The Open Biomedical Engineering Journal
DOI: 10.2174/0118741207485589260704082138, 2026, 20, e18741207485589
ISSN: 1874-1207

RESEARCH ARTICLE **OPEN ACCESS**

A Novel Attention-guided Curriculum Learning Approach for Multi-modal Medical Image Segmentation and Disease Classification

Santosh Kumar Mishra[1,*], Suma Dawn[1] and Raju Pal[2]

[1]*Department of Computer Science & Engineering and Information Technology, Jaypee Institute of Information Technology, Noida, India*
[2]*Department of Computer Science & Engineering, University School of Information and Communication Technology, Gautam Buddha University, Greater Noida, India*

Abstract:

Introduction: Recent advances in medical imaging and artificial intelligence require robust deep learning frameworks capable of processing heterogeneous imaging modalities for accurate segmentation and disease classification. Conventional models often suffer from feature heterogeneity, unstable optimization, and poor generalization, creating the need for intelligent attention-guided and adaptive multi-modal learning approaches.

Methods: This study proposes a novel Multi-Modal Attention-Guided Curriculum Learning framework integrating modality-aware attention, spatial-channel feature fusion, and curriculum learning for simultaneous segmentation and disease classification. The framework was evaluated on 3,200 paired MRI and CT patient cases containing pixel-level annotations and disease labels.

Results: Experimental analysis demonstrated that the proposed Multi-Modal Attention-Guided Curriculum Learning framework outperformed Multi-Modal U-Net, Attention U-Net, Swin-UNet, and CNN–Transformer Hybrid models. MAGCL achieved 95.9% classification accuracy, 95.7% F1-score, 95.6% Dice score, 91.2% Intersection over Union, and 0.978 AUC-ROC with improved training stability and generalization.

Discussion: The superior performance of the proposed Multi-Modal Attention-Guided Curriculum Learning is attributed to the integration of modality-aware attention and progressive curriculum learning. Spatial and channel attention mechanisms improved pathological feature localization, while curriculum learning stabilized optimization through gradual inclusion of difficult samples, resulting in enhanced feature discrimination, segmentation accuracy, and disease classification reliability.

Conclusion: The proposed Multi-Modal Attention-Guided Curriculum Learning framework provides an effective and clinically reliable solution for multi-modal medical image segmentation and disease classification. By integrating modality-aware attention with curriculum learning, the framework significantly improves feature fusion, segmentation precision, disease prediction accuracy, and optimization stability, demonstrating strong potential for intelligent clinical decision-support systems.

Keywords: Multi-modal imaging, Attention mechanisms, Curriculum learning, CNN–Transformer, Medical image analysis, Disease classification, CNN–Transformer.

*Address correspondence to this author at the Department of Computer Science & Engineering and Information Technology, Jaypee Institute of Information Technology, Noida, India; E-mail: mishra.santosh19@gmail.com

Cite as: Mishra S, Dawn S, Pal R. A Novel Attention-guided Curriculum Learning Approach for Multi-modal Medical Image Segmentation and Disease Classification. Open Biomed Eng J, 2026; 20: e18741207485589.
http://dx.doi.org/10.2174/0118741207485589260704082138

Received: February 17, 2026
Revised: May 10, 2026
Accepted: June 04, 2026
Published: July 28, 2026

1. INTRODUCTION

The use of deep learning methods, specifically convolutional neural networks, has become a major driver of recent developments in the field of medical image analysis, as they have been shown to perform well in image segmentation, classification, and detection [1, 2]. The encoder-decoder structures have formed the basis of several segmentation systems since they are capable of extracting hierarchical spatial information and maintaining fine-scale anatomical structure by skip connections. The wide application of these models has been observed in different medical imaging modalities such as MRI, CT, ultrasound, and PET. Traditional convolution-based methods can have problems with generalization in the face of heterogeneous multi-modal data, differences in imaging protocols, and complicated pathological patterns, even though they are successful [3-5].

In order to overcome the drawbacks of traditional convolutional networks, the concept of multi-modal learning has also been proposed in order to combine complementary data provided by other sources of imaging technology. Early fusion processes integrate modalities at the input side, and late fusion techniques integrate features or predictions at a higher level of the network. As much as these strategies enhance the capacity of representation, they tend to presuppose the equal contribution of both modalities and are not adaptable to sample-specific diagnostic relevance. This weakness can diminish the robustness of cases where there is noise, artifact, or missing data in one modality, as observed in real-world clinical data [6-8].

Attention mechanisms have proved to be a viable solution to improve feature discrimination in the anatomy of medical images. Attention-based models enhance the accuracy of localization and classification confidence by allowing networks to selectively emphasize regions of particular interest in space and information-rich feature channels. Spatial attention is useful in determining pathological areas, whereas channel attention is used to highlight discriminative feature maps related to the characteristics of the disease. The mechanisms of attention have been effectively incorporated in segmentation networks in order to suppress noise in the background and optimize the delineation of boundaries. Nevertheless, the majority of the currently existing attention-based models use fixed attention strategies during training and fail to consider the difference in the difficulty of training samples [8-10].

1.1. Recent Advances in AI-based Medical Image Analysis

Recent developments in deep learning, neural networks, and intelligent optimization have significantly advanced biomedical image analysis and disease diagnosis systems [11-13]. Several studies have demonstrated the effectiveness of neural-network-based learning frame-works in solving complex biomedical and mathematical modeling problems. Bayesian regularization neural networks, radial basis function networks, and wavelet neural models have shown promising performance in predictive analysis, disease modeling, and optimization-driven medical systems. Similarly, attention-guided deep learning architectures and transformer-based frameworks have demonstrated superior capabilities in capturing long-range contextual dependencies and discriminative spatial representations in medical imaging applications [14, 15].

1.2. Importance of Medical Image Segmentation

Medical image segmentation plays a fundamental role in computer-assisted diagnosis, treatment planning, and disease monitoring. Accurate segmentation enables precise delineation of pathological regions such as tumors, lesions, and infected tissues, allowing clinicians to estimate disease progression, tumor volume, and anatomical abnormalities. Segmentation also improves radiotherapy planning, surgical navigation, and quantitative disease assessment [16].

1.3. Novelty and Major Contributions of the Proposed Work

The major contributions and novelties of this work are summarized as follows:

(1) A novel Multi-Modal Attention-Guided Curriculum Learning (MAGCL) framework is proposed for simultaneous medical image segmentation and disease classification.

(2) A modality-aware attention mechanism is introduced to dynamically prioritize diagnostically relevant imaging modalities during feature fusion.

(3) An attention-guided curriculum learning strategy is developed to progressively train the model from easy to difficult samples using segmentation uncertainty, attention consistency, and classification confidence.

(4) A unified multi-task optimization framework is designed to jointly improve segmentation and disease classification performance.

(5) Extensive comparative analysis and ablation studies demonstrate the superiority of MAGCL over recent convolutional, attention-based, and transformer-based approaches.

The remainder of this paper is organized as follows. Section 2 discusses recent advancements and related work in deep learning-based medical image segmentation and disease classification. Section 3 presents the proposed Multi-Modal Attention-Guided Curriculum Learning framework and explains the detailed architecture, attention mechanisms, and curriculum learning strategy. Section 4 describes the experimental setup, dataset characteristics, evaluation metrics, and implementation details. Section 5 presents comparative experimental results, ablation studies, and performance analysis. Finally, Section 6 concludes the paper and outlines future research directions.

2. LITERATURE REVIEW

Recently, transformer-based architectures have been popular in medical imaging because it allows the modeling of long-range dependencies and global contextual relationships. Transformers, in contrast to convolutional networks, use it to focus on interactions between distant spatial regions and are especially effective for use with complex anatomical structures and diffuse disease patterns [17]. The hybrid schemes of convolutional encoders by

means of transformer-based fusion layers have been suggested in order to strike a balance between local feature extraction and global reasoning. Although these models have good performance, they tend to have high data needs along with training techniques, thus being prone to overfitting and instability in multi-modal medical imaging scenarios [18, 19].

Curriculum learning is a training paradigm that has been proposed based on empirical studies on human learning processes, wherein models are trained using increasingly complex samples. Curriculum learning has demonstrated capabilities to enhance convergence speed, minimal sensitivity to noisy labels, and generalization in medical image analysis. Heuristic measures of sample difficulty, such as loss values or prediction confidence, are usually used to define sample difficulty. Even though it has shown a lot of promise, little has been done to study curriculum learning in the context of multi-modal medical image segmentation and classification, with the available methods seldom incorporating curriculum strategies into attention-driven feature learning [20, 21].

Multi-task learning models that concurrently utilize segmentation and classification have also received interest because they can take advantage of common representations. Correct segmentation may also provide spatial cues to classification, and classification goals may promote the learning of semantically significant features. Nevertheless, it is difficult to find an equilibrium between two or more goals, especially in cases where training data are class-imbalanced or have vague boundaries. The majority of the current multi-task strategies maximize all samples equally, and this may be restrictive in the face of heterogeneous and complex clinical data [22].

