INTERNATIONAL E-CONFERENCE ON GLOBAL MULTIDISCIPLINARY RESEARCH AND INNOVATION (ICGMRI - 2026)

PROCEEDINGS

Pencil Bitz

ISBN 978-816800517-4

TABLE OF CONTENTS
ICGMRI - 26

Dynamic Swin-UNet with Adaptive Fusion for Precision HCC Delineation and Early Diagnosis

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

[2]Dr.K.Abirami, Assistant professor,
School of Computing Sciences,
VISTAS, Chennai, India
kabirami.scs@vistas.ac.in

[3]T Haripriya
Research Scholar
School of Computing Sciences,
VISTAS, Chennai, India
hariswt9@gmail.com

Abstract

Hepatocellular Carcinoma (HCC) and primary hepatic malignancies rank among the leading causes of cancer-related mortality globally. Accurate, early-stage diagnosis and automatic lesion segmentation from dynamic multiphase computed tomography (CT) and magnetic resonance imaging (MRI) are paramount for surgical planning, targeted local therapy, and treatment efficacy monitoring. However, computerized liver tumor segmentation remains an intricate challenge due to low tissue contrast, heterogeneous lesion morphology, ill-defined boundaries, microvascular invasion, and significant intra-organ structural variability. To address these critical clinical gaps, this paper presents a novel, highly innovative deep learning architecture named Multi-Phase Dynamic Swin-UNet Transformer with Adaptive Contrastive Fusion (MPD-SwinUNet-ACF). The proposed framework integrates hierarchical multi-scale Shifted Window (Swin) Transformer blocks within an encoder-decoder backbone to capture comprehensive long-range contextual spatial dependencies without incurring prohibitive computational complexity. Furthermore, an Adaptive Contrastive Dynamic Fusion (ACDF) module is engineered to synergistically fuse feature representations across arterial, portal venous, and delayed imaging phases, capturing micro-vascular dynamics essential for differential diagnosis. To refine fuzzy, ambiguous tumor margins, a specialized Boundary-Guided Gradient Refinement (BGGR) block incorporates high-frequency Sobel and Laplacian feature maps directly into the skip connections. Extensive experimental validations conducted on benchmark datasets—including the Liver Tumor Segmentation Challenge (LiTS 2017), 3DIRCADb, and an extended 2025–2026 multi-center clinical MRI/CT cohort—demonstrate that MPD-SwinUNet-ACF achieves state-of-the-art performance with a Dice Similarity Coefficient (DSC) of 97.42% for liver parenchyma and 92.18% for hepatic tumor segmentation, outperforming conventional U-Net, ResUNet++, MedNeXt, and hybrid CNN-Transformer baselines. Computational efficiency evaluations reveal an average inference latency of 41.2 milliseconds per 3D volume, verifying its suitability for real-time intraoperative navigation and clinical decision support.

Index Terms-Hepatocellular Carcinoma (HCC), Deep Learning, Swin Transformer, Multi-Phase CT/MRI Fusion, Medical Image Segmentation, Boundary Refinement, Explainable AI (XAI), Computer-Aided Diagnosis (CAD).

I. INTRODUCTION

Primary liver cancer, dominated by Hepatocellular Carcinoma (HCC) and intrahepatic cholangiocarcinoma (ICC), represents one of the most aggressive and lethal solid malignancies worldwide [1]. According to global epidemiological surveys published in 2025 and early 2026, liver cancer remains the third leading cause of cancer deaths, claiming over 800,000 lives annually [2]. The clinical prognosis of HCC is exceptionally time-sensitive; early detection of focal liver lesions significantly increases the five-year survival rate via curative interventions such as surgical resection, radiofrequency ablation (RFA), transarterial chemoembolization (TACE), or orthotopic liver transplantation [3]. Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) serve as the primary diagnostic modalities for evaluating liver tumor burden, anatomical location, and liver parenchyma condition [4].

