion methodologies were evaluated, and a gated feature fusion mechanism was implemented to integrate distinct blocks utilizing channel and spatial attention modules, thereby adaptively weighting and recalibrating the salient characteristics prior to classification. This approach outperforms all other methods of model fusion and individual models in terms of accuracy (99.93%), recall (99.92%), F1 score (99.90%), and classification accuracy (99.93%), according to the experimental data. By eliminating the necessity for segmentation, our suggested architecture shows great promise as a trustworthy clinical decision support tool for differentiating between benign and malignant liver lesions using adaptive feature fusion.*

Keywords *- Abdominal CT imaging, Hybrid Deep Learning, LiTS17 Dataset, Liver lesion classification, Segmentation-free classification, Swin Transformer, Xception.*

1. Introduction

Early detection and evaluation of liver lesions are crucial for improving patient survival and developing effective treatment plans [20, 21]. The widespread use of abdominal Computed Tomography (CT) imaging in liver diagnostics, based on its ability to generate detailed cross-sectional images of abdominal organs and abnormal structures, supports its use as an important diagnostic tool [22]. Manual evaluation of CT images is time-consuming and is subject to inter-observer variation, supporting the development of Computer-Aided Diagnosis (CAD) systems to assist clinicians in making decisions [23, 25].

Deep learning techniques, specifically Convolutional Neural Networks (CNNs), have greatly enhanced the ability to perform automatic medical image analysis by learning hierarchical representations from medical imaging data instead of relying on manually designed features [26, 27]. CNN-based approaches have demonstrated strong performance in segmentation, detection, and classification tasks for a wide range of medical imaging applications. Many liver lesion classification approaches currently depend on the successful segmentation of the liver and/or the lesions themselves prior to classification; however, these approaches add additional costs (both computationally and financially) and increase the risk of classification errors due to poor segmentation [30].

Segmentation-free classification has been gaining attention as a way to address the shortcomings of segmentation-dependent classification approaches by eliminating the need for segmentation masks prior to classification[2, 3, 6, 12]. While these approaches make it easier to develop classification pipelines, some still require a single network architecture to classify CT images, and the networks primarily focus on capturing local texture characteristics rather than modeling longer-range contextual dependencies required to accurately characterize liver lesions.

Transformer-based models like Vision Transformers (ViTs) have demonstrated their ability to learn global contextual relationships by utilizing self-attention mechanisms when compared to CNNs that primarily capture local texture patterns. Hybrid CNN-Transformer architectures have shown potential in leveraging complementary representations to improve performance by integrating CNNs that extract local textures with a Transformer that captures global contextual dependencies. Most current hybrid architectures utilize simple fusion strategies (e.g., concatenation or averaging) to combine heterogeneous features generated by CNNs and Transformers; however, these fusion strategies do not have the ability to adaptively select the most relevant features to optimize representation quality.

In recent times, there has been progress, but there is still a noticeable absence of segmentation-free frameworks. These frameworks need to effectively integrate CNN and Transformer representations through adaptive feature fusion mechanisms. The integration should be capable of emphasizing diagnostically relevant information. As a result, there is a clear need for a robust, segmentation-free hybrid framework in clinical environments. This framework should combine complementary representations using adaptive attention-driven fusion to enhance classification reliability.

1.1. Contributions of the Study

This study's primary contributions are:

1. Development of a hybrid deep learning framework that integrates CNN and Transformer architectures for segmentation-free, automated classification of liver lesions.
2. Introduction of a gated feature fusion mechanism that uses both channel and spatial attention for adaptive feature selection.
3. Evaluation of multiple different fusion strategies to determine the optimal methods to integrate features
4. Validation on the LiTS-17 dataset to demonstrate state-of-the-art performance while maintaining clinical deployment feasibility.

The rest of this paper will be organized as follows: Section 2 provides a review of the literature regarding segmentation-free liver lesion classification. Section 3 details the proposed hybrid framework and fusion methods. Section 4 details the experimental configuration and evaluation results. Finally, Section 5 concludes this study and highlights areas for future research.

2. Related Work

Recent studies in automated liver lesion analysis tend to focus more on the learning-based end-to-end classification directly from abdominal CT images without requiring dedicated liver or lesion segmentation. These segmentation-free methods can truncate pipelines, save annotated cost, and are usually more robust to the deformation of liver shape and lesion boundary. The following section covers studies (2018-2024) that fall into methodological themes, as identified through a selection of existing surveys. The Summary of these studies is shown in Table 1.

2.1. Whole-Image, Patch, or Volume-Based Classification

Numerous studies investigate the categorization of lesions using comprehensive CT slices or specific square sections without the necessity of a segmentation mask. A comprehensive discriminative deep network using InceptionV3 and residual connections was developed to differentiate cysts from metastases [2].

The generalization of CNN to small datasets was enhanced using GAN-based synthetic lesion patch synthesis [3]. For binary tumor detection, texture and SSIM characteristics were computed from whole CT slices using conventional methods [4, 5]. A Multiphase Convolutional Dense Network (MP-CDN) was proposed to combine arterial, portal, and delayed phases for directly learning enhancement dynamics [6].

The subsequent study [9] further proposed a unified end-to-end classifier for multiphase information incorporation into the quantification of PVEs. Compared to 2D and even 3D DenseNets, they have demonstrated that the proposed 2.5D slice-stacking is computationally cheaper while not significantly compromising accuracy [12].

2.2. Detection and Classification as a Unified Task

Another popular research direction is to combine detection and classification in a single pipeline, leaving out the explicit segmentation. A cascaded CNN model was presented to simultaneously localize and classify FLNs on multiphase CT imaging [1]. Lim et al. designed an anchor-free detector together with dynamic-texture learning for end-to-end large-scale lesion detection and classification [7].

An attention-driven model enhanced the robustness to unregistered multiphase CT data [8]. An additional study using a CNN classifier applied to portal vein CT showed the ability to distinguish benign from malignant lesions without performing segmentation at all [17]. These studies demonstrate the possibility of joint classification pipelines all the way.

2.3. ROI-Based and Annotation-Light Classification

Segmentation masks are very time-consuming to annotate; therefore, many methods use lesion-center coordinates or coarse ROIs. ROI-cropped CNNs with small CT patches centered on the lesions showed high accuracy, working with small annotations [2, 3, 6, 12]. Such annotation-light models can preserve enough context information but drastically decrease the requirement of labeling and computation.

2.4. Lightweight and Handcrafted Feature Approaches

Former segmentation-free classification relied on handcrafted statistical features. Texture-Based SVM classifiers and SSIM-based detection frameworks that provided reasonable accuracy on small data [4, 5]. Hybrid models that combine the wavelet and texture features later achieved better discrimination of lesions [13]. While inferior to CNN results, these interpretable methods are still relevant in low-data and interpretability settings.

