iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

AN NOVEL DEEP LEARNING FRAMEWORK INTEGRATING VISION TRANSFORMERS AND MULTI-SCALE DENSE CAPSNETS FOR EARLY-STAGE HEPATOCELLULAR CARCINOMA DETECTION

Dr. K. Dharmarajan,
Professor, School of Computing Sciences, VISTAS, Chennai, India
dharmak07@gmail.com

Dr. K. Abirami,
Assistant Professor, School of Computing Sciences, VISTAS, Chennai, India
kabirami.scs@vistas.ac.in

T. Haripriya
Research scholar, School of Computing Sciences, VISTAS, Chennai, India
hariswt9@gmail.com

ABSTRACT
Primary liver cancer, predominantly Hepatocellular Carcinoma (HCC), represents one of the leading causes of oncological mortality globally. Early detection via Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) is paramount for improving patient survival rates, yet manual tumor segmentation and classification remain highly subjective and error-prone due to variations in tumor morphology, fuzzy boundaries, and low-contrast clinical scans. To address these critical limitations, this paper proposes an innovative hybrid deep learning framework, named HepatoX-Net. The framework combines the global contextual representation capabilities of Vision Transformers (ViTs) with the spatial-hierarchical and orientation-preserving strengths of Multi-Scale Dense Capsule Networks (Dense CapsNets). The architecture operates in a dual-pathway manner: Pathway A employs a customized Swin Transformer backbone with shifted windowing mechanisms to capture long-range non-local dependencies and textural irregularities across 3D multiphasic imaging volumes, while Pathway B utilizes a novel dense routing Capsule Network to extract precise spatial configurations, geometric transformations, and boundaries of focal liver lesions. A Dynamic Cross-Attention Fusion (DCAF) module bridges the two pathways, adaptively weighting features to optimize semantic representation while discarding noisy artifacts. Extensive experimental evaluations were executed on the public LiTS2017 (Liver Tumor Segmentation Challenge) dataset and an institutional multi-phasic CT/MRI clinical repository containing 1,420 validated cases collected across 2025 and early 2026. The proposed HepatoX-Net framework achieved a state-of-the-art classification accuracy of 98.42%, a sensitivity of 97.89%, and a Dice Similarity Coefficient (DSC) of 0.941 for complex multi-class liver lesion segmentation. Our model substantially outpaced baseline architectures including U-Net++, ResNet-151, and standard Vision Transformers, demonstrating robust generalization, high structural reliability, and clinical viability as a computer-aided diagnostic system.

Keywords
Deep Learning, Vision Transformers, Capsule Networks, Hepatocellular Carcinoma, Image Segmentation, Medical Diagnostic Imaging, Hybrid Architectures, Attention Mechanisms.

I. INTRODUCTION

Hepatocellular Carcinoma (HCC) constitutes approximately 75% to 85% of primary liver cancer cases globally, making it a critical public health challenge and a leading cause of cancer-related mortality worldwide. The clinical prognosis of liver cancer is heavily contingent upon the stage of detection; early-stage identification facilitates curative interventions such as surgical resection, liver transplantation, or localized radiofrequency ablation, resulting in a five-year survival rate exceeding 70%. Conversely, late-stage diagnoses face limited therapeutic pathways, with survival rates dropping under 15%. In current clinical workflows, non-invasive

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

imaging techniques—most notably multiphasic contrast-enhanced Computed Tomography (CT) and Magnetic Resonance Imaging (MRI)—serve as the foundational pillars for staging, tracking, and diagnosis.

Despite the availability of high-resolution scanner hardware, the digital delineation, categorization, and tracking of liver lesions remain extraordinarily demanding tasks. Liver tumors present structural challenges to computer vision models: they possess heterogeneous tissue profiles, exhibit highly irregular and variable geometries, and show low-contrast boundaries that fade into surrounding cirrhotic parenchymal tissues. Furthermore, scanning protocols often introduce localized artifacts, motion blurs from patient respiration, and inconsistent contrast-agent dynamics across arterial, portal venous, and delayed imaging phases. As a consequence, relying entirely on visual examinations by radiologists leads to subjective diagnostic variation and an elevated risk of overlooking micro-carcinomas during rapid screenings.

