International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 482-498

https://doi.org/10.70517/ijhsa46444

Innovative Application of Interpretable Artificial Intelligence Based on Image Processing Algorithms in Medical Image Detection

Li Gong[1,*]
[1] Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, Universiti of Malaya, Kuala Lumpur, 50603, Malaysia
Corresponding authors: (e-mail: gongl24121@gmail.com).

Abstract Medical image detection plays a crucial role in disease diagnosis; however, traditional manual interpretation is often limited by subjectivity and low efficiency. Interpretable artificial intelligence (AI) techniques, grounded in image processing algorithms, demonstrate strong potential for enhancing objectivity and reliability in this domain. In this study, we propose an enhanced YOLOv12 algorithm that integrates both attention mechanisms and a residual feedback structure. Combined with a selective contextual transformer-enhanced SCTNet segmentation network, these components form a unified and intelligent medical image processing system. Methodologically, the AGs-ECSA hybrid attention module enhances feature extraction by incorporating Efficient Channel Attention (ECA) and spatial attention mechanisms. A loopback residual structure is introduced to preserve original feature information, while bounding box regression is improved using the Complete Intersection over Union (CIoU) loss function. Additionally, the Selective Contextual Transformer (SCT) module captures both local and global semantic dependencies.Experimental results on the DeepLesion dataset demonstrate that the proposed method achieves an average sensitivity of 88.97%, surpassing the strong baseline MULAN by 0.46% and consistently achieving higher detection accuracy at lower false positive rates. On the GLAS segmentation dataset, SCTNet achieves an mF1 score of 0.8285 and an mIoU of 0.7246, representing improvements of 4.98% and 8.64%, respectively, over existing mainstream methods. The system was further validated on cerebral hemorrhage CT scans, accurately localizing lesions and estimating their size. These findings demonstrate the effectiveness of interpretable AI in medical image detection and offer a reliable framework to support clinical diagnosis.

Index Terms Medical image detection, interpretable artificial intelligence, attention mechanism, image segmentation, deep learning, lesion detection

I. Introduction

In recent years, with the continuous development of China's medical care, the automation and intelligent detection technology of medical images have received more attention [1]. The combination of deep learning technology and medical image analysis has become one of the research hotspots in the field of intelligent medical treatment and has a broad application prospect [2], [3]. On the one hand, intelligent recognition of medical images can reduce the work pressure of medical workers, improve the efficiency of medical detection, and alleviate the problem of limited access to medical resources [4], [5]. On the other hand, automated detection of medical images can improve its recognition level, prevent misdiagnosis due to the lack of energy and cognitive level of medical staff, so as to improve the medical experience of patients [6]. However, at this stage, medical imaging still cannot be applied to the clinic. This is because the current accumulation of medical image data for specific diseases is insufficient, chaotic structure, lack of unified data interface, can not be directly applied to the existing image processing and recognition algorithms [7], [8].

With the development of medical imaging technology, medical imaging techniques such as PET, CT, Chest Photography (X-ray), Magnetic Resonance Imaging (MRI) and Single Photon Emission Computed Tomography (SPECT) have been rapidly developed [9], [10]. Medical image processing algorithms, mainly include traditional medical image processing algorithms and deep learning image processing algorithms [11]. For example, Houssein et al. proposed the Golden Jackal Optimization (GJO) algorithm applied to medical image detection of skin cancer, which solves a multilayer thresholding problem using the opposing Golden Jackal Optimizer (IGJO) with Otsu's method as the objective function [12]. As another example, Nadimi-Shahraki et al. proposed an Enhanced Whale Optimization Algorithm (E-WOA) and its binary version (BE-WOA) to improve the feature selection and enhance the efficiency of COVID-19 detection in order to cope with the huge COVID-19 lung infection medical images [13].

Unified algorithms usually require manual design of features manually, and traditional algorithms are limited by the complexity and diversity of medical images, resulting in traditional medical image processing algorithms, which are no longer applicable to the complex medical images of today.