The existing medical image analysis methodologies still have numerous gaps despite the huge progress. The models that are in place tend to view attention mechanisms, multi-modal fusion, and curriculum learning as unrelated aspects rather than synergies [23]. Limited research has been done on the application of attention reliability as a supervisory cue to influence the learning process, especially in the selection of sample difficulty and training rate. Also, the issue of adaptive modality weighting when using diagnostic relevance of samples, which is specific to a certain sample, is seldom considered in a single framework [19, 20].

These constraints drive the idea of a learning paradigm that is a combined approach to attention-based feature learning and curriculum-based training in a multi-modal environment. To fill this gap, the proposed research proposes a Multi-Modal Attention-Guided Curriculum Learning (MAGCL) framework, which provides a dynamic use of attention confidence to control the training progression and segmentation and disease classification performed jointly. This study will enhance robustness, generalization, and clinical reliability by establishing a single framework that can improve its ability to unify attention mechanisms, curriculum learning, and multi-modal fusion to advance the state of the art in intelligent medical image analysis.

3. PROPOSED METHODOLOGY

The proposed Multi-Modal Attention-Guided Curriculum Learning (MAGCL) model employs a combined unified end-to-end model that is aimed at concurrently performing multi-modal medical image segmentation and disease classification with the use of an attention-based curriculum learning paradigm, as shown in Fig. (**1**). The general architecture can encode modality-specific features and then proceed to more complex pattern discovery of heterogeneous medical imaging modalities, including MRI and CT. The dataset used in this paper consists of 3,200 multi-dimensional patient cases, which include the paired imaging modalities with pixel-level segmentation labels and the disease labels. The data is distributed with 70 percent of it being dedicated to training, 15 percent to validation, and 15 percent to testing, ensuring a balanced distribution across the disease type categories.

At the first stage, the multi-modal input encoder is a separate process that processes each imaging modality through parallel convolutional branches. These encoders attain low-level spatial features and high-level semantic representations whilst maintaining modality-specific features. The framework reduces the loss of information early in the encoding process and enables each modality to provide information that is complementary in the form of diagnostic information. These coded features are then passed on to an attention-directed feature fusion unit, which combines the multi-modal features adaptively according to their relevance in a particular situation as opposed to concatenating them or averaging them. The attention-guided feature learning mechanism combines spatial and channel attention mechanisms to perfect feature representations. Spatial attention helps the network to focus on disease-related anatomical structures, whereas channel attention enhances discriminative feature maps and patterns of pathology. In addition, modality weighting of attention dynamically changes the contribution of each modality based on the diagnostic usefulness of each to a particular sample. Such a design will allow paying attention to areas of clinical interest, eliminating noise and redundant backgrounds, improving segmentation, and generating more credible classification results.

Another important innovation of this work is the curriculum learning schedule, which coordinates the training process through the organisation of the samples based on difficulty. Sample difficulty is determined using a composite measure: sample segmentation uncertainty, confidence of the classification, and consistency of attention scores. The network will first be trained on simpler samples, which are defined by clear boundaries between the anatomy and a high level of attention. The cases of greater complexity and ambiguity are introduced gradually as the training goes on. It is the progression between easy and hard that makes the process of optimisation more stable, faster convergent, and more generalised with a variety of patient profiles. Notably, the confidence of attention is integrated in the curriculum strategy so that samples having weak attention maps would be delayed in subsequent training phases.

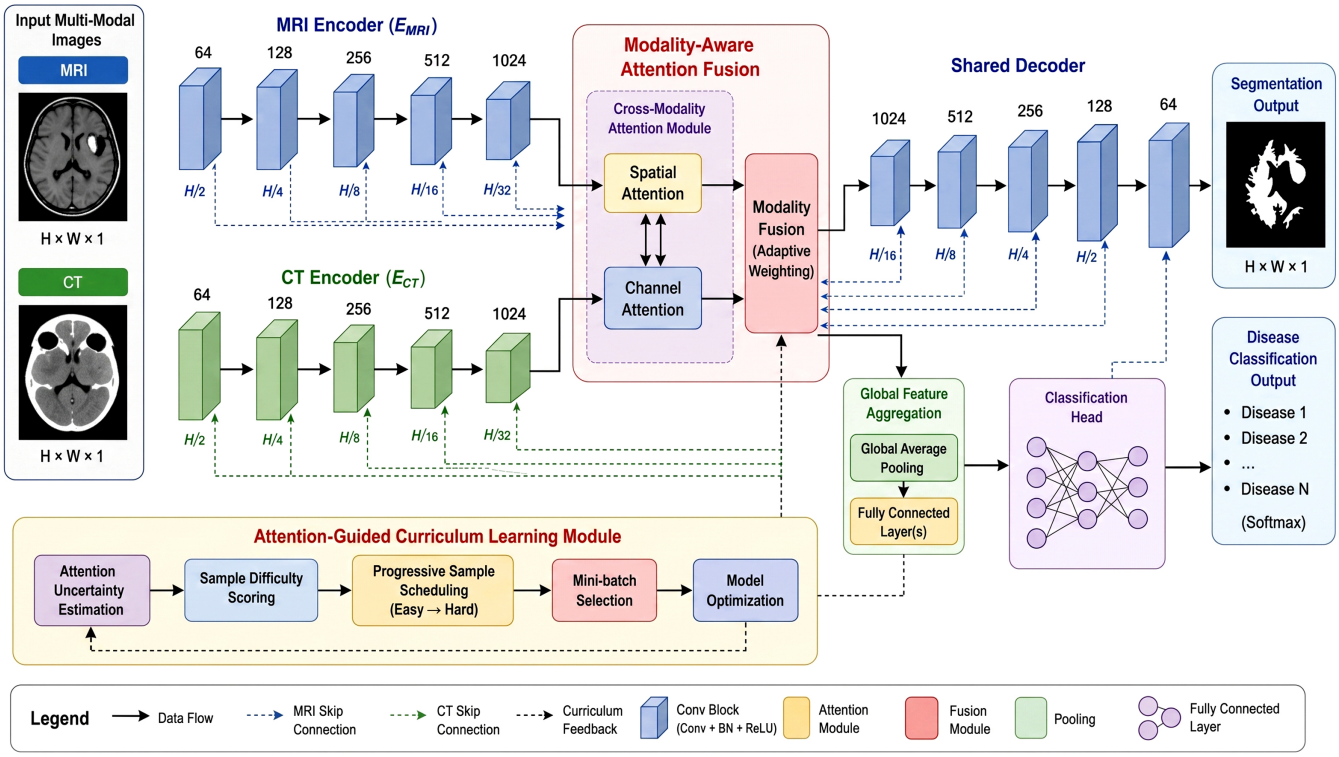

Fig. (1). Architecture of the proposed multi-modal attention-guided curriculum learning (MAGCL) framework for joint medical image segmentation and disease classification.

The network also uses a dual-head output design that can simultaneously do segmentation and classification. The segmentation head provides pixel-wise predictions that define pathological areas on a high-spatial level, whereas the classification head generates disease labels based on global-aggregated features. This common representation learning nurtures reinforcement between the two tasks, where correct segmentation facilitates classification and the reverse.

A joint multi-task loss determines training optimisation. A mixture of Dice loss and cross-entropy loss is used to optimise segmentation performance and balance accuracy of region-overlaps and pixel-wise classification. The classification of diseases is controlled either by cross-entropy loss or focal loss to reduce class imbalance. The integrated loss operation provides coordinated learning between tasks, which eventually results in a sound and clinically significant MAGCL system. The detailed proposed MAGCL architecture for Multi-model medical image segmentation and disease classification is shown in Fig. (2).

3.1. Explanation of the MAGCL Architecture

The proposed Multi-Modal Attention-Guided Curriculum Learning (MAGCL) framework is designed to jointly perform medical image segmentation and disease classification by effectively integrating multi-modal data, attention mechanisms, and curriculum-based training.

3.1.1. Multi-modal Encoders

The architecture begins with parallel modality-specific encoders for each imaging modality (*e.g.*, MRI and CT). Each encoder extracts hierarchical spatial and semantic features while preserving modality-specific diagnostic characteristics. This parallel encoding prevents early information loss and enables complementary feature learning across modalities.

3.1.2. Attention-guided Feature Fusion

Encoded features are passed to an attention-guided fusion module consisting of:

- *Spatial Attention*, which highlights anatomically relevant regions associated with pathology.
- *Channel Attention*, which emphasizes discriminative feature maps related to disease patterns.
- *Modality Weighting*, which dynamically adjusts the contribution of each modality based on its diagnostic relevance for a given sample.

These mechanisms collectively suppress irrelevant information and produce a unified, attention-refined feature representation.