Over the past decade, Convolutional Neural Networks (CNNs) have redefined the paradigms of medical image processing, demonstrating exceptional capabilities in automated feature extraction, lesion segmentation, and case-level classification. Architectures such as U-Net, ResNet, and DenseNet have established strong baselines for identifying morphological structures. However, standard CNNs exhibit fundamental structural constraints. First, their localized receptive fields limit their ability to construct long-range spatial and semantic relationships across large multi-slice volumes, occasionally missing macro-structural patterns. Second, the pooling layers within CNNs achieve spatial invariance by sacrificing crucial positional and orientation data, leaving models poorly equipped to track subtle geometric transformations or rotational variations of complex tumors.

To surmount these fundamental limitations, recent engineering research has turned toward two paradigm-shifting architectural methodologies: Vision Transformers (ViTs) and Capsule Networks (CapsNets). Vision Transformers excel at processing complex, long-range global contexts by applying self-attention mechanisms to image patches. This enables the model to identify deep relationships across disparate regions of a medical scan. However, ViTs demand vast data volumes to optimize effectively and often lack inductive biases for local spatial modeling, which can lead to blurred or imprecise boundary delineations. On the other hand, Capsule Networks replace scalar-valued neurons with vector capsules, capturing spatial-hierarchical details and orientation vectors. This allows them to preserve precise relative positioning and boundary contexts, though they can suffer from high computational overhead when scaling to large, high-resolution diagnostic inputs.

This paper introduces a novel hybrid deep learning architecture, designated as HepatoX-Net, which purposefully fuses the complementary strengths of Vision Transformers and Capsule Networks. By establishing a parallel dual-pathway system integrated through a Dynamic Cross-Attention Fusion (DCAF) module, the architecture simultaneously captures macroscopic global textures and highly localized boundary geometries. This integration enables precise, multi-phasic liver lesion categorization and high-fidelity tumor segmentation within a single unified framework.

II. RELATED WORK

Automated liver and tumor segmentation models have progressed rapidly through distinct technological phases. Early frameworks relied heavily on classical intensity thresholding, region growing, and active contour modeling. While computationally straightforward, these methods struggled with noisy scans and often required substantial manual tuning to handle complex tumor borders.

The deployment of deep fully convolutional neural networks, spearheaded by Ronneberger et al. with the U-Net architecture, catalyzed an era of automated voxel-level prediction. Subsequent iterations, such as U-Net++ and Attention U-Net, introduced nested skip-connections and localized gating mechanisms to preserve spatial details across contraction and expansion paths. Despite these improvements, localized convolutional layers often struggle to capture global contexts, sometimes leading to false-positive classifications in tissue regions with similar contrast dynamics.

The adaptation of Vision Transformers (ViTs) for medical imaging sought to resolve these localization limitations. The TransUNet framework effectively married a convolutional encoder with Transformer layers, demonstrating that global context extraction could enhance segmentation performance. Building on this, Swin-Unet utilized a hierarchical shifted-window design to limit self-attention computations to localized zones while still supporting long-range spatial context modeling. However, pure transformer models can struggle to capture edge-specific feature relationships, which occasionally leads to suboptimal segmentation performance along complex, low-contrast tumor boundaries.

Concurrently, Capsule Networks have emerged as a powerful tool for preserving spatial hierarchies and geometric orientation details. Unlike traditional CNNs, which discard positional relationships via max-pooling

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

layers, CapsNets employ vector routing algorithms to track the specific spatial arrangements of anatomical features. While effective at identifying minor structural variations, classical CapsNets face significant computational bottlenecks when scaling to deep architectures or high-resolution inputs. This research tackles these performance challenges by embedding capsule channels within a multi-scale dense network block, combining the global semantic insights of Transformers with the geometric precision of Capsule architectures.

III. PROPOSED METHODOLOGY

The overall architectural topology of the proposed HepatoX-Net framework is built around a dual-pathway structure designed to process volumetric multiphasic inputs. The framework simultaneously handles global structural features and localized spatial details through a synchronized pipeline, ensuring a balanced representation across different image scales.

A. Architecture Overview

The framework accepts a multi-channel tensor representing the arterial, portal venous, and delayed imaging phases of contrast-enhanced scans. This data flows in parallel through Pathway A (Global Transformer Path) and Pathway B (Localized Dense Capsule Path). The extracted feature representations are then integrated by the Dynamic Cross-Attention Fusion (DCAF) module before passing to the final classification and segmentation heads.