With the development of deep learning techniques, deep learning image processing algorithms have achieved great success in various fields, especially in the field of medical images [14]. For example, Rammurthy et al. proposed a technique combining Whale-Red-headed Hawk Optimization Algorithm (WHHO) with Deep Convolutional Neural Networks (DeepCNN) for brain tumor detection in MRI images, which has very high accuracy, specificity, and sensitivity [15]. Yazhinian et al. proposed a technique called Whale Optimization Driven Generative Convolutional Neural Network (WO-GCNN) medical image detection framework for detecting anemia to address the data scarcity problem with potential for early identification and improved patient care [16]. Anand et al. used Recurrent Neural Network (RNN) algorithm for MRI images and compared them with existing datasets based on age and gender criteria to improve the brain's intracerebral meningioma tumor detection accuracy, thus meeting the growing demand for effective tumor identification methods [17]. Salehin et al. introduced a novel customized convolutional neural network model (MedvCNN), combined with a long short-term memory network (LSTM)), to construct an efficient medical image detection method aimed at improving the performance of medical image classification, thus improving the efficiency and accuracy [18]. Liu et al. proposed a multiscale CNN for GGO 3D lung CT image detection that fuses different scale features with convolutional kernels of multiple sizes to capture features at each scale [19]. These methods can automatically recognize unimodal medical images for assisted diagnosis and greatly improve the efficiency of image processing [20].

As an important part of the modern medical system, medical imaging diagnosis plays an irreplaceable role in the early detection of diseases, the development of treatment programs and prognosis assessment. The wide application of medical imaging technologies, such as computed tomography, magnetic resonance imaging, and ultrasound imaging, provides clinicians with rich anatomical structure and pathology information. However, the traditional manual way of reading films faces many challenges: on the one hand, the volume of medical imaging data is growing exponentially, and the long training cycle and relative shortage of professional imaging physicians lead to a mismatch between diagnostic efficiency and the growth rate of imaging data; on the other hand, there are subjective differences in manual diagnosis, and the experience levels and knowledge backgrounds of different physicians may lead to inconsistencies in diagnostic results; on the other hand, manual diagnosis is prone to subjective variation, particularly in detecting early-stage or complex lesions, and different physicians' experience level and knowledge background may lead to inconsistent diagnostic results, especially for early lesions or complex cases. In addition, prolonged reading of films is prone to visual fatigue, which may affect the quality of diagnosis. The breakthrough development of artificial intelligence technology, especially deep learning algorithms, provides a new technical path to solve these problems. Image recognition technology based on convolutional neural networks has achieved remarkable success in the field of natural image processing, and its strong ability in feature extraction, pattern recognition and classification tasks has brought attention to its application prospects in the field of medical image analysis. Interpretable AI, as an important direction in the development of AI technology, not only requires algorithms to have high accuracy, but also emphasizes the transparency and understandability of the algorithmic decision-making process, which is particularly important for medical application scenarios.

Building upon this context, the present study proposes a unified technical framework to enhance intelligent medical image analysis from both detection and segmentation perspectives. To improve the accuracy of lesion localization, we refine the YOLOv8 object detection architecture by incorporating a multi-level attention mechanism and a loopback residual feedback structure, enabling better representation of complex lesion features. For segmentation tasks, we design a selective contextual transformer (SCT)-enhanced network aimed at capturing both global context and fine-grained spatial structures, thereby improving semantic segmentation performance. These two components are seamlessly integrated into an end-to-end medical image processing system capable of automated lesion identification and delineation. The proposed approach is rigorously evaluated across multiple publicly available datasets to validate its generalization capability and clinical applicability.

II. Medical image detection based on attention mechanism and YOLOv12 improvement

The YOLOv12 framework [21] represents a significant advancement over previous YOLO iterations by incorporating a transformer-based backbone [22], a fully anchor-free detection head, and an enhanced cross-scale feature aggregation module [23]. These architectural improvements make it particularly effective for detecting small, dense, and irregularly shaped targets—challenges that are commonly encountered in medical image analysis [24].