In routine radiological workflows, multi-phase contrast-enhanced imaging is mandatory. Contrast agents pass through arterial, portal venous, and delayed phases, highlighting distinct hemodynamic patterns. For example, classic HCC manifests characteristic hypervascular enhancement in the arterial phase followed by contrast 'washout' in the portal venous and delayed phases [5]. However, manual visual inspection and slice-by-slice manual annotation of high-resolution 3D CT/MRI scans are labor-intensive, time-consuming, subjective, and highly susceptible to inter- and intra-observer variability. Consequently, fully automated, accurate, and robust deep learning-based diagnostic systems have emerged as a non-invasive imperative in modern digital healthcare and precision oncology [6].

Challenges in Clinical AI: Despite rapid developments in artificial intelligence for medical image analysis, automated liver lesion segmentation continues to face severe technical challenges:

1) Morphological Heterogeneity and Ill-Defined Boundaries: Hepatic tumors vary drastically in shape, spatial location, and texture across patients. Furthermore, tumor margins are frequently diffuse due to microvascular infiltration or surrounding hepatic steatosis, causing severe pixel ambiguity [7].

2) Multi-Phase Temporal Discrepancy and Alignment Noise: Standard single-phase networks fail to utilize the complementary physiological information presented across multiphase scans. Naive concatenation of multi-phase images often leads to feature redundancy or spatial misalignment artifacts [8].

3) Spatial Context vs. High-Resolution Detail Trade-off: Conventional Convolutional Neural Networks (CNNs) excel at extracting localized texture patterns but suffer from limited receptive fields, making it difficult to capture long-range global contextual relationships across large 3D abdominal volumes [9].

To overcome these limitations, this study introduces an innovative, end-to-end multi-phase deep architecture entitled Multi-Phase Dynamic Swin-UNet Transformer with Adaptive Contrastive Fusion (MPD-SwinUNet-ACF). Our approach bridges localized feature extraction with global contextual modeling by unifying Swin Transformer self-attention blocks with dynamic contrastive feature fusion and explicit boundary gradient refinement.

Core Contributions: The major scientific and technical contributions of this paper are summarized below:

• Novel Architecture Design: We engineer MPD-SwinUNet-ACF, a hybrid vision transformer network tailored specifically for dynamic multi-phase liver CT and MRI scans.

• Adaptive Contrastive Fusion (ACDF): We propose an adaptive feature fusion mechanism that dynamically weights multi-phase arterial, venous, and delayed representations to capture subtle arterial enhancement and washout kinetics.

• Boundary-Guided Gradient Refinement (BGGR): We introduce a dedicated edge-aware gradient block within skip connections to recover fine high-frequency boundary details, substantially improving Dice scores on small focal lesions.

• Comprehensive Benchmark & Multi-Center Validation: We validate our network on LiTS 2017, 3DIRCADb, and a multi-center dataset collected in 2025–2026, demonstrating superior accuracy, boundary fidelity, and inference computational efficiency.

II. RELATED WORK

A. CNN-Based Medical Image Segmentation

Deep learning revolutionized medical image analysis following the landmark introduction of U-Net by Ronneberger et al. [10]. Subsequent variants such as 3D U-Net, V-Net, and ResUNet++ integrated residual connections, dense skip connections, and squeeze-and-excitation attention mechanisms to improve abdominal organ segmentation [11]. In liver imaging, models like SHDLF [12] and Modif-SegUnet [13] demonstrated the efficacy of multi-stage cascaded pipelines where liver localization precedes tumor segmentation. However, purely CNN-based approaches rely heavily on local convolutional operations. This restricted receptive field hinders the network's capacity to capture long-range spatial context, frequently leading to false positive detections in complex multi-lesion cases [14].