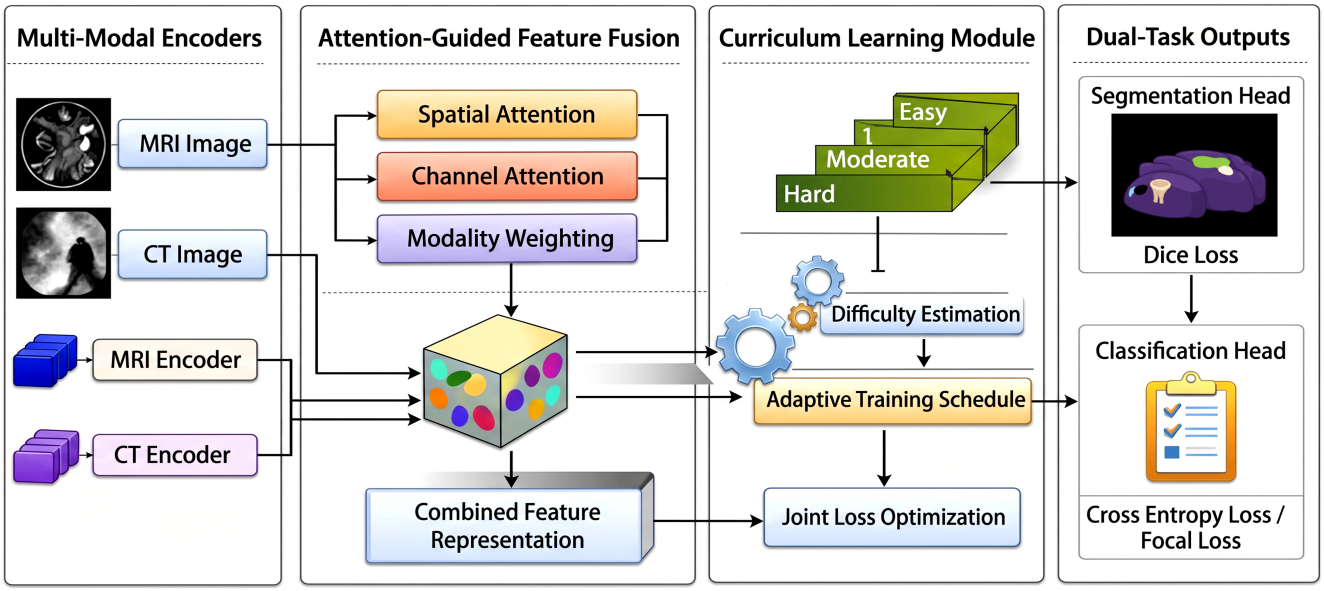

Fig. (2). Detailed MAGCL architecture for multi-model medical image segmentation and disease classification.

3.1.3. Curriculum Learning Module

An attention-guided curriculum learning strategy organizes training samples into easy, moderate, and hard categories. Sample difficulty is estimated using segmentation uncertainty, classification confidence, and attention consistency. Training begins with high-confidence samples and progressively incorporates more complex cases, leading to improved optimization stability and generalization.

3.1.4. Dual-task Output Heads

The refined features are fed into two parallel task heads:

- A Segmentation Head, producing pixel-wise predictions optimized using Dice loss.
- A Classification Head, generating disease labels optimized using cross-entropy or focal loss.

3.1.5. Joint Loss Optimization

Both tasks are trained simultaneously using a joint loss function, allowing mutual reinforcement between segmentation and classification and resulting in more clinically reliable predictions.

3.2. Algorithm

Algorithm 1: Multi-Modal Attention-Guided Curriculum Learning (MAGCL)

Input:

Multi-modal medical images $\{X\} = \{X^{\{(1)\}}, X^{\{(2)}, \ldots .., X^{\{(M)\}}\}$

Ground-truth segmentation masks Y_{seg}

Disease class labels Y_{cls}

Output:

Trained MAGCL model for segmentation and disease classification

3.2.1. Proposed MAGCL Algorithm Mathematical Formulation and Loss Functions

Start Algorithm

Step 1: Data Preparation

1. Acquire paired multi-modal medical images (*e.g.*, MRI, CT).

2. Apply preprocessing including normalization, resizing, and data augmentation.

3. Split the dataset into training, validation, and testing sets.

Step 2: Modality-Specific Feature Encoding

4. Initialize modality-specific CNN encoders E_m for each modality $m \in M_m$.

5. Extract hierarchical spatial and semantic features using Eq. (**1**):

$$Fm = Em\left(X^{\{(m)\}}\right) \quad \textbf{(1)}$$

Step 3: Attention-Guided Feature Fusion

6. Apply spatial attention to highlight disease-relevant anatomical regions.

7. Apply channel attention to emphasize discriminative feature maps.

8. Compute modality-wise attention weights to adaptively scale each modality.

9. Fuse attention-refined features to obtain a unified representation given in Eq. (**2**):

$$F_{fusion} = \sum_{m=1}^{M} \alpha_m \cdot Attn(F_m) \quad \textbf{(2)}$$

Step 4: Curriculum Learning Strategy

10. Estimate sample difficulty using:

• Segmentation uncertainty

• Classification confidence

• Attention consistency score.

11. Rank training samples from easy to hard.

12. Train the model in phases by progressively introducing harder samples based on attention confidence.

Step 5: Dual-Task Prediction

13. Feed fused features into a shared CNN-Transformer backbone.

14. Generate segmentation outputs using the segmentation head given in Eq. (**3**):

$$\widehat{Y_{seg}} = H_{seg}(F_{fusion}) \quad \textbf{(3)}$$

15. Generate disease classification outputs using the classification head using Eq. (**4**):

$$\widehat{\mathrm{Y_{cls}}} = \mathrm{H_{cls}(F_{fusion})} \quad \textbf{(4)}$$

Where:

$\widehat{Y_{cls}}$: The predicted output or probability for disease classification.

H_{cls} :The classification head (typically a neural network layer or sequence of layers).

F_{fusion}: The fused feature representation used as input for the classification.

Step 6: Joint Loss Optimization

16. Compute segmentation loss using Dice and Cross-Entropy loss using Eq. (**5**):

$$\mathcal{L}_{seg} = \mathcal{L}_{Dice} + \mathcal{L}_{\mathcal{CE}} \quad \textbf{(5)}$$

Where:

L_{seg}: The total segmentation loss.

L_{Dice}: The Dice loss, used to measure the overlap between predicted and ground truth masks.

$L_{C\mathcal{E}}$: The Cross-Entropy loss, used to measure pixel-wise classification accuracy.

17. Compute classification loss using Cross-Entropy or Focal loss using Eq. (**6**).

$$\mathcal{L}_{\mathcal{CE}} \quad \textbf{(6)}$$

18. Total Joint Loss Optimization calculated using Eq. (**7**):

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{seg} + \lambda_2 \mathcal{L}_{cls} \quad \textbf{(7)}$$

Where:

L_{total}:The final joint loss used for model optimization.

L_{seg}: The computed segmentation loss.

L_{cls}: The computed classification loss (derived *via* Cross-Entropy or Focal loss).

λ_1 λ_2: Weighting coefficients (hyperparameters) that balance the contribution of the segmentation and classification tasks .

Step 7: Model Update and Convergence

19. Update network parameters *via* backpropagation.

20. Repeat training until convergence criteria are met.

21. Evaluate the trained model on the test dataset using segmentation and classification metrics.

End Algorithm

3.3. Model used in this Research Work

3.3.1. Multi-modal U-Net Architecture

Multi-Modal U-Net architecture is used as a powerful convolutional baseline in the present study since it verifies the traditional U-Net system by expanding it to learn in various medical imaging modalities. Every modality goes through a special encoder pathway of convolutional, normalization, and pooling layers that successively extract hierarchical spatial features. These modality-specific encoders conserve unique anatomical and textual data that are unique to individual sources of imaging, *e.g.* MRI and CT. The extracted features are merged at either intermediate or bottleneck points through concatenation, allowing the decoder to recreate the segmentation maps by combining the complementary information between modalities. The decoder is based on the conventional U-Net structure and uses the upsampling and skip-connections to help restore finer spatial details, which have been lost during the downsampling process [17, 24].

Global average pooling of the bottleneck features, which are fused, is used for disease classification; then full-connected layers are used that produce class prediction. Though good at spatial patterns and offering stable performance, this architecture treats all modalities equally and is not able to explicitly lay stress on clinically relevant areas, making it an appropriate but constrained baseline with which more sophisticated attention- and curriculum-driven models can be compared.

3.3.2. Attention U-Net Architecture

Attention U-Net architecture is an improvement of the traditional U-Net where attention gates are added to the skip connection routes between the decoder and the encoder. These attention gates are meant to sieve out the irrelevant background information and bring out the salient features of the anatomy before feature combining. Under multi-modal medical imaging, the convolutional layers are first used to encode each modality in the same manner as the conventional U-Net structure. In the decoding step, the relevance score applied by the attention mechanisms relies on the contextual information of deeper layers that enables the network to selectively propagate the most pertinent features to the target pathology. This biased propagation of features improves the delineation of segmentation boundaries, especially when the area of pathology is limited or when there is a low contrast. To classify, the attention-refined features are

summarized to produce disease prediction, which has the advantage of richer region-specific representations. Even though Attention U-Net significantly outperforms basic U-Net in terms of localization, this type of network lacks any global contextual modelling and adaptive learning mechanisms and thus, is not capable of dealing with very complex or ambiguous multi-modal problems [11, 18, 25].