Table 1. Summary liver lesion classification methods using direct abdominal CT input

Authors	Input Type	Model / Key Idea	Main Contribution
Zhou et al. [1]	Multiphase CT	Hierarchical CNN (detection + classification)	Joint lesion localization and classification.
Perdigón et al. [2]	ROI slices	InceptionV3 + residual classifier	End-to-end classification
Frid-Adar et al. [3]	ROI patches	GAN-augmented CNN	Better performance with augmented limited data.
Hussain et al. [4]	Whole CT slice	Texture + SVM	Classical slice classification without segmentation.
Siddiqi et al. [5]	Whole CT	SSIM + SVM	Fast tumor-presence check.
Cao et al. [6]	Multiphase CT stack	DenseNet (MP-CDN)	Demonstrated contrast-phase dynamics end-to-end.
Huo et al. [7]	Multiphase CT	Anchor-free detector + texture learning	Joint detection + classification on a large dataset.
Lee et al. [8]	Unregistered multiphase CT	Attention-guided CNN	Phase-misalignment-tolerant classification.
Zhao et al. [9]	Multiphase CT	Unified temporal fusion CNN	Multiphase end-to-end classification.
Wu et al. [10]	Multiphase CT	MULLET framework	Pre-training resource for classification backbones.
Shen et al. [11]	CT/MRI	Explainable CNN + CAM	Visual explanation for segmentation-free models.
Stollmayer et al. [12]	CT/MRI	2D vs 3D DenseNet comparison	Showed efficiency of 2.5 D classification.
Phan et al. [13]	CT/MRI	Combines Hounsfield Unit (HU) attenuation analysis with deep learning features from CT/MRI	HU-guided deep learning for improved liver lesion classification
Lee et al. [14]	Full CT volume	CNN classification using lesion-centered augmented patches	Synthetic augmentation improves small-lesion classification accuracy.
Amitojdeep Singh et al. [15]	CT/MRI	Attention-based explainable CNN	Improved interpretability in classification.
Lyu et al. [16]	CT volume	Weakly supervised detector	minimal pixel-level annotation
Chiu et al. [17]	Portal-venous CT	CNN classifier	Binary benign vs malignant classification.
Cao et al. [6]	Multiphase CT	Multiphase fusion network	Spatio-temporal modeling of enhancement dynamics.
Qaio et al. [18]	Multiphase CT	Spatio-temporal fusion CNN	Learned inter-phase transitions without masks.
Bilic et al. [19]	CT volumes	LiTS benchmark dataset	Transfer learning
Azad et al. [37]	Multi-modality medical images	Vision Transformer review and comparative analysis	Comprehensive analysis showing benefits and limitations of ViT models in medical imaging tasks.
Guo et al. [38]	Medical image datasets	UCTNet: Uncertainty-guided CNN–Transformer hybrid	Adaptive fusion mechanism improves robustness by activating transformer modules selectively.
Al-Hejri et al. [39]	Medical image datasets	Hybrid CNN + Vision Transformer classifier	Demonstrated improved classification accuracy using a hybrid architecture under class imbalance conditions.
He et al. [40]	Liver medical images	CMT-Net: Compact CNN–MLP–Transformer hybrid	Proposed computationally efficient hybrid network achieving competitive classification performance.

2.5. Multiphase Stacking and Dynamic Enhancement Modeling

The dynamic enhancement characteristics at different CT phases serve as diagnostic indicators. Multiphase dense networks achieve over 80% accuracy by integrating multiple CT phases as input channels [6]. Attention-based fusion networks have further generalized this concept by recording joint temporal and spatial enhancement patterns throughout the phases [6, 9, 18]. These contrast-free architectures utilize the temporal evolution of physiological contrast to improve the precision of classifier labels.

2.6. Data Scarcity and Weakly Supervised Learning

The lack of data is still a limitation in segmentation-free models. GAN-based data augmentation enhanced generalization of the model [3, 14], and bootstrapped detection and weakly supervised pseudo-labeling methods enabled us to employ large-scale end-to-end training in the image domain [7, 16]. There are a few semi-automatic systems, including MULLET, that pre-train features for classifier backbones [10]. Transfer learning with the LiTS17 dataset improves feature robustness [19].

2.7. Explainability and Clinical Interpretability

Many methods are not interpretable, which is essential for clinical use. Interpretable CNNs via class activation maps show the model's reasoning process [11]. Attention-augmented architectures yield region-based saliency visualization to boost radiologist trust [15]. These visualization modules allow the mask-free model to generate clinically interpretable results without explicit masks.

In recent years, there has been an explosion of Vision Transformer (ViT) applications in medical imaging, as documented by both surveys and systematic reviews that show an increase in the number of large-scale studies and a strong advantage to model long-range dependencies versus pure CNNs [37]. ViT-based models provide enhanced global context modeling, advantageous for tasks where lesion appearance and relative spatial relationships are significant; however, they may require substantial data and can be less efficient in encoding intricate local textures in the absence of convolutional priors.

Recent hybrid CNN-Transformer architectures have aimed at combining CNNs' texture sensitivity with the transformer's ability to capture global context. Examples of such architectures include UCTNet, which uses uncertainty estimates as a basis for selectively triggering transformer modules in regions of the image that CNNs find unreliable – thus reducing redundancy and computational overhead while enhancing robustness [38]. Hybrid architectures were also developed using stage-wise or gated fusion to improve feature integration from both streams [40]. The studies cited above demonstrate that intelligent fusion (e.g., uncertainty-based, gated, or multi-stage) typically outperforms simple methods like concatenation and averaging, particularly in cases of heterogeneous clinical datasets.

A number of empirical studies have recently been conducted to evaluate how well different types of hybrid designs function in terms of modality and organ. Al-Hejri et al. [39] found that a hybrid CNN-ViT design improved the AUC value for cervical cancer diagnosis and exhibited increased robustness to class imbalances. [38] stated that their study approach, UCTNet, outperformed existing methods by dynamically directing transformer attention to the most detailed areas of the input image. Multiple researchers who conducted studies in 2025, such as CMT-Net and hybrid ensemble ViT-CNN, discovered that their models demonstrated competitive and/or state-of-the-art classification performance while meticulously managing the parameter budget, FLOPS budget, and interpretability of their models. This body of work demonstrates that (1) employing both adaptive fusion and selective application of self-attention diminishes overfitting and lowers the computational expenses of training deep neural networks and (2) despite the effectiveness of certain techniques in mitigating overfitting, the integration of data augmentation and regularization is essential when incorporating ViT components into a network architecture.