B. Vision Transformers in Biomedical Imaging

To mitigate the receptive field restrictions of CNNs, Vision Transformers (ViTs) and Shifted Window Transformers (Swin-UNet) have been widely adopted [15]. Swin Transformer architectures compute self-attention within local, shifted non-overlapping windows, yielding linear computational complexity with respect to image resolution while preserving global context capture capabilities [16]. Recent hybrid architectures, such as ACF-TransUNet [17] and G-UNETR++ [18], attempt to join CNN local encoder blocks with transformer self-attention layers. Nevertheless, existing transformer frameworks are primarily optimized for single-phase images, leaving multi-phase temporal modeling under-explored.

C. Multi-Phase Dynamic Fusion & Boundary Modeling

Multi-phase contrast imaging provides rich pathophysiological insight into tumor vascularization. Recent studies in 2025 and 2026 have highlighted that multi-phase dynamic alignment is crucial for discriminating early-stage hepatocellular carcinoma from benign hemangiomas and hepatic cysts [19], [20]. Earlier multi-phase fusion methods relied on simple channel concatenation or element-wise addition, which ignore non-linear feature interactions between phases. Furthermore, boundary ambiguity remains a prominent challenge in deep learning segmentation pipelines [21]. Our proposed MPD-SwinUNet-ACF addresses both dynamic multi-phase fusion and precise boundary recovery within a unified deep learning framework.

TABLE I
COMPARATIVE SUMMARY OF STATE-OF-THE-ART LIVER TUMOR SEGMENTATION METHODS (2021–2026)

Model Name	Architecture Type	Multi-Phase Support	Boundary Module	LiTS DSC (%)	Inference Latency
3D U-Net [10]	3D Pure CNN	No (Single Phase)	None	67.5%	112 ms
ResUNet++ [11]	Residual CNN + SE	No	Attention Gate	74.8%	85 ms
SHDLF Framework [12]	ResNet50 + UNet70	No	Basic Masking	73.6%	94 ms
G-UNETR++ [18]	Graph + ViT Hybrid	Partial	Gradient-Enhanced	84.4%	68 ms
Proposed MPD-SwinUNet	Swin-Transformer + Hybrid	Yes (Adaptive Dynamic)	Boundary Gradient (BGGR)	92.18%	41.2 ms

III. PROPOSED METHODOLOGY

The structural workflow of the proposed MPD-SwinUNet-ACF framework is depicted conceptually in Figure 1 and consists of three main stages: (1) Multi-Phase Preprocessing and Rigid/Non-Rigid

Registration, (2) Dual-Stream Swin-Transformer Encoder with Adaptive Contrastive Dynamic Fusion (ACDF), and (3) Boundary-Guided Gradient Refinement (BGGR) Decoder with Multi-Task Loss Supervision.

A. Multi-Phase Input Preprocessing & Contrast Alignment

Multi-phase contrast CT/MRI protocols capture images at specific time intervals post-injection of contrast media: arterial phase (25–35 seconds), portal venous phase (60–70 seconds), and delayed phase (180–300 seconds). Let I_A, I_V, and I_D in R^(H x W x D) represent the 3D volumetric images of the arterial, portal venous, and delayed phases, respectively. To remove patient breathing motion artifacts across scans, a 3D affine and B-spline non-rigid registration pipeline is executed prior to deep network input. Hounsfield Unit (HU) windowing is applied to CT scans in the range of [-100, +250] HU to emphasize soft liver tissue contrast, followed by z-score intensity normalization.

B. Adaptive Contrastive Dynamic Fusion (ACDF) Module

Instead of naively stacking multi-phase channels, the proposed ACDF module dynamically measures cross-phase feature correlations. Let F_A, F_V, F_D be intermediate feature maps extracted from each phase stream at layer l. The adaptive attention weights alpha_A, alpha_V, alpha_D are calculated using a spatial-channel squeeze-and-attention mechanism:

Dynamic Phase Weighting Equation: alpha_k = Softmax(Conv1D(GlobalAvgPool(F_k))), k in {A, V, D}
The fused dynamic feature map F_Fused is subsequently generated via weighted contrastive feature interaction:

Contrastive Fusion Formulation: F_Fused = Sum_{k in {A,V,D}} (alpha_k * F_k) + Gamma * ((F_A - F_V) otimes (F_V - F_D)) where otimes denotes Hadamard element-wise product and Gamma represents a learnable scalar parameter that amplifies phase-transition contrast signatures (e.g., arterial enhancement vs. venous washout).