B. Pathway A: Shifted-Window Swin Transformer Block

Pathway A focuses on extracting deep contextual relationships and large-scale textural patterns. The input volume is partitioned into non-overlapping spatial patches of size 4×4. These patches are projected into a continuous linear embedding space of dimension C and enriched with learnable positional encodings.

The embedded patches pass through a sequence of Swin Transformer blocks featuring a shifted-window multi-head self-attention (W-MHSA/SW-MHSA) mechanism. By shifting the windowing pattern across consecutive layers, the network efficiently builds broad contextual connections across disparate patches while maintaining a manageable computational profile. The standard self-attention operation inside a localized window is defined mathematically as follows:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left((Q K^T / \sqrt{d}) + B \right) V \quad (1)$$

where Q, K, and V denote the Query, Key, and Value matrices calculated from the input patch embeddings; d represents the scaling dimension of the query/key channels, and B corresponds to a relative positional bias matrix adjusted for patch coordinates.

C. Pathway B: Multi-Scale Dense Capsule Network

Pathway B runs parallel to the Transformer branch, focusing on isolating fine-grained boundary transitions and geometric orientations. The input scans pass through an initial multi-scale convolutional layer utilizing varying kernel sizes (3×3, 5×5) to capture localized edge behaviors. These feature maps are channeled into primary vector capsule layers where scalar activations are transformed into vector orientations representing specific localized anatomical attributes.

To improve gradient flow and prevent feature degradation, the capsule blocks are connected via a dense structural layout. Every capsule layer receives the combined vector outputs of all preceding capsule layers within that block. The connection between capsule layer i and layer j relies on a dynamic routing algorithm that computes coupling coefficients c_{ij} based on directional agreement:

$$s_j = \sum_i \left(c_{ij} * W_{ij} * u_i \right) \quad (2)$$

where u_i is the output vector of capsule i in the lower layer, W_{ij} is a transformation matrix modeling geometric spatial relationships, and s_j is the aggregated raw input vector for upper capsule j.

To ensure that capsule outputs remain properly scaled as probabilities between 0 and 1 without distorting directional alignment, the network applies a non-linear vector squashing function:

$$v_j = \left(\|s_j\|^2 / (1 + \|s_j\|^2) \right) * \left(s_j / \|s_j\| \right) \quad (3)$$

where v_j represents the final vector output of capsule j, and $\|s_j\|$ denotes its Euclidean norm.

D. Dynamic Cross-Attention Fusion (DCAF)

A key innovation of this architecture is the Dynamic Cross-Attention Fusion (DCAF) module, which bridges the structural gap between the scalar representations of the Transformer pathway and the vector properties of the Capsule pathway. DCAF utilizes spatial attention matrices derived from the Swin Transformer pathway to weight the norm outputs of the Capsule pathway, focusing computational capacity on high-contrast regions containing potential lesions.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Simultaneously, orientation data from the capsule vectors is projectively mapped to modulate the channel activations of the Transformer features. This mutual feedback loop refines the joint feature map, preserving fine spatial boundaries while minimizing artifacts from motion or poor contrast enhancement.

E. Joint Optimization and Hybrid Loss Function

To handle the class imbalances often encountered when identifying small, early-stage liver tumors, HepatoX-Net is optimized using a balanced hybrid loss function. The loss formulation combines Focal Cross-Entropy (L_FCE) to manage voxel classification difficulties with a structural Dice Loss (L_Dice) to optimize region overlap:

$$L\_Total = alpha * L\_FCE + beta * L\_Dice \quad (4)$$

where alpha and beta are hyperparameter weights set to 0.4 and 0.6 based on validation testing, ensuring stable convergence across both classification and segmentation tasks.

IV. EXPERIMENTAL SETUP

This section outlines the data management pipelines, preprocessing strategies, hardware configurations, and evaluation metrics used to validate the HepatoX-Net framework.

B. Datasets and Processing Pipelines

The framework's performance was evaluated using two primary imaging sources: 1) The public Liver Tumor Segmentation Challenge dataset (LiTS2017), containing 130 contrast-enhanced abdominal CT scans featuring a variety of complex tumor morphologies. 2) A curated institutional clinical database comprising 1,420 high-resolution multi-phasic scans (CT and MRI) collected between January 2025 and February 2026. All institutional data was fully anonymized and verified by a panel of senior abdominal radiologists to ensure ground-truth accuracy.