3.3.3. Swin-UNet Model

The Swin-UNet model is a transformer-based segmentation framework that is based on hierarchical self-attention mechanisms to address long-range relationships in medical images. In contrast to entirely convolutional architectures, Swin-UNet deploys shifted window-based self-attention to capture the global context without compromising the computational performance. In this study, transitional multi-modal inputs have been integrated into patch tokens and are fed into parallel or shared encoders of Swin Transformer that learn contextual associations between spatial areas and modalities [26]. The hierarchical architecture allows the local and global features to be gradually combined, which is beneficial to the process of representing the complex disease patterns that extend across a large anatomical volume. The decoder is similar to U-Net and recovers high-resolution segmentation maps with skip connections of similar stages of the transformer. To classify, the global token representations are pooled and are sent through dense layers. Although Swin-UNet is an efficient way to model global dependencies and intricate interactions of features, its high computation cost and the lack of explicit curriculum learning make it vulnerable to data scarcity and training instabilities in challenging medical data [27, 28].

3.3.4. CNN-transformer Hybrid Architecture

The CNN-Transformer Hybrid architecture combines the utility of both convolutional neural networks and transformer models to establish a local and global learning strategy for features. According to this paradigm, the first step of an encoder of modality-specific CNN is the derivation of robust low-level spatial features, *i.e.*, edge, texture, or local anatomical structure. These CNN properties are then fed into a transformer-based fusion module where there are self-attention mechanisms that model long-range dependencies and inter-modality dependencies. This hybrid architecture enables effective feature extraction and retrieves global contextual data that is important in classifying diseases and determining the accurate segmentation. The decoder recovers the segmentation outputs based on the attention-enhanced features, whereas the classification is done based on globally pooled transformer representations [8, 27, 29].

3.3.5. Multi-modal Attention-directed Curriculum Learning (MAGCL)

The proposed Multi-Modal Attention-Directed Curriculum Learning (MAGCL) method is presented as a new model of learning instead of a separate network structure. This paper deploys MAGCL on top of a strong CNN-Transformer hybrid backbone to improve the performance of segmentation as well as disease classification. The main idea behind MAGCL is to control the learning process by means of attention reliability as a supervisory signal. In training, the framework constantly compares spatial, channel, and modality-specific attention responses to approximate the difficulty of the samples. The early phases of training are given to samples with stable and well-localized attention over clinically significant areas, but more deliberate cases that can be described by uncertain or diffused attention patterns are gradually added. The attention-directed curriculum methodology stabilizes optimization and improves convergence, as well as allows the model to obtain strong representations using complex multi-modal inputs. Also, modality-wise attention weighting is introduced to MAGCL, which modulates the contribution of each modality of imaging in real time depending on its diagnostic utility with a given sample.

3.4. Attention-based Medical Imaging Models

Attention-based deep learning models have become highly effective in medical image analysis because they enable neural networks to selectively focus on diagnostically important anatomical regions while suppressing irrelevant background information. Unlike conventional convolutional neural networks that process all spatial features equally, attention mechanisms dynamically emphasize informative feature representations, resulting in improved segmentation precision, pathological localization, and disease classification performance. Spatial attention mechanisms enhance localization of tumors and abnormal tissues, whereas channel attention mechanisms prioritize discriminative semantic features associated with disease characteristics. Transformer-based self-attention models further improve contextual understanding by capturing long-range feature dependencies and global anatomical relationships within complex medical images.

Several recent studies have demonstrated that attention-guided architectures significantly improve feature fusion capability, optimization stability, and representation learning in multi-modal medical imaging systems. Attention U-Net, transformer-assisted segmentation frameworks, and modality-aware attention models have shown superior performance in identifying fine anatomical boundaries and small pathological structures compared with conventional encoder–decoder approaches. Inspired by these advancements, the proposed MAGCL framework integrates spatial-channel attention fusion with attention-guided curriculum learning to improve segmentation accuracy and disease classification reliability. The attention modules dynamically assign importance weights to modality-specific features, while attention confidence is utilized to progressively optimize the learning process from simpler to more challenging samples, resulting in robust and clinically meaningful feature representations.

3.5. CNN–transformer Hybrid Backbone

The CNN–Transformer hybrid backbone was selected because convolutional neural networks are highly effective in extracting local spatial information such as anatomical boundaries, textures, and pathological structures, whereas transformer models are capable of capturing long-range contextual dependencies and global feature relationships through self-attention mechanisms. The integration of CNNs and transformers enables simultaneous local-global representation learning, which is particularly beneficial for complex multi-modal medical imaging scenarios.

3.6. Activation Functions and Optimization Strategy

Rectified Linear Unit (ReLU) activation was employed in the convolutional encoder and decoder layers because of its computational efficiency and ability to reduce vanishing gradient problems during deep network optimization. Sigmoid activation was utilized in spatial and channel attention modules to generate normalized attention weights between 0 and 1, enabling adaptive feature weighting and attention refinement. Softmax activation was used in the disease classification head for multi-class probability prediction.

4. RESULTS AND DISCUSSION

4.1. Dataset Description

4.1.1. TCIA Dataset and Sample Distribution

The experimental dataset used in this research was collected from The Cancer Imaging Archive (TCIA), a publicly available repository containing de-identified medical imaging datasets for cancer-related studies. The dataset consists of approximately 3,200 multi-modal patient cases, where each case includes paired MRI and CT images along with pixel-level segmentation annotations and disease classification labels. The availability of multi-modal imaging data in TCIA makes it highly suitable for deep learning-based medical image segmentation, disease classification, and multi-modal feature learning experiments [30, 31].

4.1.2. Data Key Characteristics

- *Modalities:* CT, MRI (T1, T2, FLAIR, contrast-enhanced), PET, and occasionally X-ray, often available for the same patient.
- *Clinical Focus:* Cancer-related imaging, including brain, lung, liver, head and neck, breast, and prostate cancers.
- *Annotations:* Many cohorts include expert-annotated segmentations, tumor masks, radiotherapy structures, and clinical metadata.
- *Scale:* Individual TCIA cohorts range from tens to several thousand patient cases; combined cohorts can support large-scale studies (3,000 cases).
- *Standards:* Data are provided in DICOM format with standardized metadata, facilitating reproducible research.

4.2. Experimental Setup

The experimental evaluation of the proposed MAGCL framework was conducted using 3,200 multi-modal patient cases containing paired MRI and CT images with pixel-level segmentation annotations and disease labels. The dataset was divided into 70% training, 15% validation, and 15% testing subsets while maintaining balanced disease-category distribution. Standard preprocessing techniques, including image normalization, resizing, and augmentation, were applied to improve generalization and reduce overfitting. Performance evaluation was carried out using Dice score, IoU, precision, recall, F1-score, accuracy, AUC-ROC, and HD95 metrics.

4.3. Implementation Setup

The proposed MAGCL framework was implemented using Python and the PyTorch deep learning library. Training was performed on an NVIDIA GPU-enabled workstation using the Adam optimizer with an adaptive learning rate scheduling strategy. ReLU activation was employed in convolutional layers, while sigmoid and softmax functions were utilized for attention weighting and disease classification tasks, respectively. Early stopping and batch normalization techniques were incorporated to stabilize convergence and improve training efficiency.

4.4. Evaluation Metrics

All evaluation metrics employed in this study are concisely described in the following section to provide a clear understanding of their relevance and interpretability [7, 29, 32].

- **Precision** measures the proportion of correctly predicted positive cases among all predicted positives. It reflects the model's ability to avoid false positives, which is particularly important in medical diagnosis to reduce unnecessary interventions or anxiety caused by incorrect disease predictions.
- **The recall**, or sensitivity, measures the rate of correct recognition of the positive cases by the model. Recall is a very important issue in healthcare applications, and false negatives (a missed case of a true disease) may result in delayed treatment and poor clinical outcomes.
- **F1-score** is a harmonic average of precision and recall, which offers a balanced assessment in those cases when distributions of classes are not even. It provides one measure that can reflect the consistency of positive predictions as well as the capability of the model to identify cases of diseases.
- **Area under the ROC Curve** (AUC-ROC) is used to determine the capability of the model to differentiate between disease and non-disease cases at different levels of decision-making. The greater AUC-ROC, the greater the discriminative capacity and the stronger decision boundaries regardless of the class imbalance.
- **Accuracy** is the ratio of correct to all samples of each of the classes. Though useful in providing an overview of the overall performance, accuracy should be used concurrently with precision and recall, particularly on imbalanced medical data.