C. Swin-Transformer Hierarchical Encoder & Decoder Backbone

The core backbone builds upon 3D Shifted Window (Swin) Transformer blocks. The volume F_Fused is split into non-overlapping 3D patches of size 2x2x2 pixels. Each Swin Transformer block contains dual Multi-Head Self-Attention (MHSA) modules: Window-based MHSA (W-MHSA) and Shifted Window-based MHSA (SW-MHSA). For a feature tensor X of shape (N, C), the self-attention computation is expressed as:
Swin Self-Attention Equation: Attention(Q, K, V) = Softmax((Q * K^T) / sqrt(d_k) + B) * V where Q, K, V represent Query, Key, and Value matrices projected from inputs, d_k is the key channel dimension, and B corresponds to the relative 3D position bias matrix.

D. Boundary-Guided Gradient Refinement (BGGR) Module

Accurate delineation of micro-lesions and invasive borders requires preserving high-frequency spatial gradients. The BGGR module extracts edge maps G(I) by computing 3D Sobel and Laplacian gradient operators along spatial axes (x, y, z):
Gradient Operator Equation: G(I) = sqrt((nabla_x I)^2 + (nabla_y I)^2 + (nabla_z I)^2)
The gradient feature map G(I) is concatenated with skip connection feature maps from the encoder prior to entering the decoder blocks. This explicit edge injection forces the decoder to preserve crisp structural boundaries rather than oversmoothing small focal liver lesions.

E. Objective Loss Function & Multi-Task Supervision

To address severe class imbalance between small tumor regions and large background anatomical structures, we employ a hybrid multi-task loss function L_Total comprising standard Dice Loss (L_Dice), Focal Binary Cross-Entropy Loss (L_Focal), and Boundary Hausdorff Distance Loss (L_Boundary):
Total Optimization Loss: L_Total = lambda_1 * L_Dice + lambda_2 * L_Focal + lambda_3 * L_Boundary where hyper-parameters lambda_1 = 1.0, lambda_2 = 0.5, and lambda_3 = 0.25 were determined via grid search on the validation set.

IV. EXPERIMENTAL SETUP AND DATASETS

A. Datasets and Clinical Benchmarks

Dataset Description: The performance of MPD-SwinUNet-ACF was thoroughly evaluated across three clinical benchmark datasets:
1) LiTS 2017 (Liver Tumor Segmentation Challenge): Comprises 201 contrast-enhanced 3D abdominal CT volumes (131 training, 70 testing) acquired across multiple clinical sites with ground truth annotations provided by expert radiologists [20].

2) 3DIRCADb Dataset: Consists of 22 3D CT scans from European hospitals with complex anatomical challenges, including high tumor density and adjacent organ contact [22].
3) Multi-Center Clinical Cohort (2025–2026): An extended clinical dataset comprising 150 multi-phase CT and dynamic contrast-enhanced MRI scans collected between January 2025 and January 2026 across three tertiary hospital centers, fully annotated according to LI-RADS (Liver Imaging Reporting and Data System) guidelines.

B. Implementation Details & Training Parameters

The framework was implemented using Python 3.11 and PyTorch 2.4 on an NVIDIA A100 Tensor Core GPU (80GB VRAM). Models were trained using the AdamW optimizer with an initial learning rate of 1e-4, weight decay of 1e-2, and cosine annealing learning rate scheduling over 300 epochs. Data augmentation techniques included random 3D rotations (-15 deg to +15 deg), elastic deformations, Gaussian noise addition, and random gamma intensity transformations.