However, lesion detection in clinical imaging involves additional complexity due to the fuzzy morphology, low contrast, and highly variable anatomical positioning of pathological regions. To better address these challenges, we propose an improved medical lesion detection network based on YOLOv12. Specifically, the model integrates both

spatial and channel attention mechanisms [25] along with attention gates (AGs) to selectively enhance the representation of lesion-relevant features while suppressing irrelevant background noise. In addition, a residual feedback structure [26] is introduced to improve semantic information preservation and facilitate more stable gradient propagation. These enhancements collectively enable the network to more accurately identify complex lesion regions in high-resolution medical images.

II. A.Attention mechanisms

II. A. 1) Channel Attention Mechanisms

In channel attention computation, 3D feature maps are first aggregated into one-dimensional descriptors using global average pooling. These descriptors are then passed through a downscaling and activation process via fully connected (FC) layers to compute the attention weights. Finally, the weights are reshaped and broadcasted to match the original feature dimensions.

SENet [27] utilizes two fully connected layers to capture non-linear cross-channel dependencies. However, the dimensionality reduction step in SE blocks can result in a loss of inter-channel correlation, potentially limiting the model's capacity to capture fine-grained channel relationships.

To address this limitation, the Efficient Channel Attention (ECA) module omits dimensionality reduction and instead models local cross-channel interaction by applying a one-dimensional convolution of size

k across the channel descriptors obtained via global average pooling. This approach enables each channel to attend to its immediate neighbors while maintaining the original feature dimensionality. As a result, the ECA module achieves a more favorable balance between performance and model complexity, enhancing attention precision without introducing significant computational overhead.

II. A. 2) Spatial attention mechanisms

Although the YOLOv12 framework [21] incorporates a transformer-based backbone to capture long-range dependencies, it lacks explicit spatial attention mechanisms for emphasizing localized features—an aspect that is particularly critical in medical image analysis, where subtle spatial variations can be diagnostically significant.

To address this limitation, we augment the YOLOv12 architecture with a lightweight spatial attention module designed to enhance the representation of salient anatomical regions. Specifically, average pooling and max pooling operations are first applied along the channel axis to produce two complementary spatial descriptors of size $H \times W \times 1$, encoding both global and salient contextual information. These descriptors are concatenated to form a feature map of size $H \times W \times 2$, which is then passed through a 7×7 convolution to learn an adaptive attention map over spatial positions.

The resulting attention map is used to reweight the original input feature map via element-wise multiplication, thereby enhancing the model's sensitivity to lesion-relevant areas while suppressing irrelevant background. This spatial attention augmentation allows the detection network to better localize lesions with irregular boundaries or low contrast—common challenges in CT and MRI datasets.

When integrated into the YOLOv12 backbone, this spatial attention mechanism complements the transformer's global context modeling by introducing fine-grained spatial discrimination, leading to more accurate lesion detection.

II. A. 3) Improved hybrid attention module

According to the attention from the channel or spatial dimension introduced above, of course, it is also possible to use both channel and spatial attention mechanisms, and CBAM [28] is a representative network among them, which combines channel and spatial attention in a sequential manner to improve the detection accuracy from both perspectives.

This chapter draws on the CBAM idea and proposes a hybrid attention module. The channel attention mechanism ECA and spatial attention mechanism above are concatenated to form the channel spatial attention mechanism (ECSA), and combined with attention gates (AGs) to form the attention module AGs-ECSA.

II. B.YOLOv8 detection algorithm incorporating attention and residual feedback

II. B. 1) Loopback residual module

Earlier versions of YOLO, such as YOLOv3, employed DBL blocks (consisting of convolution, batch normalization, and LeakyReLU activation) as fundamental building units, with each block typically comprising five stacked DBLs. While effective at the time, this architecture exhibited limitations in modeling capacity and feature reuse, particularly when applied to complex detection tasks like medical lesion identification.