The emerging hybrid literature demonstrates potential; however, the majority of current hybrids either (a) concentrate on segmentation or volumetric tasks rather than segmentation-free slice-level classification, (b) employ static fusion methods (such as concatenation or fixed attention blocks) instead of image-adaptive gated fusion, or (c) present results on modalities or datasets dissimilar to LiTS17, complicating direct comparisons. Our research tackles all three deficiencies by " (i) assessing segmentation-free binary classification on LiTS-17, (ii) developing a novel dynamic fusion strategy that combines gated fusion with sequential channel and spatial attention (CBAM/SE style) to dynamically adjust the contribution weights of each backbone component for each image and (iii) conducting an extensive comparative analysis of seven different fusion strategies".

3. Proposed Hybrid Model

The suggested methodology for this research utilizes a hybrid deep learning model. A segmentation-free hybrid deep learning model will be developed, which can utilize abdominal CT images in order to classify liver lesions. Therefore, the model does not require explicit liver or lesion segmentation. In addition to this, the model utilizes two different architectures. One is a CNN model (Xception), and the other is a Transformer-based model (Swin Transformer). These architectures learn complementary feature representations. As opposed to previous studies that have relied upon segmentation-dependent pipelines or utilized static fusion methodologies, the proposed approach employs an adaptive gated fusion with both channel and spatial attention mechanisms. The use of these attention mechanisms allows the model to emphasize diagnostically relevant features as needed adaptively. The combination of these architectures, along with their use of adaptive gated fusion with channel and spatial attention mechanisms, provides an ability to effectively integrate the local texture patterns and the global contextual information. This results in a more robust classification system when comparing it to previous hybrid or single network models.

The framework of the proposed method is composed of four main stages: pre-processing, feature extraction, feature fusion, and classification, as illustrated in Figure 1.

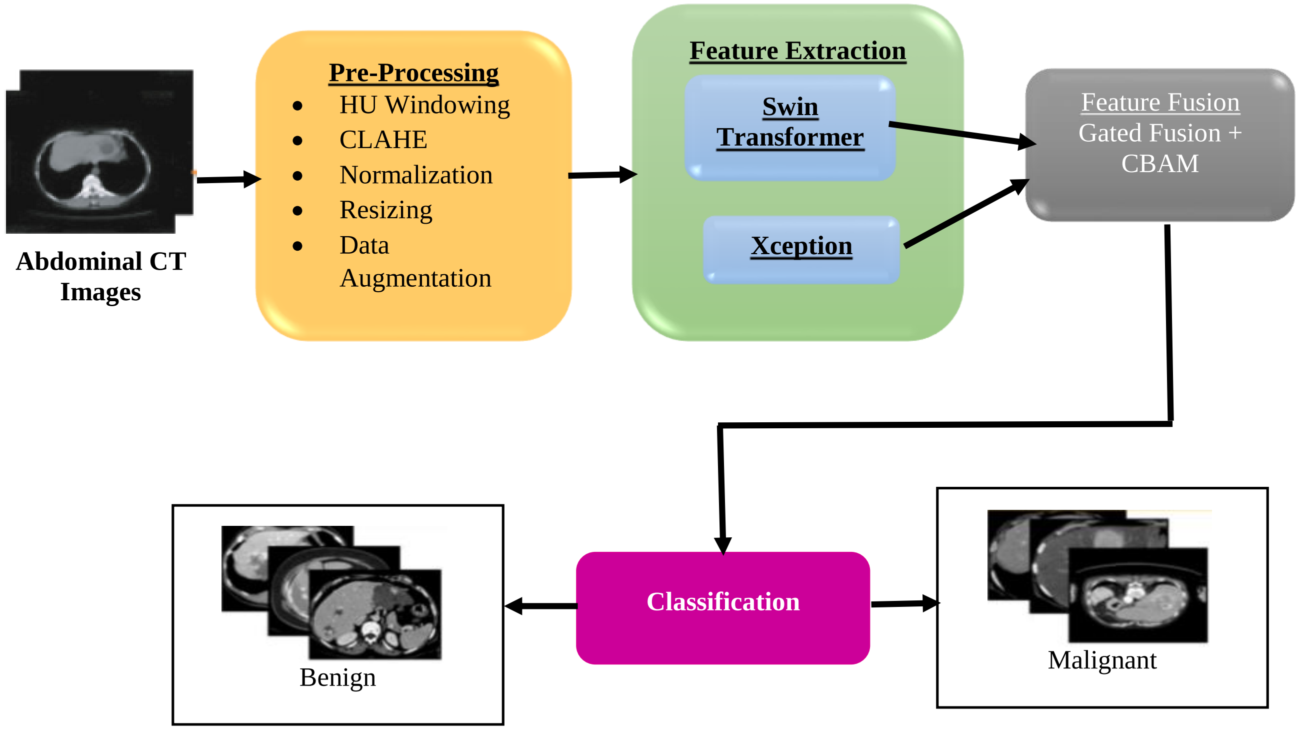

Fig. 1 Hybrid deep learning framework for classification of liver lesions with abdominal CT images

3.1. Data Set

We employ in this work the Liver Tumor Segmentation Challenge 2017 (LiTS17) [19] dataset, a benchmark dataset provided for liver lesion segmentation from CT scans. The Dataset Consists of a total of 130 and 70 contrast-enhanced 3D abdominal CT scans from several clinical centers with different scanners and protocols for training and testing, respectively. Accordingly, the datasets exhibit highly varying spatial resolution and fields of view.

The in-plane resolution ranges from 0.60 mm to 0.98 mm, and the slice spacing ranges from 0.45 mm to 5.0 mm. All the scan axial slices are the same size, 512 x 512, but the number of slices varies significantly and ranges from 42 to 1026. The publicly available LiTS17 dataset includes only the ground truth (GT) of liver and lesion segmentation for the training split. A GT lesion type is not available.

Khaled et al [30] published a paper where lesions are divided into three categories: normal, benign, and malignant (4 Normal, 41 Benign , and 85 Malignant) for the LiTS17 training dataset.

3.2. Pre-Processing

To enhance the quality of CT images and improve feature learning, the following pre-processing steps are applied:

- HU Windowing: CT images are windowed to [-30, 150] Hounsfield Units to emphasize the organ of interest, like the liver parenchyma and lesions. It serves to suppress irrelevant structures (e.g., bones or air), enhance contrast, and better model feature extraction.
- CLAHE (Contrast Limited Adaptive Histogram Equalization): It is applied to improve the local contrast of CT images. When compared to global histogram equalization, CLAHE's 7x7 image tile diffusion effectively eliminates noise while revealing fine features. It comes in use for medical imaging, in particular when there is poor contrast between different types of tissue and a crucial feature (like a tumor) that needs to be examined. The approach additionally made use of a clip limit (0.2) to prevent the histogram equalization inside each tile from amplifying noise to an unreasonable degree. At last, the produced tiles are combined using bilinear interpolation.
- Normalization: To ensure that the numerical representation of the CT pixel intensities was consistent

across all input images and to keep the relative contrast differences intact, Min-Max normalization was used to scale them into the range [0,1]. Training gradient updates and model convergence are less affected by differences in scanners, acquisition schemas, and patient anatomy when intensities are normalized.