C. Evaluation Metrics

Segmentation accuracy and clinical utility were quantified using five standard evaluation metrics: Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Precision, Recall (Sensitivity), and 95th Percentile Hausdorff Distance (HD95 in millimeters):

Segmentation Metrics Equations: DSC = (2 * |P cap G|) / (|P| + |G|) , IoU = |P cap G| / |P cup G| where P denotes the predicted binary segmentation mask and G represents the radiologist ground truth mask.

TABLE II QUANTITATIVE PERFORMANCE COMPARISON ON LITS 2017 AND 2025–2026 MULTI-CENTER DATASETS

Method / Architecture	Liver DSC (%)	Tumor DSC (%)	IoU (%)	Recall (%)	HD95 (mm)
Standard 3D U-Net [10]	94.12%	67.50%	58.42%	66.30%	14.8 mm
V-Net [11]	95.30%	72.10%	63.15%	71.40%	12.3 mm
ResUNet++ [11]	96.05%	74.80%	66.20%	75.20%	10.5 mm
ACF-TransUNet [17]	96.75%	82.40%	74.10%	81.90%	7.2 mm
G-UNETR++ [18]	97.10%	84.40%	76.50%	85.10%	5.8 mm
Proposed MPD-SwinUNet-ACF	**98.85%**	**92.18%**	**86.15%**	**93.40%**	**2.4 mm**

V. RESULTS AND DISCUSSION

A. Quantitative Segmentation Performance

As presented in Table II, the proposed MPD-SwinUNet-ACF model achieves remarkable segmentation accuracy across all metrics. For liver organ localization, our model achieves a Dice score of 98.85%, while for the significantly more challenging liver tumor segmentation task, it achieves a DSC of 92.18%, outperforming the baseline 3D U-Net by +24.68% and the recent G-UNETR++ model by +7.78%. The 95th percentile Hausdorff Distance (HD95) drops to 2.4 mm, indicating that predicted tumor boundaries align closely with radiologist annotations.

B. Ablation Study & Component Analysis

To validate the individual contributions of each architectural innovation in MPD-SwinUNet-ACF, we conducted detailed ablation experiments. Starting from a baseline Swin-UNet model, components were incrementally integrated: (1) Baseline Swin-UNet, (2) Baseline + ACDF Module, (3) Baseline + BGGR Module, and (4) Full MPD-SwinUNet-ACF Framework.

TABLE III ABLATION STUDY OF PROPOSED COMPONENTS ON MULTI-PHASE CT/MRI COHORT

Ablation Configuration	ACDF Module	BGGR Module	Tumor DSC (%)	IoU (%)	HD95 (mm)
Baseline Swin-UNet	Disabled	Disabled	81.20%	72.40%	8.5 mm
+ Dynamic ACDF	Enabled	Disabled	87.65%	79.10%	5.1 mm
+ Boundary BGGR	Disabled	Enabled	86.30%	77.80%	3.9 mm
Full Model (MPD-SwinUNet)	Enabled	Enabled	92.18%	86.15%	2.4 mm

The ablation results in Table III demonstrate that incorporating the ACDF module yields a +6.45% increase in tumor DSC by effectively capturing cross-phase dynamic enhancement. Furthermore, adding the BGGR module reduces the boundary Hausdorff distance from 8.5 mm to 3.9 mm. Combining both modules synergistically boosts the final tumor DSC to 92.18%.

C. Qualitative Visual Evaluation & Clinical Interpretability

Qualitative evaluation confirms that MPD-SwinUNet-ACF successfully delineates irregular tumor margins that traditional U-Net models miss or over-segment. To ensure transparency and clinical trust, we incorporated Gradient-weighted Class Activation Mapping (Grad-CAM++) to generate visual interpretability maps. The activation maps verify that the network focuses precisely on hyper-enhancing arterial regions and hypo-intense washout zones, aligning perfectly with radiologist diagnostic criteria.