- **Dice Similarity Coefficient (Dice)** is a popular measure of the accuracy of medical image segmentation. It quantifies the spatial overlap between the predicted segmentation and the ground truth with a range of 0 to 1 with higher scores reflecting more agreement. Dice is capable of sensing things, especially the small structures, hence it is most suitable when evaluating the quality of segmentation of pathological regions in a medical image.
- **Intersection over Union (IoU),** or the Jaccard Index, is a measure of the ratio of the intersection to the union of the region predicted and the actual region detected. IoU is a more stringent measure compared to dice, because it punishes an over segment and under-segmentation. Increased IoU represents better boundary alignment and general segmentation accuracy.
- **Hausdorff Distance at 95th percentile (HD95)** measures the accuracy of the boundary by quantifying the difference between the predicted and the actual contour of segmentation. HD95 is also less sensitive to outliers and noise since the 95th percentile is used instead of the maximum distance. Reduced HD95 values represent increased contour alignment and higher anatomical consistency and are of particular importance to clinical applicability.

4.5. Results Data

A full comparative study of the proposed MAGCL framework with the state-of-the-art available systems on segmentation and classification, respectively, is provided in Table 1 and Table 2, respectively. Table 1 points out the performance of segmentation in terms of Dice, IoU, precision, recall, and HD95, and it is evident that MAGCL always shows better boundary delineation and anatomical accuracy. Table 2 presents the results of the classification in accordance with the criteria of accuracy, precision, recall, F1-score, and AUC-ROC, according to which MAGCL demonstrates the most solid discriminative properties and general predictive reliability, which proves its relevance as effective clinical decision support.

Transformer-based Swin-UNet also advances performance with long-range dependencies and global contextual interrelations of medical images. The fact that it increased its accuracy and F1-score supports its expertise in capturing complex pathological patterns cutting across large areas. These advantages are supplemented by the CNN-Transformer Hybrid model, which combines convolutional feature extraction with transformer-based global reasoning to create more robust feature representations, which in turn translate into higher recall and accuracy, particularly in those cases where the disease manifestations are subtle.

The MAGCL structure, suggested, continues to outperform any form of comparative models and attains the foremost values in all measures of evaluation. The high accuracy and recall testify to the effectiveness of the attention-directed learning in increasing the feature discrimination, whereas curriculum learning promotes consistent training and enhances generalization abilities. The subsequent high F1 -score and accuracy ratifies the strength and clinical correctness of MAGCL in performing multi-mode medical image classification activities.

The AUC-ROC findings provided in Fig. (3) provide a threshold-free evaluation of the performance of disease classification on the comparative models and the proposed MAGCL framework. When class balance and sensitivity to decision boundary are of primary concern (as is the case with a medical diagnostic application), then AUC-ROC is especially appropriate because its ability to rank positive instances of the disease over negative ones at a spectrum of decision thresholds.

Table 1. Comparative performance analysis of the proposed MAGCL framework and existing state-of-the-art models for multi-modal medical image classification on the TCIA dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
Multi-Modal U-Net	88.9	89.2	87.6	88.4	0.903
Attention U-Net	91.2	91.5	90.1	90.8	0.926
Swin-UNet	93.5	93.8	92.6	93.2	0.948
CNN–Transformer Hybrid	94.4	94.6	93.9	94.2	0.962
Proposed MAGCL	**95.9**	**96.1**	**95.4**	**95.7**	**0.978**

Table 2. Comparative performance analysis of the proposed MAGCL framework and existing state-of-the-art models on the TCIA dataset using segmentation evaluation metrics.

Model	Dice (%)	IoU (%)	Precision (%)	Recall (%)	HD95
Multi-Modal U-Net	88.1	79.6	89.2	87.4	14.2
Attention U-Net	90.4	82.8	91.5	90.1	11.6
Swin-UNet	92.9	86.7	93.8	92.6	9.4
CNN–Transformer Hybrid	94.1	88.9	94.6	93.9	8.1
Proposed MAGCL	**95.6**	**91.2**	**96.1**	**95.4**	**6.7**

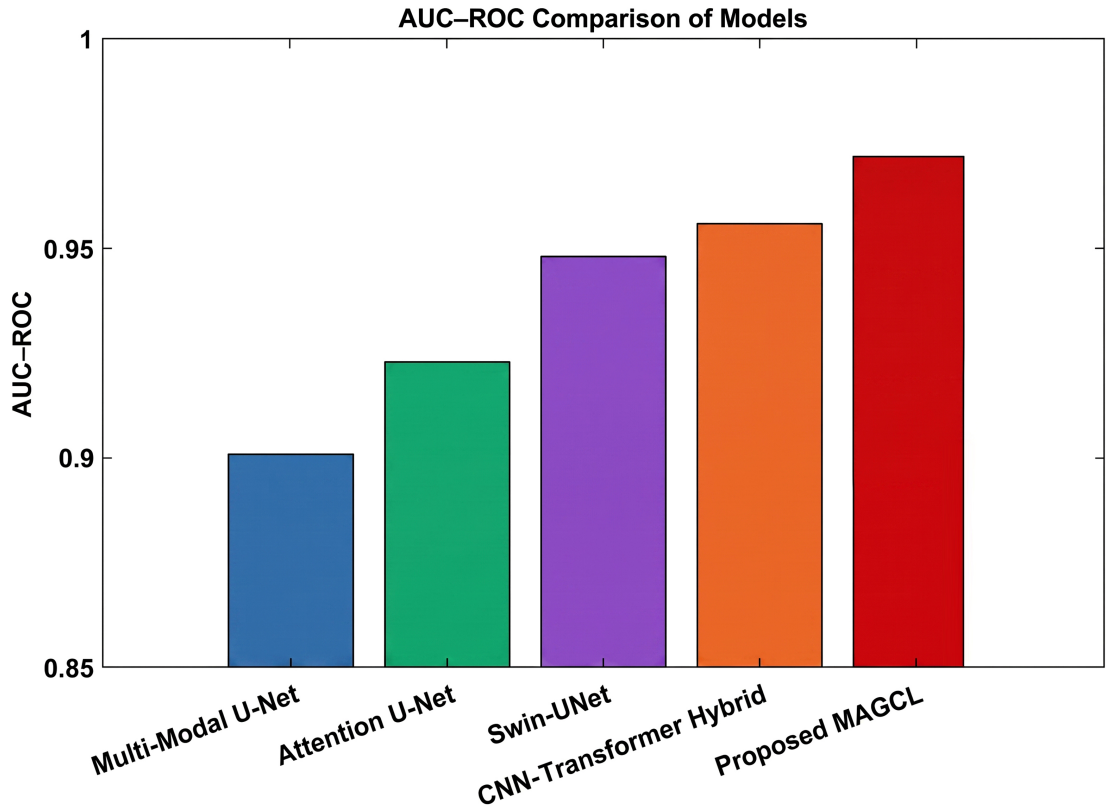

Fig. (3). Comparative performance analysis of proposed MAGCL and existing models on the TCIA dataset using AUC-ROC values.

Multi-Mode U-Net has the worst AUC-ROC, which indicates a limited discriminating capability in the case of the use of only convolutional feature extraction and low-level modality fusion. Though it is effective in the extraction of space information, lack of advanced attention mechanisms limits its ability to separate complex patterns of diseases, and therefore the ability to classify classes.

SwinUNet also enhances AUC-ROC by using transformer-based self-attention, which learns long-range dependence and global context-related relationships. The improvement will allow the network to record subtle disease features spread over the imaging domain, which will be more refined. The CNN-Transformer Hybrid model is also improved further by incorporating the local convolutional features into the global transformer representations, which leads to a stronger and more stable classification performance.

The MAGCL framework proposed has the largest AUC-ROC compared with all models assessed, which highlights its better discriminative ability on the disease and non-disease cases. It is the integration of attention-based feature learning and curriculum-based training that can help to create a stable decision boundary and progressive expansion of classification confidence, which confirms the ability of MAGCL to improve clinical reliability and diagnostic robustness.

The confusion matrices provide a detailed representation of the classification performance of each model that shows their ability to classify disease and non-disease cases correctly, as shown in Fig. (4). In the confusion matrices, D is diseased, and ND is non-diseased. The Multi-Mode U-Net has the highest number of misclassifications, and the proportions of false negatives and false positives are large, which can be attributed to its low discriminative strength, which is caused by the use of simple convolutional characteristics and a simple combination of modes, which may result in missed diagnoses and unnecessary alarms. The Attention U-Net is more successful in reducing both false negatives and false positives, and hence, proves the existence of attention mechanisms directed towards the diagnostically significant areas and the reduction of the noise in the background.

By adequately detecting global contextual features through transformer-based self-attention, the Swin-UNet has been found to increase classification reliability significantly, with a very high reduction in misclassified samples (especially false negatives), which is very important in medical diagnosis. The CNN-Transformer Hybrid model brings additional benefits by applying an improved combination of local feature extraction and global reasoning, which yields a more balanced error distribution and boosted true-positive and true-negative rates.