Building on the limitations of YOLOv3, the YOLOv8 architecture introduced several notable improvements that significantly enhanced its suitability for real-time object detection. It adopts an anchor-free detection head, eliminating the need for manual anchor box configuration, and leverages a C2f-based lightweight backbone derived

from YOLOv7's ELAN design to facilitate more efficient feature reuse and stable gradient propagation. The model also includes the SPPF (Spatial Pyramid Pooling - Fast) module to improve multi-scale feature aggregation without introducing excessive computational overhead.

As illustrated in Figure 1, YOLOv8's architecture supports a decoupled detection head, enabling the separate optimization of object classification and bounding box regression. These architectural refinements result in faster inference speeds and competitive accuracy across various detection benchmarks. However, when applied to medical imaging—where lesion boundaries are often ambiguous, sizes vary widely, and visual contrast is low—YOLOv8 may still struggle to capture fine-grained spatial dependencies and retain high-level semantic cues, particularly for small or irregularly shaped targets. These limitations highlight the need for more expressive backbones and specialized attention mechanisms.

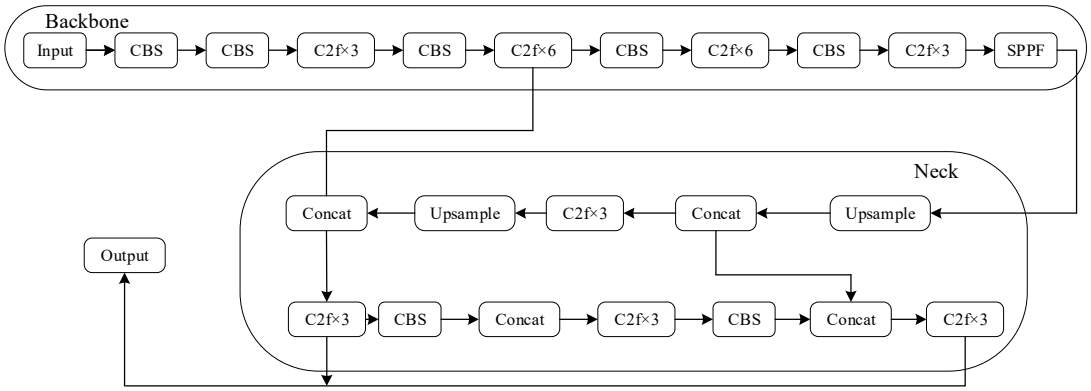

Figure 1: YOLOv8 network framework diagram

To address these limitations, the present study adopts the advanced YOLOv12 detection framework, which introduces several architectural enhancements over its predecessors. YOLOv12 integrates a transformer-enhanced backbone, a fully anchor-free detection head, and a refined multi-scale feature aggregation strategy, making it well-suited for detecting small, ambiguous, and morphologically diverse lesions common in medical imaging. In contrast to YOLOv3's reliance on manually defined anchor boxes and fixed receptive fields, YOLOv12 provides an end-to-end trainable pipeline that simplifies optimization and improves generalization across varied clinical datasets.

Figure 2 illustrates the overall architecture of the YOLOv12 backbone, which incorporates a refined C2f module based on cross-stage partial connections (CSP). Derived from the ELAN strategy used in YOLOv7, the C2f module improves gradient flow and computational efficiency by deepening network pathways without increasing model complexity. This design ensures sufficient feature reuse and improved convergence while maintaining a lightweight structure—essential for real-time medical image processing. At the end of the backbone, the network retains the SPPF (Spatial Pyramid Pooling—Fast) module, which enhances receptive field diversity and maintains robust detection performance across varying lesion scales.