- Resizing: The input size requirement of deep learning architectures dictated that each CT slice be resized to 224×224 pixels after normalization.
- Data Augmentation: The dataset contains 6930 CT images, with 2,602 benign cases and 4,766 malignant instances, resulting in a class imbalance, which can be addressed through data augmentation. Classical image manipulation techniques were used to enhance data from the minority (benign) class in order to increase the model's generalizability and address this problem. The classical augmentation methods (Horizontal Flip, Vertical Flip, Rotations (90°, 180°, and 270°), and Random Shifting by 10-15 pixels in horizontal and vertical directions)were employed: These transformations were applied to synthetically increase the number of benign samples and ensure a more balanced distribution between classes, thereby reducing bias during model training

3.3. Feature Extraction

The framework employs two parallel fine-tuning networks (Swin Transformer and Xception) for complementary feature representations. The Swin Transformer branch is fine-tuned from pre-trained weights on ImageNet and does not contain a classification head. The model adopts a window-based hierarchical self-attention to capture global context information and long-range spatial dependencies.

Meanwhile, the Xception network adopts depthwise separable convolutions that can adaptively extract local texture details and structure data, which are sensitive to liver injury descriptions. The corresponding feature maps of two branches are globally averaged-pooled to generate a compact feature vector. Then these vectors are projected into the same 512-dimensional embedding space with fully connected layers to maintain consistent dimensions.

The generated feature representations from the two paths pertaining to global context and local texture information are fused and input into the classification module for the prediction of benign and malignant lesions.

3.4. Feature Fusion

To strengthen the discriminative power of these deep features, various fusion strategies were designed. Each approach serves to fuse features from the Swin transformer network and the Xception network together to construct an overall representation for better liver lesion classification. The learned dimensional feature vectors of both branches are then combined in the following fusion operations and evaluated.

3.4.1. Method 1: Concatenation

It simply stacks feature vectors (or maps) from multiple backbones along the channel dimension and passes the result to the classifier (or to further projection layers).

$$if\ F_1 \in R^{H\ x\ w\ x\ C_1}\ and\ F_2 \in R^{H\ x\ w\ x\ C_2},$$
$$fused\ \ F = [F_1, F_2] \in\ R^{H\ x\ w\ x\ (C_1+C_2)}$$

Effect on image features: No interference; it retains all the information in each backbone. The classifier can then be learned to exploit whichever channels are found predictive.

Pros / Cons / Failure modes.

- Very simple, parameter-free
- Keeps full representational capacity.
- There is no explicit interaction between backbones; redundancy might be present, and the classifier needs to learn to ignore the useless channels, which may depend on more data and parameters. Memory and parameter count increase. It tends to over-fit on small datasets.

3.4.2. Method 2: Average

It computes the element-wise average from the two feature vectors, effectively merging their information in a balanced way. Element-wise average of corresponding features: F=1/2*(F_1+F_2). Works when feature maps are spatially aligned and have matching channel size.

Effect on image features: Generates a smoothed, consensus representation. Mitigates noise coming from a given backbone but can wash out complementary signal (e.g., different texture attributes from one network may be suppressed).

Pros / Cons / Failure modes.

- Very light-weight, no added parameters. Helps when both backbones provide similar, redundant signals.
- Destroys complementary features; if two models specialize in different cues, averaging can harm discriminative power.

3.4.3. Method 3: Hierarchical / Multi-Level Feature Fusion

This method extracts intermediate and deep-level features from both backbones, enabling multi-scale representation learning. Features from different depths are concatenated and passed through dense projection layers to capture both local and global contextual information. Each feature map undergoes GlobalAveragePooling2D and Dense projection to reduce dimensionality. These projected representations — capturing both mid-level (texture/edges) and deep-level (semantic) information — are concatenated for final fusion.

Effect on image features: Multilevel fusion is able to consolidate information at different scales: small lesions and fine textures from early/mid layers/early hierarchy levels, and global lesion context abstractions from deeper/highest levels.

For CT lesions, this helps combine boundary sharpness and structural context.

Pros / Cons / Failure modes.

- Usually yields better robustness to lesion size and heterogeneity, empirically strong.
- More complex: needs careful alignment (spatial sizes and channels), more parameters, and possible redundancy across scales.

3.4.4. Method 4: Concatenation + Channel Attention

In this method, Concatenate as above, but then have a channel-wise attention module (e.g., SE block) that calculates global descriptors (GAP) → bottleneck MLP → sigmoid gating per-channel and multiplies these channel weights to the fused map.

Effect on image features: The SE module learns "which feature channels matter" and reweights them, effectively suppressing unhelpful channels introduced by concatenation and boosting discriminative ones. This implicitly performs a learned feature selection over concatenated channels.

Pros / Cons / Failure modes.

- Lightweight and effective at removing redundant or noisy channels. Improves performance over plain concatenation in many settings.
- Only models channel-wise importances—not spatial localization. If the unwanted features are spatially localized rather than entire channels, channel attention alone may underperform.

3.4.5. Method 5: Concatenation + Spatial Attention

Both network features were concatenated and enriched by spatial attention, which enables the model to concentrate on regions in the CT slices pertinent to lesions without employing additional channel weighting schemes. Compute a spatial attention map after concatenation: average and max across channels → concatenate → conv7×7 → sigmoid → multiply with the fused map. This highlights the spatial locations, rather than channels.

Effect on image features: Improves localization by learning where in the image to focus (e.g., lesion pixels vs background). Especially useful when lesions are in small image areas or the background is cluttered.

Pros / Cons / Failure modes.

- Helps focus on lesion regions, complements channel attention.
- If the two backbones disagree about location (misalignment), spatial attention might produce diffused maps. Spatial attention alone does not solve channel redundancy.

3.4.6. Method 6: Concatenation + CBAM Fusion (Channel + Spatial Attention)

This approach incorporates the channel attention and spatial attention with the CBAM (Convolutional Block Attention Module). The Swin Transformer and Xception features are spatially rescaled, projected to the same channel sizes, and concatenated. The concatenated tensor is sent through the channel attention sub-module, as well as the spatial attention sub-module. The CBAM progressively refines features along both the channel and spatial dimensions in terms of expressiveness and focusing capability. Following the application of attention, the final map is then followed by global-pooling and classification. Compared to the single-attention designs, this architecture is an improvement that might be seen as an improvement since it enhances the selectivity of individual features.

Effect on image features: Combines the benefits of channel and spatial reweighting by selecting informative filters and then putting an emphasis on informative spatial locations, which ultimately results in more discriminative fused representations from the combined representations.

Pros / Cons / Failure modes.

- The fusion operators are both empirically robust and very inexpensive in comparison to those that are more sophisticated. Enhances both precision and recall in a wide variety of medical imaging jobs.
- It is still a post-hoc refinement of concatenation, which means that it does not describe explicit cross-backbone interactions. For example, one backbone's attention to another's position is not modeled.