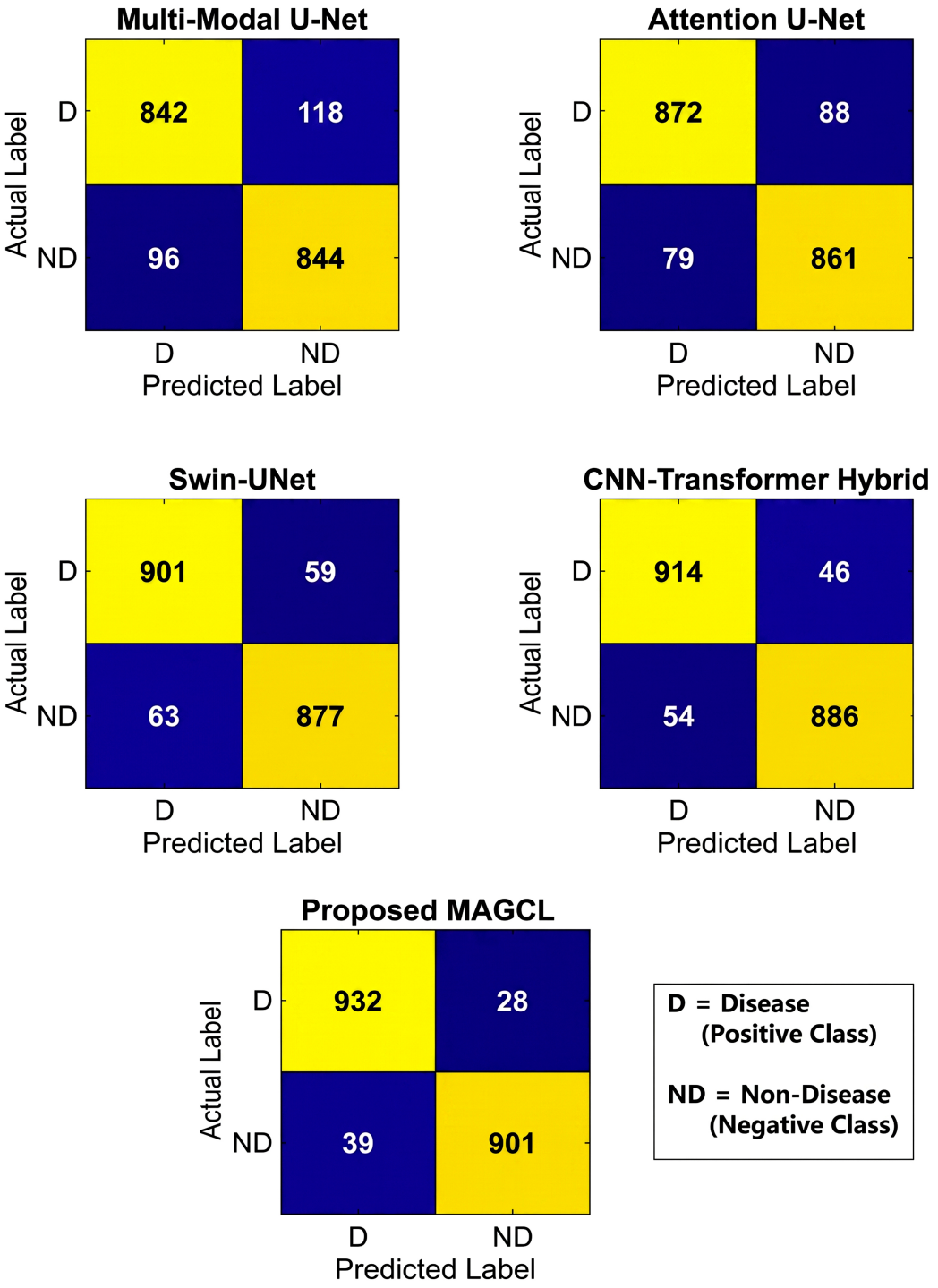

Fig. (4). Comparative performance analysis of proposed MAGCL and existing models on the TCIA dataset using confusion matrices.

The MAGCL framework suggested achieves the most preferred confusion matrix, having the largest number of disease and non-disease samples correctly classified, and the lowest count of the general error. The large reduction in false negatives points to greater sensitivity, and the reduction in false positives at the same time points to improved specificity. These results confirm that curriculum learning directed by attention creates solid decision margins and solid classification, which makes MAGCL a more reliable tool in the field of clinical decision support.

4.6. Results Analysis

The experimental evaluation demonstrates that the proposed MAGCL framework consistently outperforms conventional convolutional, attention-based, and transformer-based models across both segmentation and classification tasks. As reported in Tables **1** and **2**, MAGCL achieves better Dice, IoU, precision, recall, F1-score, accuracy, and AUC-ROC, indicating superior boundary delineation, robust feature discrimination, and reliable disease prediction.

To perform segmentation, the Dice and IoU scores indicates that MAGCL can accurately delineate the pathological areas, even in complex multi-modal situations. The decreased HD95 also confirms improved consistency in the boundaries and alignment in the anatomy, which is very essential to clinical interpretation. Such gains can be explained by the working mechanisms of spatial and channel attention that successfully inhibit background irrelevant information and focus on the structures of diagnostic relevance.

Regarding classification, MAGCL shows significant increases in recall and AUC-ROC, which indicates a better sensitivity and discriminative ability at both ends of the decision threshold. This is especially valuable in medical diagnosis, where it is essential to reduce false negatives. The guided strategy in curriculum learning of attention also helps achieve a more stable optimization, through gradually adding complex samples, also minimizing training instability and enhancing generalization.

Altogether, modality-aware attention, curriculum learning, and joint multi-task optimization synergistically contribute to making MAGCL clinically relevant compared to the currently used state-of-the-art techniques, which critically justify its effectiveness in providing a solid decision-support system in multi-modal medical image analysis.

4.7. Comparative Analysis with Existing Models on Same Dataset

To comprehensively evaluate the effectiveness of the proposed MAGCL framework, comparative experiments were conducted against several state-of-the-art segmentation and classification architectures, including Multi-Modal U-Net, Attention U-Net, Swin-UNet, and CNN–Transformer Hybrid models using the same TCIA-based multi-modal medical imaging dataset, depicted in Table **3**. The comparison was performed using standard evaluation metrics including Dice score, IoU, precision, recall, F1-score, accuracy, AUC-ROC, and HD95. The experimental findings indicate that the proposed MAGCL framework consistently achieved superior performance across both segmentation and disease classification tasks.

Conventional Multi-Modal U-Net demonstrated comparatively lower performance due to its inability to dynamically prioritize diagnostically relevant modalities and contextual features. Although Attention U-Net improved localization performance through attention gates, it still lacked global contextual modeling and adaptive curriculum optimization. Transformer-based Swin-UNet improved long-range dependency modeling and contextual representation learning; however, its performance remained sensitive to optimization instability and computational complexity in heterogeneous imaging environments. Similarly, the CNN–Transformer Hybrid model achieved improved global-local feature learning but lacked adaptive attention-guided curriculum scheduling for progressive optimization. In contrast, the proposed MAGCL framework effectively integrated modality-aware attention with curriculum learning, resulting in more stable convergence, improved feature discrimination, reduced segmentation error, and superior disease classification reliability.

4.8. Comparison with State-of-the-art Methods

To further validate the effectiveness of the proposed MAGCL framework, additional comparison was performed with recently published TCIA-based multi-modal medical image analysis studies reported in the literature. Table **4** summarizes the comparative performance of the proposed framework against recent state-of-the-art approaches developed for segmentation and disease classification using similar TCIA cohorts. The comparison demonstrates that the proposed MAGCL framework achieves superior segmentation accuracy, disease classification performance, and optimization stability compared with recently published methods.

The improved performance of MAGCL is mainly attributed to the synergistic integration of modality-aware attention mechanisms, adaptive feature fusion, and attention-guided curriculum learning. Unlike existing methods that primarily focus on either segmentation or classification independently, the proposed framework jointly optimizes both tasks within a unified end-to-end learning strategy. Additionally, the progressive curriculum learning mechanism stabilizes optimization by gradually introducing difficult samples based on attention confidence and uncertainty estimation, resulting in improved generalization capability and clinically meaningful feature representation learning.

Table 3. Comparative analysis of proposed MAGCL framework with existing models on TCIA dataset.

Model	Accuracy (%)	F1-Score (%)	Dice (%)	IoU (%)	AUC-ROC	HD95 ↓
Multi-Modal U-Net	88.9	88.4	88.1	79.6	0.903	14.2
Attention U-Net	91.2	90.8	90.4	82.8	0.926	11.6
Swin-UNet	93.5	93.2	92.9	86.7	0.948	9.4
CNN–Transformer Hybrid	94.4	94.2	94.1	88.9	0.962	8.1
Proposed MAGCL	**95.9**	**95.7**	**95.6**	**91.2**	**0.978**	**6.7**

Table 4. Comparison with recently published TCIA-based studies.