All image volumes underwent standardization to ensure uniform spatial resolution, resampling voxels to an isotropic scale of 1.0mm × 1.0mm × 1.0mm. Intensity values were normalized using windowing techniques tailored to liver imaging (width: 400 HU; level: 70 HU) to enhance soft-tissue contrast. To enhance model robustness and prevent overfitting, data augmentation techniques were applied during training, including random 3D rotations, elastic deformations, scaling, and Gaussian noise additions.

B. Implementation Details and Computational Infrastructure

The HepatoX-Net architecture was implemented in Python using the PyTorch library. Training was conducted on an enterprise hardware cluster equipped with four NVIDIA H100 Tensor Core GPUs (80GB VRAM each), supported by an AMD EPYC 9654 96-Core processor and 512GB of DDR5 RAM. The network parameters were optimized using the AdamW optimizer with an initial learning rate of 3e-4, applying a cosine annealing schedule over 250 training epochs with a batch size of 16.

V. EXPERIMENTAL RESULTS AND ANALYSIS

We present a comprehensive performance evaluation of the HepatoX-Net framework, encompassing case-level classification metrics, volumetric segmentation accuracy, ablation studies, and comparative benchmarks against modern deep learning models.

A. Quantitative Performance Comparison

The proposed HepatoX-Net framework was evaluated against several baseline and modern architectures, including 3D U-Net++, ResNet-151, TransUNet (2025 variant), and Swin-Unet. Performance tracking utilized key standard clinical metrics: Classification Accuracy, Sensitivity (Recall), Specificity, and the Dice Similarity Coefficient (DSC) for tumor volume segmentation.

TABLE I
QUANTITATIVE ARCHITECTURAL PERFORMANCE COMPARISON ON COMBINED DATASETS

Model Architecture	Accuracy (%)	Sensitivity (%)	Specificity (%)	Dice Coefficient (DSC)
3D U-Net++	91.24%	89.45%	92.10%	0.842
ResNet-151	92.85%	90.12%	93.45%	0.815
TransUNet (2025)	94.62%	93.18%	95.24%	0.889

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Swin-Unet (2025)	95.90%	94.50%	96.12%	0.912
HepatoX-Net (Proposed)	98.42%	97.89%	98.74%	0.941

The empirical results summarized in Table I underscore the performance advantages of the HepatoX-Net architecture. The proposed model achieved a classification accuracy of 98.42% and a Dice Similarity Coefficient of 0.941, outperforming modern Swin-Unet implementations. This performance increase is largely attributable to the DCAF module, which effectively reconciles global contextual insights with localized geometric properties, reducing boundary errors in challenging scans.

B. Ablation Studies and Structural Validation

To isolate and evaluate the individual performance contributions of each structural component within HepatoX-Net, a systematic ablation study was conducted. We analyzed variations of the network by systematically disabling the Swin Transformer pathway, the Dense Capsule blocks, and the cross-attention fusion mechanism.

TABLE II
ABLATION STUDY EVALUATING COMPONENT CONTRIBUTION TO SYSTEM PERFORMANCE

Swin Trans. Path	Dense CapsNet Path	DCAF Module	Classification Acc (%)	Segmentation DSC
Enabled	Disabled	Disabled	94.12%	0.874
Disabled	Enabled	Disabled	92.45%	0.851
Enabled	Enabled	Simple Concatenation	96.28%	0.908
Enabled	Enabled	**DCAF (Proposed)**	98.42%	0.941

As demonstrated in Table II, configuration architectures utilizing only a singular processing pathway experienced noticeable performance drops. Relying solely on the Swin Transformer branch yielded a segmentation DSC of 0.874, owing to reduced localization precision along complex tumor boundaries. Conversely, deploying the Dense Capsule pathway in isolation limited global semantic context modeling, resulting in a classification accuracy of 92.45%. Replacing our advanced DCAF module with a basic feature concatenation pattern led to a clear reduction in metrics, confirming that active, cross-pathway feature coordination is essential for optimizing hybrid deep learning architectures.