3.4.7. Method 7: Proposed Method: Gated Fusion + Squeeze-and-Excitation (SE) + Spatial Attention

To improve lesion discrimination, the suggested hybrid model combines gated features with sequential channel- and spatial-attention refinement. The initial step in merging features from Xception and Swin Transformer is to utilize a learnable gating mechanism that determines the relative importance of each backbone dynamically. In order for the network to concentrate on diagnostically valuable feature channels, the fused feature is then passed via a Squeeze-and-Excitation (SE) block, which performs channel-wise recalibration using global average pooling and bottleneck fully linked layers.

Then, to isolate areas that are important for lesion detection, a spatial attention block creates a two-dimensional attention map while hiding structures in the backdrop. With this sequential SE-spatial attention approach, the model can learn which characteristics are most important and how to focus on specific areas of the image, creating a feature space that is both discriminative and highly aware of its context. At last, a completely connected classification head receives the revised feature vector.

Effect on image features: Gated fusion enables adaptive selection—per image and per feature dimension—of which backbone should contribute more. This avoids the "both-or-none" behaviour of concatenation and the information loss of averaging. Coupled with CBAM, it then refines "what" and "where" in the fused map. For CT lesion tasks, gating lets the model prefer shape features for some lesions and texture features for others.

Pros / Cons / Failure modes.

- Highly adaptive, empirically shown to give the best performance.
- Slightly more parameters and computation for gate MLPs. If training data is extremely small, gates might overfit unless regularized.

4. Results and Discussion

This section demonstrates the outcome attained by the proposed model. In addition, a comparative analysis of the respective model with existing mechanisms is illustrated.

4.1. Experimental Setup

All of the tests were done in Google Colab Pro using Accelerated for NVIDIA GPU. The LiTS17 dataset, which is publicly available, was used as the source of the data. Only abdominal CT images were used as the input modality; no Region Of Interest (ROI) extraction or segmentation mask was used as input. The total number of abdominal CT slices in the dataset, after pre-processing and augmentation, was 9,532. The total number of abdominal CT slices was then randomly divided into a training subset, validation subset, and test subset in a ratio of 70%, 10% and 20%, respectively. The training subset was used to learn the model, the validation subset was used to select optimal hyperparameters and prevent overfitting of the model; the test subset was used to evaluate the performance of the final model. To further limit the effect of sampling bias, random subdivision was performed multiple times with different seed values. Table 2 shows the parameters used in the framework.

Table 2. Model Parameter

Parameter Name	Value
Optimizer	Adam for Xception and AdamW for Swin Transformer
Learning Rate	0.00013
Batch Size	16
Epochs	50-100
Loss Function	Binary Cross-Entropy
Evaluation Metrics	Accuracy, Precision, Recall, and F1-Score

4.2. Quantitative Assessment of Distinct Models

The LiTS17 dataset was used to fine-tune six cutting-edge deep learning models, which are as follows: VGG16, InceptionResNetV2, ResNet50, EfficientNetB2, Xception, and Swin Transformer. The results of these efforts are presented in Table 3. Swin Transformer (97.82%), InceptionResNetV2 (97.5% accuracy), and Xception (96.5% accuracy) were the three models that displayed the highest level of competence in capturing liver lesion characteristics, as shown in Table 2. EfficientNetB2 and ConvNeXt-Tiny, on the other hand, demonstrated inadequate generalizability, which indicates that they lacked sufficient sensitivity to minor lesion patterns that are typical in abdominal CT imaging investigations. These findings justify the need for hybrid feature learning rather than relying on a single architecture. The hybrid models were experimented with concatenation as fusion techniques, where fusion of the Swin transformer and Xception gives comparatively good performance with other experimented models.

Table 3. Experiment result on Individual Models

Model	Precision	Recall	F1-Score	Accuracy
VGG16	96%	96%	96%	96%
InceptionResNetV2	97.5%	96.5%	97%	98%
ResNet50	89%	89%	89%	89%
EfficientNetB2	68.5%	68.5%	68%	68%
Xception	96.5%	97%	97%	97%
Swin Transformer	97.82%	97.20%	97.47%	98.3%
Hybrid: Xception+Swin Transformer	**98%**	**98%**	**98%**	**98.78%**
Hybrid: InceptionResNetV2+Swin Transformer	97.90%	97.50%	97.70%	98.50%

4.3. Performance of the Hybrid Model

Multiple fusion strategies were examined to determine the most effective feature integration method, as represented in Table 4. Simple concatenation and averaging resulted in competitive but restricted performance caused by the lack of feature adaptability. Hierarchical fusion and attention-based approaches made small improvements by focusing on the most important representations. Concatenation + CBAM had the best performance of the group, with an accuracy of 99.39%, which shows that spatial–channel recalibration works. The suggested Gated Fusion + CBAM mechanism had the best results: 99.87% precision, 99.92% recall, 99.90% F1-

score, and 99.93% accuracy. This shows that it is better at getting rid of unnecessary information and bringing attention to clinically important lesion cues. This enhancement is especially beneficial for borderline instances where the margins of lesions are visually unclear.

Table 4. Experiment results on different fusion techniques

Fusion Technique	Precision	Recall	F1-Score	Accuracy
Concatenation	98.00%	98.00%	98.00%	98.78%
Average	97.79%	98.10%	97.95%	98.37%
Hierarchical / Multi-Level Fusion	98.40%	98.30%	98.47%	98.83%
Concat + Channel Attention	98.20%	98.50%	98.34%	98.81%
Concat + Spatial Attention	98.40%	98.80%	98.73%	98.90%
Concat + CBAM	99.23%	99.38%	99.30%	99.39%
Proposed: Gated Fusion + CBAM	**99.87%**	**99.92%**	**99.90%**	**99.93%**

We also analyzed how each of the techniques for fusing the information from the two models with respect to the model parameters, the amount of memory required by the GPU to perform the computations, and the training times, as can be seen in Table 5. The simple concatenation technique had a relatively small memory and computation footprint; however, the simple concatenation technique did not support adaptive learning and was less accurate than the other techniques. The attention-based fusion technique resulted in increased computational overhead; however, the results showed significant performance gains. The proposed technique is the most computationally expensive (50-100 K+ model parameters) and requires the longest training time; however, the improved accuracy provides justification for the additional costs associated with this technique in a diagnostic environment where reliability is greater than speed.