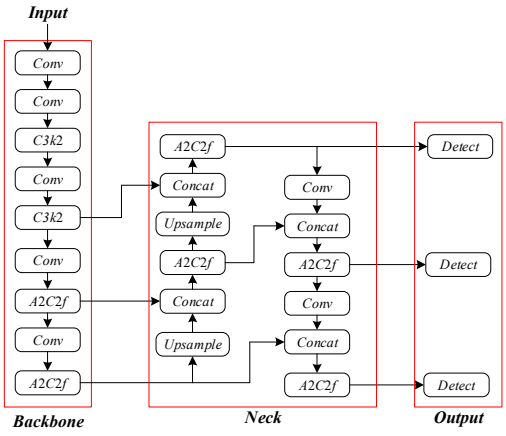

Figure 2: YOLOv12 network framework diagram

To further improve the model's ability to retain fine-grained features and stabilize training, we incorporate a loopback residual structure. This design introduces residual feedback by reintegrating the output of intermediate feature maps into earlier network stages. Such feedback not only reinforces important semantic cues but also facilitates gradient propagation in deep layers, thereby preserving lesion-related information that may otherwise degrade in depth. This mechanism proves particularly beneficial in medical contexts, where subtle image features—such as texture variation or lesion boundary irregularity—are crucial for accurate diagnosis.

We propose a novel loopback residual structure that combines residual propagation with residual feedback. It enhances feature retention by integrating input and mapped features through skip connections, allowing the network to preserve more original information. The residual feedback mechanism enhances the extracted features by adding the learned features to the input information and performing the feature extraction again and the output is defined as equation (1):

$$y_s = G(y_f) + x \tag{1}$$

x is the input vector of the module; the function G represents the residual mapping; y_f and y_s are the output vectors of the first residual propagation and the enhancement vectors after residual feedback, respectively.

In medical human lesion detection, the loopback residual block is connected by jumps to retain the original medical human lesion features; the feature information is consolidated through the loopback connection, so that the medical human lesion region and other regions of the image have obvious differences, to ensure that the discriminative ability of the image features will not become weaker with the increase of the depth of the network, and to make the network have a better and more stable performance of detecting medical human lesions.

II. B. 2) Improved network structure

In this section, we enhance the conventional Feature Pyramid Network (FPN) by incorporating additional semantic features extracted from the Darknet-53 backbone after two downsampling operations. This modification enriches the multi-scale representation by leveraging deeper contextual information. To further strengthen the network's ability to capture lesion-specific cues, we embed the AGs-ECSA attention module into the deepest feature level. This integration enables more focused and effective extraction of high-level semantic features associated with pathological regions.

II. B. 3) Improved loss function

The loss function is very important for the training of the network model, which determines the subsequent optimization direction of the network. The loss function in the YOLOv12 algorithm contains three parts: regression loss, classification loss and confidence loss. Among them, when calculating the regression loss, the positive and negative samples are determined based on the IoU values of the prediction frame and the real frame, i.e., the overlap between the prediction frame and the real frame is measured, and the calculation formula is shown in equation (2):

$$IoU_{(A,B)} = \frac{A \cap B}{AUB} \tag{2}$$

A and B are the coordinates of the predicted and true frames, respectively, $A \cap B$ denotes the intersection of the two, and $A \cup B$ denotes the concatenation of the two. The network usually carries out the following steps when classifying the samples, the first step needs to determine the location of the grid cell that contains the center point of the target, calculate the IoU value between all the a priori frames that contain this grid and the target location, and select the a priori frame that has the largest IoU of these values, that is, the a priori frame is used to predict this target, then this a priori frame is a positive sample. In the second step after the above calculation for the IoU between the remaining a priori frames that are not selected and all the targets, if the maximum IoU of the a priori frames and all the targets is less than the threshold value, then it is considered that this a priori frame does not contain the target and is labeled as a negative sample with a confidence level of 0. The third step is that the remaining a priori frames in the second step are identified as non-positive and non-negative samples and are not involved in the network calculations.

Based on the above analysis, it can be concluded that when there is no intersection of the predicted and real frames, $A \cap B = 0$, leading to $IoU_{(A,B)}$ is always zero, and the gradient of the loss function is zero, which indicates that IoU will not have a gradient shift of the loss function in the case of non-overlapping of the predicted and real frames, and the model will not be able to continue to optimize.