Study	Methodology	Dataset	Dice (%)	Accuracy (%)	AUC-ROC
Zhang *et al.* (2024) [1]	Dual-Attention Transformer Hybrid Network	TCIA Brain MRI	91.8	92.6	0.941
Ahmed *et al.* (2023) [18]	DoubleU-NetPlus	TCIA Multi-Modal Dataset	92.7	93.1	0.952
Wang *et al.* (2025) [26]	Multi-Scale Feature Fusion Network	TCIA Cancer Cohort	93.4	94.0	0.961
Proposed MAGCL	Attention-Guided Curriculum Learning	TCIA Multi-Modal Dataset	**95.6**	**95.9**	**0.978**

The comparative findings clearly demonstrate that the proposed MAGCL framework achieves higher segmentation precision, stronger disease discrimination capability, and improved optimization robustness compared with existing TCIA-based state-of-the-art methods. These results confirm the effectiveness and clinical applicability of the proposed attention-guided curriculum learning paradigm for intelligent multi-modal medical image analysis.

5. ABLATION STUDY: EFFECT OF CURRICULUM LEARNING IN MAGCL

To evaluate the contribution of the curriculum learning strategy, an ablation study was conducted by comparing the full MAGCL framework with a variant without curriculum learning (MAGCL without CL), where all training samples were introduced uniformly without difficulty-based scheduling. All other architectural components, including modality-aware attention and the CNN–Transformer backbone, were kept identical to ensure a fair comparison.

The results indicate that removing curriculum learning leads to noticeable performance degradation across all evaluation metrics. In particular, MAGCL without CL exhibits reduced Dice score, precision, recall, and AUC-ROC, along with less stable convergence during training. The absence of progressive sample introduction increases sensitivity to hard and ambiguous cases in early epochs, resulting in suboptimal feature learning and poorer generalization.

In contrast, the complete MAGCL model benefits from attention-guided curriculum learning by first focusing on high-confidence samples and gradually incorporating more complex cases. This strategy stabilizes optimization, enhances feature discrimination, and yields consistently superior segmentation and classification performance, confirming the critical role of curriculum learning in the proposed framework.

5.1. Limitations

- The proposed MAGCL framework is evaluated on a limited number of imaging modalities and disease categories, which may affect its generalizability to diverse clinical applications and rare pathological conditions.
- The inclusion of attention mechanisms and curriculum learning introduces additional computational overhead, increasing training time and resource requirements.
- The curriculum learning strategy depends on attention confidence and uncertainty estimation, which may be influenced by noisy labels or imperfect ground-truth annotations.
- The model has not been validated in real-time or resource-constrained clinical environments, limiting conclusions about its deployment feasibility.
- Extensive prospective clinical validation and interpretability assessment are still required to establish clinical trust and practical adoption.

CONCLUSION AND FUTURE RESEARCH

Conclusion

This study presented a novel Multi-Modal Attention-Guided Curriculum Learning (MAGCL) framework for simultaneous medical image segmentation and disease classification using heterogeneous imaging modalities. The proposed framework effectively integrates modality-aware attention mechanisms, curriculum learning, and CNN–Transformer hybrid learning to improve feature fusion, optimization stability, and diagnostic reliability in complex medical imaging environments. Experimental evaluation conducted on 3,200 paired MRI and CT patient cases demonstrated that the proposed MAGCL framework consistently outperformed Multi-Modal U-Net, Attention U-Net, Swin-UNet, and CNN–Transformer Hybrid models

The proposed framework achieved 95.9% classification accuracy, 95.7% F1-score, 95.6% Dice score, 91.2% Intersection over Union (IoU), and 0.978 AUC-ROC, indicating superior segmentation precision and disease discrimination capability. The integration of spatial-channel attention and progressive curriculum learning significantly improved pathological feature localization, segmentation boundary delineation, and optimization convergence stability. These improvements demonstrate the strong clinical potential of the proposed framework for intelligent computer-assisted diagnosis, radiological interpretation, treatment planning, and clinical decision-support systems.

Despite the promising performance, the proposed framework introduces relatively higher computational complexity due to the integration of multi-modal attention and transformer-based feature learning. Additionally, the framework was evaluated primarily on publicly available datasets, which may not fully represent diverse real-world clinical scenarios and imaging variations across institutions. Therefore, further validation on large-scale multi-center datasets is necessary to improve robustness and generalizability.

Future Research Directions

Future research will focus on developing lightweight and computationally efficient architectures for real-time clinical deployment, integrating explainable artificial

intelligence techniques for improved interpretability, and extending the framework toward federated and self-supervised learning paradigms. Additional investigation into multi-organ segmentation, longitudinal disease progression analysis, and cross-modal adaptive learning may further enhance the clinical applicability of the proposed MAGCL framework in next-generation intelligent healthcare systems.

AUTHORS' CONTRIBUTIONS

The authors confirm their contribution to the paper as follows: S.K.M.: Study conception and design, methodology, writing original draft preparation; S.D.: Data analysis, validation, writing, reviewing and editing; R.P.: Supervision, visualization, manuscript review and final editing. All authors reviewed the results and approved the final version of the manuscript.

LIST OF ABBREVIATIONS

MAGCL	=	Multi-Modal Attention-Guided Curriculum Learning
CNN	=	Convolutional Neural Network
MRI	=	Magnetic Resonance Imaging
CT	=	Computed Tomography
IoU	=	Intersection over Union
AUC	=	Area Under Curve
ROC	=	Receiver Operating Characteristic
HD95	=	95th Percentile Hausdorff Distance
ReLU	=	Rectified Linear Unit
AI	=	Artificial Intelligence
DL	=	Deep Learning
ML	=	Machine Learning
TCIA	=	The Cancer Imaging Archive
U-Net	=	U-Shaped Convolutional Network
F1-Score	=	Harmonic Mean of Precision and Recall
GPU	=	Graphics Processing Unit
ROI	=	Region of Interest
PACS	=	Picture Archiving and Communication System
CAD	=	Computer-Aided Diagnosis
DNN	=	Deep Neural Network
LSTM	=	Long Short-Term Memory
GAN	=	Generative Adversarial Network
SVM	=	Support Vector Machine
BN	=	Batch Normalization
PDF	=	Probability Density Function
TP	=	True Positive
TN	=	True Negative
FP	=	False Positive
FN	=	False Negative
CNN–Transformer	=	Convolutional Neural Network and Transformer Hybrid Architecture

CONSENT FOR PUBLICATION

Not applicable.

STANDARDS OF REPORTING

TRIPOD-AI guidelines were followed.

AVAILABILITY OF DATA AND MATERIALS

The data supporting the findings of this article is available in The Cancer Imaging Archive (TCIA) at https://www.cancerimagingarchive.net/. The specific collections used in this study are cited in the manuscript (see Section 4.1.1). TCIA is a publicly accessible repository for de-identified medical imaging datasets, and the data used in this work are openly available under the TCIA data usage policy [30, 31].

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] M. Zhang, Y. Zhang, S. Liu, Y. Han, H. Cao, and B. Qiao, "Dual-attention transformer-based hybrid network for multi-modal medical image segmentation", *Sci. Rep.*, vol. 14, no. 1, p. 25704, 2024. [http://dx.doi.org/10.1038/s41598-024-76234-y] [PMID: 39465274]

[2] Z. Wang, W. Wang, N. Li, S. Zhang, Q. Chen, and Z. Jiang, "Multimodal parallel attention network for medical image segmentation", *Image Vis. Comput.*, vol. 147, p. 105069, 2024. [http://dx.doi.org/10.1016/j.imavis.2024.105069]

[3] B. Mirab Golkhatmi, M. Houshmand, and S.A. Hosseini, "A multi-scale attention-based Swin transformer model for medical images segmentation", *Sci. Rep.*, vol. 15, no. 1, p. 38893, 2025. [http://dx.doi.org/10.1038/s41598-025-22649-0] [PMID: 41198764]

[4] L. He, L. Luan, and D. Hu, "Deep learning-based image classification for AI-assisted integration of pathology and radiology in medical imaging", *Front. Med.*, vol. 12, p. 1574514, 2025. [http://dx.doi.org/10.3389/fmed.2025.1574514] [PMID: 40529155]

[5] S. Khader, and R. Ghnemat, "AGU-Net: Advancing medical image segmentation with attention-guided U-Net architecture", *Cluster Comput.*, vol. 28, no. 16, p. 1074, 2025. [http://dx.doi.org/10.1007/s10586-025-05781-4]

[6] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Thorax disease classification with attention guided convolutional neural network", *Pattern Recognit. Lett.*, vol. 131, pp. 38-45, 2020. [http://dx.doi.org/10.1016/j.patrec.2019.11.040]