Table 5. Summary of feature fusion and selection strategies

Fusion Method	Extra Parameters	Memory Usage	Training Time	Trade-Off Summary
Concatenation	~0	Low	Fastest	Simple, no learning in fusion, good baseline, but lower adaptiveness.
Average Fusion	~0	Low	Fast	Very cheap, loses feature diversity, and slightly lower performance.
Hierarchical / Multi-Level Fusion	+0.2–0.4M	Medium	Moderate	Better feature utilization across levels, improved robustness & accuracy.
Concat + Channel Attention (SE)	+5K–20K	Medium	Moderate	Learns per-channel importance, small overhead, and improves discriminability.
Concat + Spatial Attention	+10K–30K	Medium–High	Moderate–Slow	Adds spatial focus, effective for lesion localization, with a higher computational cost.
Concat + CBAM	+20K–40K	Medium–High	Slower	Best among non-gated methods, improves both spatial & channel selectivity.
Proposed: Gated Fusion + CBAM	**+50K–100K**	**High**	**Slowest**	**Most adaptive fusion, best performance, higher computational cost, but worth it for accuracy.**

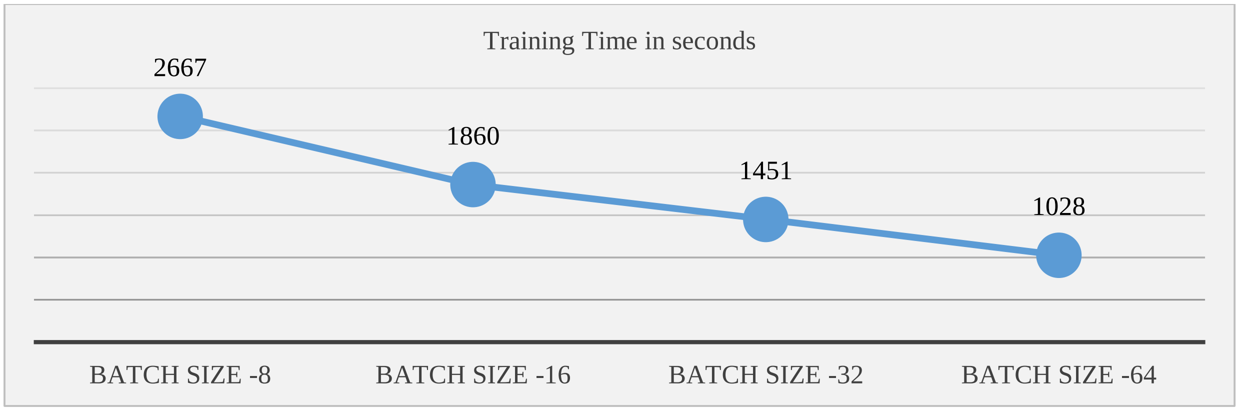

Fig. 2 Comparison of different batch sizes for the proposed method

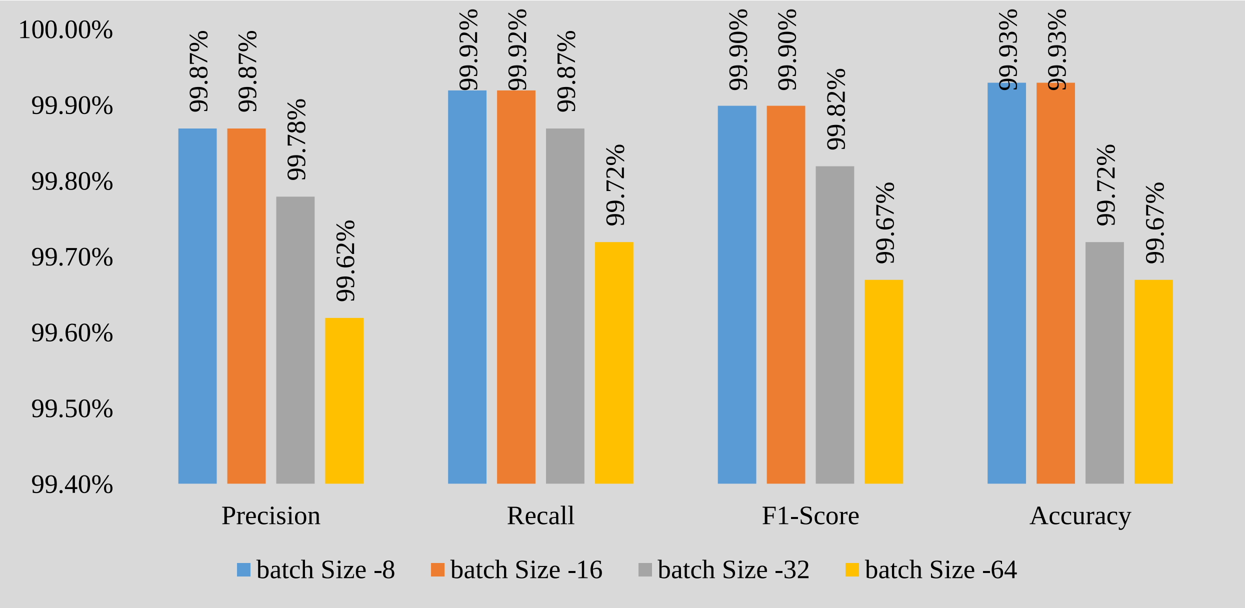

Fig. 3 Comparison of different batch sizes for the proposed method

We experimentally evaluated the proposed fusion technique using four different batch sizes- 8, 16, 32, and 64 to measure their effect on model accuracy, precision, recall, and F1-score. We found that small batch sizes, e.g., 8, gave higher accuracy at higher time/costs, and larger batch sizes of 32 & 64 reduced the costs at lower performance metrics due to averaging effects. The medium batch size of 16 struck a balance between computation efficiency and performance, demonstrating steady convergence and balanced results on all metrics. The training time and performance comparison are presented in Figures 2 and 3, respectively.

Table 6 provides a summary of the performance comparison between the suggested method and a number of recently published liver lesion categorization techniques. With accuracies ranging from 96 to 100%, traditional segmentation-dependent pipelines [30, 35] use CNN-based classifiers or manual or U-Net-based liver segmentation, followed by traditional machine learning. These methods, however, have a high pre-processing overhead, rely heavily on expert annotations, and have limited scalability for clinical use in the real world. More sophisticated segmentation–classification hybrid frameworks [32-34] have reported performance gains of roughly 92–99.25%; however, segmentation accuracy, single-stream feature extraction, or the lack of adaptive feature fusion mechanisms continue to limit their efficacy. An accuracy of 98.79% was attained by a segmentation-free deep learning method [31] that used transformer-based feature extraction in conjunction with GAN-driven data augmentation. In contrast, the proposed segmentation-free hybrid architecture integrating Swin Transformer and Xception backbones with gated fusion and CBAM achieved superior performance, with an accuracy of 99.93%, precision of 99.87% and recall of 99.92%, outperforming all existing methods. These results demonstrate that adaptive multi-branch feature fusion with attention mechanisms not only removes segmentation dependency but also yields more robust and clinically reliable liver lesion classification.