To address this, we adopt the Complete IoU (CIoU) loss, which considers overlap area, center distance, and aspect ratio to improve localization accuracy. The calculation formula is shown in equation (3):

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \tag{3}$$

where av as the influence factor CIoU integrates the IoU term with penalties based on the Euclidean distance between box centers and the aspect ratio discrepancy. b and b^{gt} denote the centroids of the predicted and real frames, ρ^2 denotes the Euclidean distance, and c denotes the diagonal distance between the predicted and real frames from the smallest outer bounding rectangle. α and v are calculated as in equations (4) and (5):

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{4}$$

$$v = \frac{4}{\pi^2}(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h})^2 \tag{5}$$

III. SCTNet-based medical image segmentation algorithm

III. A. UNet algorithm

UNet[29] is a convolutional neural network (CNN) architecture specifically designed for biomedical image segmentation. It follows a symmetric U-shaped structure comprising two main paths: the encoder (contracting path) and the decoder (expanding path). The encoder path progressively reduces the spatial resolution while increasing the number of feature channels to extract high-level semantic representations. In contrast, the decoder path performs upsampling to recover spatial resolution and restore fine-grained localization details.

Formally, let $X \in R^{H \times W \times C}$ denote the input image, where H, W, and C are the height, width, and number of channels, respectively. The encoder applies a sequence of convolutional operations and downsampling functions (typically max pooling), which can be represented as in equations (6):

$$f_l = Down(\sigma(W_l * f_{l-1} + b_l)), l = 1, 2, ...L \tag{6}$$

where Wl and bl are the learnable parameters at layer l, σ is a non-linear activation function (e.g., ReLU), and Down denotes a pooling operation.

The decoder path restores the spatial dimensions using transposed convolution (also known as deconvolution), while skip connections transfer feature maps from corresponding encoder layers to the decoder layers, thus enabling the network to leverage both coarse and fine features. These skip connections can be mathematically expressed as in equations (7):

$$g_l = Concat(f_l, Up(g_{l+1})), l = L - 1, ..., 1 \tag{7}$$

where Up denotes the upsampling operation and Concat represents channel-wise concatenation. This structure enables the model to perform pixel-wise classification while preserving spatial coherence.

As in equations (8), the final segmentation map is obtained by applying a 1×1 convolution to the last decoder output, projecting it into an N-channel output, where N is the number of segmentation classes:

$$S = Soft\max(W_s \times g_l + b_s) \tag{8}$$

During training, the network is typically optimized using the pixel-wise cross-entropy loss, defined as in equations (9):

$$L_{CE} = -\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{c=1}^{N} y_{ij}^{(c)} \log \hat{y}_{ij}^{(c)} \tag{9}$$

where $y_{ij}^{(c)} \in \{0,1\}$ is the ground truth label for class c at pixel (i,j), and $\hat{y}_{ij}^{(c)}$ is the predicted probability for that class.

Owing to its elegant architecture and strong performance, UNet has become a foundational model for medical image segmentation tasks, particularly under conditions of limited annotated data.

The decoder portion of the UNet architecture is composed of a sequence of transposed convolutional layers—commonly referred to as upsampling layers—interleaved with skip connections that link corresponding stages in the encoder and decoder. These transposed convolutions serve to incrementally restore the spatial resolution of feature maps, bringing them back to the original image dimensions while maintaining semantically rich representations acquired during the encoding phase.

The skip connections, a defining feature of UNet, directly transfer high-resolution features from the encoder to the decoder. This pathway compensates for the spatial information that may be diminished or lost during downsampling. By merging the encoder's fine-grained spatial cues with the decoder's abstract semantic understanding, the network becomes capable of generating highly detailed and accurate segmentation outputs at the pixel level.

In the final stage of the network, a 1×1 convolution is typically applied to the output of the decoder to produce a feature map with N channels, where N denotes the number of target classes. This operation enables the network to assign class probabilities to each pixel, thereby facilitating accurate differentiation of anatomical structures or lesion types. The model is trained using a pixel-wise cross-entropy loss function, and its parameters are optimized using standard gradient-based algorithms such as Adam or stochastic gradient descent (SGD).