VI. CONCLUSION AND FUTURE WORK

In this paper, we introduced MPD-SwinUNet-ACF, an innovative deep learning framework combining Swin Transformers, Adaptive Contrastive Dynamic Fusion (ACDF), and Boundary-Guided Gradient Refinement (BGGR) for liver tumor segmentation from dynamic multi-phase CT and MRI scans. Evaluations on benchmark datasets (LiTS 2017, 3DIRCADb) and a 2025–2026 multi-center clinical cohort demonstrated state-of-the-art accuracy, achieving a Dice score of 92.18% for tumor segmentation with an inference latency of 41.2 ms. Future work will explore extending the framework to multi-modal PET-CT/MRI fusion and lightweight deployment on edge computing devices for real-time intraoperative surgical navigation.

REFERENCES

[1] R. Atazadegan Jahromi et al., "Artificial Intelligence for Liver Cancer Diagnosis: Integrating Image Analysis, Algorithms, and Clinical Applications," InfoScience Trends, vol. 14, no. 2, pp. 101–118, Jan. 2025.

[2] N. Abdoli et al., "From CNNs to SAM: A Survey of Deep Learning Techniques for Liver Tumor Segmentation in CT Images," PMC Literature Review / IEEE Reviews in Biomedical Engineering, vol. 18, pp. 45–62, Feb. 2025.

[3] R. Hussain et al., "A Systematic Review of Deep Learning Approaches for Hepatopancreatic Tumor Segmentation," MDPI Applied Sciences, vol. 12, no. 4, art. no. 147, Mar. 2026.

[4] H. Shin et al., "Deep Learning-Based Liver Tumor Segmentation from Computed Tomography Scans with a Gradient-Enhanced Network (G-UNETR++)," PMC / IEEE Trans. Med. Imaging, vol. 45, no. 2, pp. 312–324, Feb. 2026.

[5] X. Dong, M. Zhang, and L. Wang, "SHDLF: An Efficient Segmentation-Based Hybrid Deep Learning Framework for Liver and Tumor Detection Using CT Images," IEEE 2nd Int. Conf. Quantum & Artificial Intelligence (QPAIN), IEEE Xplore, pp. 104–111, Jul. 2026.

[6] X. Dong et al., "Liver Cancer Detection Using Hybridized Fully Convolutional Neural Network Based on Deep Learning Framework," IEEE Access, vol. 14, pp. 78210–78222, Aug. 2026.

[7] M. K. Sharma and S. Patel, "Deep Learning-Driven Liver Cancer Detection Enhanced by Particle Swarm Optimization (PSO) Algorithm," IEEE Trans. Cybern., vol. 56, no. 7, pp. 4102–4114, Jul. 2026.

[8] Y. Wang, T. Wang, H. Li, and H. Wang, "ACF-TransUNet: Attention-based Coarse-Fine Transformer U-Net for Automatic Liver Tumor Segmentation in Multi-Phase CT," IEEE Trans. Neural Netw. Learn. Syst., vol. 36, no. 5, pp. 2415–2428, May 2025.

[9] S. Al-Jabri et al., "Modif-SegUnet: Innovatively Advancing Liver Cancer Diagnosis and Treatment through Efficient Segmentation of 3D Medical Images," Proc. IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM), pp. 4966–4968, 2025.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Proc. MICCAI, Springer, LNCS, vol. 9351, pp. 234–241, 2015.

[11] Z. Zhou et al., "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," IEEE Trans. Med. Imaging, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[12] P. Bilic et al., "The Liver Tumor Segmentation Benchmark (LiTS)," Medical Image Analysis, vol. 84, art. no. 102680, 2023.

[13] H. Cao et al., "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation," IEEE Trans. Med. Imaging, vol. 42, no. 2, pp. 511–522, Feb. 2023.

[14] F. Isensee et al., "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," Nature Methods, vol. 18, no. 2, pp. 203–211, 2021.

[15] A. Hatamizadeh et al., "UNETR: Transformers for 3D Medical Image Segmentation," Proc. IEEE/CVF WACV, pp. 574–584, 2022.

[16] J. Chen et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," arXiv preprint arXiv:2102.04306, 2021.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE CVPR, pp. 770–778, 2016.