By comparing the results from this paper with some recent literature, it is clear that the advantages of using segmentation-free classification frameworks far outweigh the advantages of segmentation-dependent frameworks. Segmentation-dependent frameworks can deliver good results, but they also require extra work in terms of annotating data and, therefore, create more opportunities for errors. This research demonstrates that it is possible to achieve a high degree of accurate classification without segmenting images first, which should make it easier to implement these types of algorithms on a larger scale in clinical settings. In addition, the level of accuracy achieved by this algorithm is higher than many other segmentation-based and segmentation-free algorithms reported in the literature, and as such, suggests that adaptive hybrid fusion is an effective strategy for integrating information from multiple sources.

Another important practical observation was made concerning the selection of training parameters. When the impact of various batch size configurations was examined, the use of a moderate-sized batch resulted in a reasonable tradeoff between model training speed and model ability to generalize. Batch sizes that were too small increased the amount of time required to train the model, while very large batch sizes reduced gradient diversity and slightly decreased performance. These results should be useful to those who wish to develop image processing algorithms to run on equipment with limited resources.

Table 6. Comparison of existing methods and proposed approach for liver lesion classification on LiTs'17 dataset

Author	Method used	Performance
Khalid et al. (2022) [30]	Input: Segmented Liver CT image (Manual Segmentation) Feature Extraction: LiverNet, Classifier: SVM	Accuracy: 100%, Precision: 100%, Recall: 100%
Joshi et al. (Sep 2025) [31]	Segmentation Free Classification CustomLiverNet: GAN for Augmentation + Residual Networks and Vision Transformer models as Feature Extractor + Customized Fusion Layer	Accuracy: 98.79%, Precision: 98.64%, Recall: 98.58%
Archana R et al. (2025) [32]	Segmentation: Multi-Scale Cascaded Spatial Segmentation Transformer (M-SCSST), Classification: Global Average + Dense (Softmax)	Accuracy: 98.1%, Precision: 95.9%, Recall: 96.9%
Mei G et al. (2025) [33]	AM-Unet Model for segmentation and CNN for Classification	Accuracy: 92%, Precision: 94%, Recall: 95%
Srinivas Kolli et al. (2024) [34]	Segmentation: Improved meta-heuristic technique, Classification: PNN with Bayesian Optimization	Accuracy: 99.25%, Recall: 98.63%
Aparna P R et al. (2023) [35]	Segmentation: Modified Dense U-Net, Classification: Novel Deep CNN with Pre-Trained VGG16	Accuracy: 96%, Precision: 95.80%, Recall: 95.80%
Proposed Approach	**Hybrid: Swin Tranformer + Xception + Gated Fusion + CBAM**	**Accuracy: 99.93%, Precision: 99.87%, Recall: 99.92%**

5. Conclusion

This study assessed seven hybrid feature fusion methods for liver lesion segmentation-free classification based upon abdominal CT image data from the LiTS-17 dataset. In comparison to the performance of the individual networks, the hybrid fusion method combining the strengths of CNNs and Transformers demonstrated improved classification accuracy, due to their ability to provide complementary architecture strengths. The strategy of gating feature fusion with both channel and spatial attention was the most successful of all tested strategies, demonstrating the capability to identify benign and malignant lesions while eliminating the need for explicit segmentation of the lesion. Additionally, this study provided evidence that balancing batch size provides stable performance and computational efficiency during training.

Considering the clinical and health care policy ramifications of the proposed segmentation-free framework, it has the potential to provide a high degree of assistance to radiologists in clinical environments. Automated lesion classification systems have the potential to provide radiologists with accurate second opinions, to alleviate the workload associated with the diagnosis of liver lesions, and to decrease the inter-radiologist variability in lesion interpretation, which are significant issues in low-resource health care environments. By decreasing the reliance of such systems on the requirement for segmentation annotations, the computational resources required for processing large volumes of images can be greatly decreased, allowing for wider-scale implementation within hospitals that lack specialized imaging expertise. These types of systems have the potential to enhance the effectiveness of national cancer screening programs by enhancing the triage and prioritization of suspicious cases for earlier detection and more effective use of health care resources.

There are several limitations that must be addressed regarding the findings of this study. Although the LiTS-17 dataset is one of the most commonly used datasets for testing liver lesion classification algorithms, it does not contain all possible variations in scanners, acquisition protocols, and patient populations that are typically found in clinical environments. The classification algorithm implemented in this study only classified liver lesions into two categories (benign vs. malignant), when liver lesion classification in clinical practice is often performed at multiple levels. All experiments were performed using a 2D slice-based approach, which likely did not capture the full extent of the volumetric context of the CT scan images. While the segmentation-free pipeline decreases the amount of annotation required for model training, the performance of models trained with these pipelines may still be impacted by imaging artifacts and/or slices that do not clearly display the presence of liver lesions.

Future studies will continue to examine multi-class liver lesion classification, where each class includes an increasing number of lesion categories to increase the correspondence between the classification algorithm output and clinical practice. Future studies may extend the current framework to include 3D volumetric analysis, which would potentially allow the classification algorithm to incorporate additional morphological features of liver lesions. Future studies will require cross-institutional validation of the proposed pipeline across a variety of scanner manufacturers and scanning protocols to assess its generalizability to real-world clinical environments. Additional future research may focus on developing lightweight or efficient architectures to support the clinical integration of these models in real time, as well as developing explainable AI techniques to improve the confidence of clinicians in the reliability of automated diagnosis systems. In conclusion, the proposed hybrid framework offers a promising direction toward the development of reliable and scalable automated liver lesion classification systems to aid clinicians in medical imaging decision-making support applications.

Data Availability

The proposed word used publicly available data set LiTS17.(https://academictorrents.com/details/27772adef6f563a1ecc0ae19a528b956e6c803ce).