The overall architecture of UNet is illustrated in Figure 3, clearly exhibiting its characteristic symmetric U-shaped design. The encoder on the left progressively compresses spatial dimensions while deepening semantic understanding through the combined operation of convolutional and pooling layers. In contrast, the decoder on the right restores spatial resolution through upsampling and feature fusion. Of particular importance are the skip connections between mirrored layers in the encoder and decoder, which play a key role in preserving crucial localization cues. This architectural design aligns closely with the specific demands of medical image segmentation, where both precise boundary delineation and contextual awareness are essential to producing clinically meaningful results.

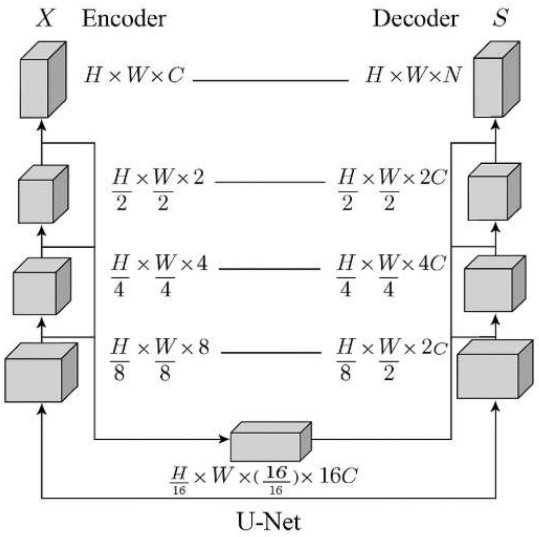

Figure 3: Architecture of UNet

III. B. Selective Context Transformer Enhanced SCTNet

III. B. 1) Network structure

The overall architecture of SCTNet adopts a symmetrical encoder–decoder structure, specifically designed for high-precision medical image segmentation tasks. The encoder is responsible for extracting hierarchical feature representations rich in semantic context, laying the groundwork for accurate pixel-level predictions during the decoding phase. Structurally, the encoder primarily consists of Double CBS modules, Selective Contextual Transformer (SCT) units, and two downsampling stages implemented via max pooling operations.

Each Double CBS module comprises two cascaded CBS blocks. A single CBS block integrates a 3×3 convolutional layer, batch normalization (BN), and the SiLU activation function, forming an efficient feature extraction unit. The SCT module innovatively combines a CBS block with a Selective Contextual Transformer attention mechanism, with the core objective of enhancing the network's capacity to model long-range dependencies. Before attention computation, a channel expansion layer is applied to enrich the feature map's expressive power. Subsequently, cross-spatial domain contextual relationships are captured through the SCT mechanism, enabling effective modeling of global semantic information.

The decoder path performs progressive upsampling of the high-level semantic features produced by the encoder. At each stage, skip connections are established between encoder and decoder layers to facilitate the fusion of semantic and spatial information, ensuring fine-grained detail preservation. The network outputs a full-resolution segmentation map through a final 1×1 convolution layer followed by softmax activation for pixel-wise classification.

The complete structure is illustrated in Figure 4, where the encoder (left) begins with the input image and passes through successive Double CBS blocks and SCT attention modules, while the decoder (right) mirrors the structure to reconstruct the output segmentation map. The dotted arrows represent skip connections that link encoder and decoder stages at corresponding resolutions, enabling efficient fusion of multilevel feature representations.