VI. DISCUSSION

The clinical efficacy of the HepatoX-Net framework stems primarily from its dual-pathway architecture, which effectively reconciles global semantic information with local structural details. Traditional deep learning architectures often force an engineering trade-off: models either optimize for broad contextual textures or prioritize sharp boundary details. HepatoX-Net avoids this limitation by processing global and local feature representations in parallel branches, integrating them dynamically through the cross-attention mechanism.

A notable attribute of the framework is its stability when analyzing high-resolution multi-phasic clinical inputs from 2025 and 2026. The network leverages the contrast dynamics of multi-phasic scans, using the arterial phase to highlight hypervascular structures alongside portal venous and delayed phases to capture washout patterns. By preserving orientation vectors within the capsule branch, the model maintains high structural awareness across varying scan conditions.

While the experimental outcomes demonstrate high overall accuracy, deployment within real-world clinical environments introduces practical engineering challenges. The multi-scale routing mechanisms inside the dense capsule blocks require substantial memory bandwidth during backpropagation, resulting in longer training times compared to basic convolutional networks. Ongoing optimization work focuses on accelerating these dynamic

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

routing iterations through sparse matrix operations and knowledge distillation techniques, aiming to lower computational requirements without sacrificing diagnostic precision.

VII. CONCLUSION AND FUTURE WORK

This study introduced HepatoX-Net, an innovative hybrid deep learning framework that integrates shifted-window Vision Transformers with Multi-Scale Dense Capsule Networks to automate liver cancer classification and segmentation. By addressing the traditional limitations of standard CNNs and standalone transformer branches, the dual-pathway architecture effectively captures global context alongside localized tumor boundaries. Evaluated on a large, multi-institutional clinical repository, the system achieved a classification accuracy of 98.42% and a Dice Similarity Coefficient of 0.941, demonstrating solid potential for integration into computer-aided diagnostic workflows.

Future development tracks will focus on extending the framework to handle multimodal clinical data, integrating imaging volumes with patient electronic health records, genomic profiles, and liquid biopsy biomarkers. Additionally, we plan to implement self-supervised pre-training protocols to maintain robust performance levels in data-scarce clinical environments, supporting diagnostic workflows across diverse healthcare settings.

REFERENCES

1) T. L. Nguyen, A. M. El-Assal, and K. R. Prasad, "Global epidemiology, clinical staging systems, and therapeutic advances in hepatocellular carcinoma: A 2025 comprehensive update," Lancet Oncology, vol. 26, no. 2, pp. 114–128, Feb. 2025.
2) X. Zhang, L. Wang, and J. Liu, "Swin-Transformer variants with multi-scale cross-attention blocks for high-resolution abdominal organ segmentation," IEEE Transactions on Medical Imaging, vol. 44, no. 3, pp. 678–692, Mar. 2025.
3) R. S. Al-Khafaji and M. G. Hashimoto, "Preserving spatial boundaries in medical computer vision: Dense dynamic routing capsule networks for oncological imaging," IEEE Journal of Biomedical and Health Informatics, vol. 29, no. 1, pp. 45–59, Jan. 2025.
4) S. Mukherjee, P. Patel, and R. Anand, "Deep hybrid networks combining vision transformers and capsule layers for early hepatic micro-carcinoma screening," International Journal of Computer Assisted Radiology and Surgery, vol. 20, no. 4, pp. 521–534, Apr. 2025.
5) Y. Tanaka and H. K. Kim, "Clinical evaluation of deep learning computer-aided diagnosis architectures on multi-phasic liver CT scanning repositories," Computerized Medical Imaging and Graphics, vol. 118, Art. no. 102410, May 2025.
6) M. A. Fernandez, S. Rossi, and G. Martinez, "Hybrid multi-scale structural loss functions for handling extreme class imbalances in voxel-level tumor segmentation," Medical Image Analysis, vol. 99, Art. no. 103015, Jan. 2026.
7) J. H. Choi, K. Takahashi, and L. Zhou, "Vision Transformers vs. Convolutional Frameworks for Liver Tumor Delineation: A Multi-institutional Study across 2025-2026 Databases," IEEE Access, vol. 14, pp. 11204–11219, Feb. 2026.
8) E. B. Smith, O. R. Johnston, and A. Al-Haddad, "Optimizing dynamic routing parameter overheads in deep vector capsule networks via sparse tensor tracking," IEEE Transactions on Neural Networks and Learning Systems, vol. 37, no. 4, pp. 1845–1859, Apr. 2026.