References

[1] Jiarong Zhou et al., "Automatic Detection and Classification of Focal Liver Lesions based on Deep Convolutional Neural Networks: A Preliminary Study," *Frontiers in Oncology*, vol. 10, pp. 1-11, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] Francisco Perdigón Romero et al., "End-To-End Discriminative Deep Network for Liver Lesion Classification," *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, Venice, Italy, pp. 1243-4246, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[3] Maayan Frid-Adar et al., "GAN-based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification," *Neurocomputing*, vol. 321, pp. 321-331, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[4] Mubasher Hussain, Najia Saher, and Salman Qadri, "Computer Vision Approach for Liver Tumor Classification Using CT Dataset," *Applied Artificial Intelligence*, vol. 36, no. 1, pp. 1-24, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[5] Ayesha Amir Siddiqi, Attaullah Khawaja, and Adnan Hashmi, "Classification of Abdominal CT Images bearing Liver Tumor Using Structural Similarity Index and Support Vector Machine," *Mehran University Research Journal of Engineering and Technology*, vol. 39, no. 4, pp. 751-758, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[6] Su-E Cao et al., "Multiphase Convolutional Dense Network for the Classification of Focal Liver Lesions on Dynamic Contrast-Enhanced Computed Tomography," *World Journal of Gastroenterology*, vol. 26, no. 25, pp. 3660-3672, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[7] Yuankai Huo et al., "Harvesting, Detecting, and Characterizing Liver Lesions from Large-scale Multi-phase CT Data via Deep Dynamic Texture Learning," *arxiv Preprint*, pp. 1-10, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[8] Sang-Gil Lee et al., "Robust End-to-End Focal Liver Lesion Detection Using Unregistered Multiphase Computed Tomography Images," *IEEE Access*, vol. 7, no. 2, pp. 319-329, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[9] Ling Zhao et al., "A Unified End-to-End Classification Model for Focal Liver Lesions," *Biomedical Signal Processing and Control*, vol. 86, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Lei Wu et al., "Beyond Radiologist-level Liver Lesion Detection on Multi-Phase Contrast-Enhanced CT Images by Deep Learning," *iScience*, vol. 26, no. 11, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Zhehan Shen et al., "An Explainable Deep Learning Model for Focal Liver Lesion Diagnosis Using Multiparametric MRI," *Radiology: Artificial Intelligence*, vol. 7, no. 6, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[12] Robert Stollmayer et al., "Diagnosis of Focal Liver Lesions with Deep Learning-based Multi-channel Analysis of Hepatocyte-Specific Contrast-Enhanced Magnetic Resonance Imaging," *World Journal of Gastroenterology*, vol. 27, no. 35, pp. 5978-5988, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[13] Anh-Cang Phan et al., "Improving Liver Lesions Classification on CT/MRI Images Based on Hounsfield Units Attenuation and Deep Learning," *Gene Expression Patterns*, vol. 47, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Hansang Lee et al., "Classification of Focal Liver Lesions in CT Images using Convolutional Neural Networks with Lesion Information Augmented Patches and Synthetic Data Augmentation," *Medical Physics*, vol. 48, no. 9, pp. 5029-5046, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[15] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan, "Explainable Deep Learning Models in Medical Image Analysis," *Journal of Imaging*, vol. 6, no. 6, pp. 1-19, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[16] Fei Lyu et al., "Weakly Supervised Liver Tumor Segmentation using Couinaud Segment Annotation," *IEEE Transactions on Medical Imaging*, vol. 41, no. 5, pp. 1138-1149, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[17] Chiu Sung-Hua et al., "Binary Classification of Benign and Malignant Hepatic Lesions with Portal Venous Phase Computed Tomography Images with Deep Learning: A Single-institution Study," *Journal of Medical Sciences*, vol. 45, no. 2, pp. 33-37, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[18] Shaohua Qiao et al., "Four-phase CT Lesion Recognition Based on Multi-phase Information Fusion Framework and Spatiotemporal Prediction Module," *BioMedical Engineering OnLine*, vol. 23, pp. 1-18, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Patrick Bilic et al., “The Liver Tumor Segmentation Benchmark (LiTS),” *Medical Image Analysis*, vol. 84, pp. 1-24, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Hyuna Sung et al., “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209-249, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[21] Alejandro Forner, María Reig, and Jordi Bruix, “Hepatocellular Carcinoma,” *The Lancet*, vol. 391, pp. 1301-1314, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[22] Khaled Y. Elbanna, and Ania Z. Kielar, “Computed Tomography Versus Magnetic Resonance Imaging for Hepatic Lesion Characterization/ Diagnosis,” *Clinical Liver Disease*, vol. 17, no. 3, 159-164, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[23] Wenya Linda Bi et al., “Artificial Intelligence in Cancer Imaging: Clinical Challenges and Applications,” *CA: A Cancer Journal for Clinicians*, vol. 69, no. 2, pp. 127-157, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[24] Koichiro Yasaka, and Osamu Abe, “Deep Learning and Artificial Intelligence in Radiology: Current Applications and Future Directions,” *PLOS Medicine*, vol. 15, no. 11, pp. 1-14, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[25] Hayit Greenspan, Bram van Ginneken, and Ronald M. Summers, “Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153-1159, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[26] Geert Litjens et al., “A Survey on Deep Learning in Medical Image Analysis,” *Medical Image Analysis*, vol. 42, pp. 60-88, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[27] Patrick Ferdinand Christ et al., “Automatic Liver and Tumor Segmentation of CT and MRI Volumes using Cascaded Fully Convolutional Neural Networks,” *arXiv Preprint*, pp. 1-20, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[28] Kaiming He et al., “Deep Residual Learning for Image Recognition,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770-77, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[29] Christian Szegedy et al., “Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning,” *Thirty-First Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, pp. 4278-4284, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[30] Khaled Alawneh et al., “LiverNet: Diagnosis of Liver Tumors in Human CT Images,” *Applied Sciences*, vol. 12, no. 11, pp. 1-16, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[31] Shivani Joshi et al., “Enhancing Liver Cancer Detection: An Innovative Deep Learning Approach Combining GAN, ResNet, and Vision Transformer,” *Expert Systems with Applications*, vol. 298, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[32] R. Archana, and L. Anand, “Multi-Scale Cascaded Spatial Segmentation Transformer for Liver Cancer Classification,” *International Journal of Computational Intelligence Systems*, pp. 1-32, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[33] Guang Mei et al., “Research on CT Image Segmentation and Classification of Liver Tumors based on Attention Mechanism and Improved U-Net Model,” *Technology and Health Care*, vol. 33, no. 5, pp. 2468-2483, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[34] Srinivas Kolli et al., “A Novel Liver Tumor Classification using Improved Probabilistic Neural Networks with Bayesian Optimization,” *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 8, pp. 1-9, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[35] P.R. Aparna, and T.M. Libish, “Automatic Segmentation and Classification of the Liver Tumor using Deep Learning Algorithms,” *2023 3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, Kalady, Ernakulam, India, pp. 334-339, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[36] François Chollet, “Xception: Deep Learning with Depthwise Separable Convolutions,” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 1800-1807, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[37] Reza Azad et al., “Advances in Medical Image Analysis with Vision Transformers: A Comprehensive Review,” *Medical Image Analysis*, vol. 91, pp. 1-72, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[38] Xiayu Guo et al., “UCTNet: Uncertainty-Guided CNN-Transformer Hybrid Networks for Medical Image Segmentation,” *Pattern Recognition*, vol. 152, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[39] Aymen M. Al-Hejri et al., “A Hybrid Vision Transformer with Ensemble CNN Framework for Cervical Cancer Diagnosis,” *BMC Medical Informatics and Decision Making*, vol. 25, pp. 1-20, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[40] Xiaolei He et al., “A Novel Liver Image Classification Network for Accurate Diagnosis of Liver Diseases,” *Scientific Reports*, vol. 15, pp. 1-16, 2025. [CrossRef] [Google Scholar] [Publisher Link]