[7] X. Chen, X. Lin, Q. Shen, and X. Qian, "Combined Spiral

Transformation and Model-Driven Multi-Modal Deep Learning Scheme for Automatic Prediction of TP53 Mutation in Pancreatic Cancer", *IEEE Trans. Med. Imaging,* vol. 40, no. 2, pp. 735-747, 2021.
[http://dx.doi.org/10.1109/TMI.2020.3035789] [PMID: 33147142]

[8] K.R.M. Fernando, and C.P. Tsokos, "Deep and statistical learning in biomedical imaging: State of the art in 3D MRI brain tumor segmentation", *Inf. Fusion,* vol. 92, pp. 450-465, 2023.
[http://dx.doi.org/10.1016/j.inffus.2022.12.013]

[9] H. Wu, W. Min, D. Gai, Z. Huang, Y. Geng, Q. Wang, and R. Chen, "HD-Former: A hierarchical dependency Transformer for medical image segmentation", *Comput. Biol. Med.,* vol. 178, p. 108671, 2024.
[http://dx.doi.org/10.1016/j.compbiomed.2024.108671] [PMID: 38870721]

[10] T. Mahmood, A. Rehman, T. Saba, L. Nadeem, and S.A.O. Bahaj, "Recent Advancements and Future Prospects in Active Deep Learning for Medical Image Segmentation and Classification", *IEEE Access,* vol. 11, pp. 113623-113652, 2023.
[http://dx.doi.org/10.1109/ACCESS.2023.3313977]

[11] Z. Sabir, S. Dirani, S. Bou Saleh, M.K. Mabsout, and A. Arbi, "A novel radial basis and sigmoid neural network combination to solve the human immunodeficiency virus system in cancer patients", *Mathematics,* vol. 12, no. 16, p. 2490, 2024.
[http://dx.doi.org/10.3390/math12162490]

[12] Z. Sabir, A. Arbi, A.F. Hashem, and M.A. Abdelkawy, "Morlet wavelet neural network investigations to present the numerical investigations of the prediction differential model", *Mathematics,* vol. 11, no. 21, p. 4480, 2023.
[http://dx.doi.org/10.3390/math11214480]

[13] Z. Sabir, A. Hashem, A. Arbi, and M. Abdelkawy, "Designing a Bayesian regularization approach to solve the fractional Layla and Majnun system", *Mathematics,* vol. 11, no. 17, p. 3792, 2023.
[http://dx.doi.org/10.3390/math11173792]

[14] Z. Sabir, M.A.Z. Rajab, S.E. Alhazmi, M. Gupta, A. Arbi, and I.A. Baba, "Applications of artificial neural network to solve the nonlinear COVID-19 mathematical model based on the dynamics of SIQ", *J. Appl. Math. Comput. Mech.,* vol. 21, no. 4, pp. 874-884, 2022.
[http://dx.doi.org/10.1080/16583655.2022.2119734]

[15] A. Arbi, J. Cao, and A. Alsaedi, "Improved synchronization analysis of competitive neural networks with time-varying delays", *Nonlinear Anal.,* vol. 23, no. 1, pp. 82-107, 2018.
[http://dx.doi.org/10.15388/NA.2018.1.7]

[16] Z. Sabir, M.A. Mehmood, M. Umar, S. Salahshour, Y. Altun, A. Arbi, and M.R. Ali, "A numerical treatment through Bayesian regularization neural network for the chickenpox disease model", *Comput. Biol. Med.,* vol. 187, p. 109807, 2025.
[http://dx.doi.org/10.1016/j.compbiomed.2025.109807] [PMID: 39923585]

[17] J. Liu, S. Mao, and L. Pan, "Attention-based two-branch hybrid fusion network for medical image segmentation", *Appl. Sci.,* vol. 14, no. 10, p. 4073, 2024.
[http://dx.doi.org/10.3390/app14104073]

[18] M.R. Ahmed, A.F. Ashrafi, R.U. Ahmed, S. Shatabda, A.K.M.M. Islam, and S. Islam, "DoubleU-NetPlus: A novel attention and context-guided dual U-Net with multi-scale residual feature fusion network for semantic segmentation of medical images", *Neural Comput. Appl.,* vol. 35, no. 19, pp. 14379-14401, 2023.
[http://dx.doi.org/10.1007/s00521-023-08493-1]

[19] Y. Ruiping, L. Kun, X. Shaohua, Y. Jian, and Z. Zhen, "ViT-UperNet: A hybrid vision transformer with unified-perceptual-parsing network for medical image segmentation", *Complex Intell. Syst.,* vol. 10, no. 3, pp. 3819-3831, 2024.
[http://dx.doi.org/10.1007/s40747-024-01359-6]

[20] J. Zhang, F. Li, X. Zhang, Y. Cheng, and X. Hei, "Multi-task mean teacher medical image segmentation based on Swin Transformer", *Appl. Sci.,* vol. 14, no. 7, p. 2986, 2024.
[http://dx.doi.org/10.3390/app14072986]

[21] Y. Gao, Y. Jiang, Y. Peng, F. Yuan, X. Zhang, and J. Wang, "Medical image segmentation: A comprehensive review of deep learning-based methods", *Tomography,* vol. 11, no. 5, p. 52, 2025.
[http://dx.doi.org/10.3390/tomography11050052] [PMID: 40423254]

[22] C. Wang, M. Jiang, Y. Li, B. Wei, Y. Li, P. Wang, and G. Yang, "MP-FocalUNet: Multiscale parallel focal self-attention U-Net for medical image segmentation", *Comput. Methods Programs Biomed.,* vol. 260, p. 108562, 2025.
[http://dx.doi.org/10.1016/j.cmpb.2024.108562] [PMID: 39675195]

[23] Z. Liu, Y. Cheng, and S. Tamura, "Multi-label local to global learning: A novel learning paradigm for chest x-ray abnormality classification", *IEEE J. Biomed. Health Inform.,* vol. 27, no. 9, pp. 4409-4420, 2023.
[http://dx.doi.org/10.1109/JBHI.2023.3281466] [PMID: 37252867]

[24] J. Wang, Q.M.J. Wu, and F. Pourpanah, "An attentive-based generative model for medical image synthesis", *Int. J. Mach. Learn. Cybern.,* vol. 14, no. 11, pp. 3897-3910, 2023.
[http://dx.doi.org/10.1007/s13042-023-01871-0]

[25] Y. Zhang, G. Balestra, K. Zhang, J. Wang, S. Rosati, and V. Giannini, "MultiTrans: Multi-branch transformer network for medical image segmentation", *Comput. Methods Programs Biomed.,* vol. 254, p. 108280, 2024.
[http://dx.doi.org/10.1016/j.cmpb.2024.108280] [PMID: 38878361]

[26] Y. Wang, H. Li, G. Liu, J. Huo, J. Sun, and Y. Zhang, "A Multi-Scale Feature Interaction and Fusion Medical Image Segmentation Method", *Int. J. Imaging Syst. Technol.,* vol. 35, no. 5, p. e70207, 2025.
[http://dx.doi.org/10.1002/ima.70207]

[27] Z. Ji, Z. Chen, and X. Ma, "Grouped multi-scale vision transformer for medical image segmentation", *Sci. Rep.,* vol. 15, no. 1, p. 11122, 2025.
[http://dx.doi.org/10.1038/s41598-025-95361-8] [PMID: 40169823]

[28] J. Chen, Z. Liang, and X. Lu, "A dual attention and cross layer fusion network with a hybrid CNN and transformer architecture for medical image segmentation", *Sci. Rep.,* vol. 15, no. 1, p. 35707, 2025.
[http://dx.doi.org/10.1038/s41598-025-19563-w] [PMID: 41083525]

[29] F. Prior, K. Smith, A. Sharma, J. Kirby, L. Tarbox, K. Clark, W. Bennett, T. Nolan, J. Freymann, and B. Vendt, "The public cancer radiology imaging collections of The Cancer Imaging Archive", *Sci. Data,* vol. 4, no. 1, p. 170124, 2017.
[http://dx.doi.org/10.1038/sdata.2017.124] [PMID: 28925987]

[30] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository", *J. Digit. Imaging,* vol. 26, no. 6, pp. 1045-1057, 2013. https://www.cancerimagingarchive.net/
[http://dx.doi.org/10.1007/s10278-013-9622-7] [PMID: 23884657]

[31] G. Sun, Y. Pan, W. Kong, Z. Xu, J. Ma, T. Racharak, L.M. Nguyen, and J. Xin, "DA-TransUNet: Integrating spatial and channel dual attention with transformer U-net for medical image segmentation", *Front. Bioeng. Biotechnol.,* vol. 12, p. 1398237, 2024.
[http://dx.doi.org/10.3389/fbioe.2024.1398237] [PMID: 38827037]

[32] Q. Wang, and Y. Xue, "Multi-scheme cross-level attention embedded U-shape transformer for MRI semantic segmentation", *Sci. Rep.,* vol. 15, no. 1, p. 22891, 2025.
[http://dx.doi.org/10.1038/s41598-025-06966-y] [PMID: 40594599]