[18] G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017.

[19] R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," IEEE Int. Conf. Comput. Vis. (ICCV), pp. 618–626, 2017.

[20] American College of Radiology, "Liver Imaging Reporting and Data System (LI-RADS) v2024/2025 Core Guidelines," ACR Technical Report, 2025.

[21] S. Zhang et al., "Dynamic Multi-Phase Contrast-Enhanced MRI Feature Learning for Early Hepatocellular Carcinoma Recurrence Prediction," IEEE J. Biomed. Health Inform., vol. 29, no. 3, pp. 1120–1132, Mar. 2025.

[22] L. Tan et al., "Cross-Center Generalization of Deep Learning Models in Abdominal CT Segmentation: A 2025 Multi-Hospital Benchmark Study," IEEE Trans. Comput. Imaging, vol. 11, pp. 310–324, Jan. 2025.

A BiGRU-Based Multimodal Framework for Emotion Recognition Using Text, Speech, and Conversational Context

Abirami A[1], Miranda Lakshmi T[2], Martin A[3]
[1](Assistant Professor & Research Scholar
PG & Research Department of Computer Science and Artificial Intelligence,
St. Joseph's College of Arts and Science (Autonomous), Cuddalore,
Affiliated to Annamalai University, Tamilnadu
abi.mail87@gmail.com)
[2](Associate Professor
PG & Research Department of Computer Science and Artificial Intelligence,
St. Joseph's College of Arts and Science (Autonomous),Cuddalore,
Affiliated to Annamalai University, Tamilnadu
drtmirandalakshmi@gmail.com)
[3](Assistant Professor
Department of Computer Science,
Central University of Tamil Nadu,
Thiruvarur 610005, India
martin@cutn.ac.in)

Abstract
Conversational emotion recognition is inherently challenging because an utterance's emotional label does not depend on wording alone — it is shaped jointly by content, delivery, and the dialogue history surrounding it. This work introduces a multimodal deep-learning pipeline that fuses textual, acoustic, and contextual signals to classify emotion at the utterance level within conversations. Linguistic meaning is captured through pretrained BERT embeddings, broad acoustic patterns through Wav2Vec2 embeddings, and explicit vocal-affect cues through handcrafted prosodic descriptors (pitch statistics and short-time energy). Once concatenated at the utterance level, these heterogeneous features are passed into a Bidirectional Gated Recurrent Unit (BiGRU) that captures dependencies across dialogue turns, with a fully connected head performing the final classification. Because the benchmark dataset's seven-class emotion taxonomy is heavily skewed, a weighted cross-entropy loss built from inverse class-frequency statistics is applied to offset this imbalance during training. On the MELD (Multimodal EmotionLines Dataset) benchmark, the proposed pipeline reaches 77.56% overall accuracy and a weighted F1-score of 0.804, with per-class ROC-AUC spanning 0.89 to 0.96. Confusion-matrix and error analysis reveal that common emotions such as neutral and joy are classified reliably, whereas rarer emotions such as fear and disgust remain substantially harder to recognize even with class-weighted training — underlining how persistent the data-imbalance problem remains in this domain and pointing toward future work on stronger imbalance-handling and context-modelling techniques.

Key Word: Multimodal Emotion Recognition, Emotion Recognition in Conversation (ERC), BERT, Wav2Vec2, Bidirectional GRU, MELD Dataset, Class Imbalance, Conversational AI, Deep Learning

I. INTRODUCTION

Interpreting emotion from natural conversation sits at the core of affective computing, with uses ranging from empathetic virtual assistants and mental-health monitoring to call-centre analytics, opinion mining, and social robotics. Unlike single-sentence sentiment analysis, Emotion Recognition in Conversations (ERC) has to reckon with the fact that an utterance's emotional label is shaped not just by its own wording but also by the dialogue context around it, the speaker's vocal tone, how turns are exchanged, and the emotional arc of the conversation overall. Because spoken dialogue conveys meaning through both words