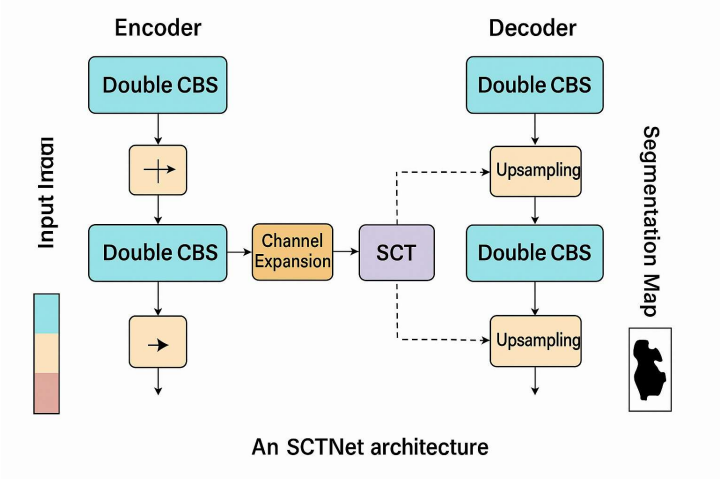

Figure 4: The architecture of SCTNet

III. B. 2) Selective contextual transformers

To obtain rich contextual semantic information, it is necessary to take into account both global and local information of the feature map. Convolutional neural network gradually expands the sensory field by continuously stacking convolutional layers. At the same time, feature information with high-level semantics in the image is gradually extracted from shallow features to deep features, and each layer learns its respective feature representation to enhance the characterization capability. However, studies have shown that the effective receptive field of stacked convolutional layers is often much smaller than the theoretical one, which limits their ability to capture global contextual dependencies. In order to fully utilize the local context information and global context information of the feature graph, the attention mechanism SCT is proposed:

(1) Inter-nearest-neighbor local context expression features K^1

In order to obtain the inter-nearest neighbor local context information of the target, the AMSK module is designed. AMSK enables the network to adaptively select the appropriate convolutional kernel size, so as to model the different scales of information in the feature maps and extract x the context information.

AMSK integrates the information streams from multiple branches through the add operation and inputs them into the next layer of neurons. The computational process is shown in Equation (10)(11):

$$D_i = C_{k=3}^{r_i}(X^{H\times W\times C}) \tag{10}$$

$$M = Add[D_1, D_2, D_3, D_4] \tag{11}$$

$i \in [1,4]$, i are integers representing the i rd branch. k represents the convolution kernel size. C is the convolution. D_i represents the feature map after the i th inflated convolution. Add is the summation operation of the feature map.

The average pooling operation preserves the location information of the feature map, and the global pooling operation improves the network's noise immunity to the feature map, and the two parallel methods are more effective than single pooling. The fused feature map is used to extract the salient and detailed features of the feature map using global maximum pooling and global average pooling in parallel, respectively. The computational process is shown in (12):

$$P = Add[AvgPool(M), MaxPool(M)] \tag{12}$$

Selection of most suitable kernel: Convolutional weights for each branch are obtained by using 1*1 convolution component, softmax function.1*1 convolution component is used to generate selective parameters to guide the sensory field of the convolution kernel.softmax function is used to assign weights on the channel based on the features.The weights for different convolution kernel channels are represented as follows (13):

$$W_i = Softmax[SiLU(BN(C^i_{k=1}(P)))] \tag{13}$$

BN stands for Batch Normalization, SiLU stands for SiLU activation function, and W_i is the weight of the i rd branch feature map.

By multiplying the attention weights with the elements of the corresponding channels to weight the feature maps of different branches, the important contextual information in different sensory fields is captured adaptively. These feature maps with rich contextual information are then summed to get the local contextual expression feature K^1 that reflects the local context among the nearest neighbors.

(2) Enhanced features with global context information

The resulting K^1 and Q are concatenated and the attention matrix is computed by two 1×1 convolutions (14):

$$W_V = C_{k=1}(SiLU(BN(C_{k=1}([K^1, Q])))) \tag{14}$$

For each head, the local attention matrix at each spatial location is learned based on Q and the local context expression feature K^1 among the nearest neighbors, instead of isolated query key pairs in the self-attention mechanism. Self-attention learning is enhanced by mining the local contextual information of the target. Based on the contextualized attention matrix W_V, we dot-multiply it with K^2 to obtain the augmented feature V, as it shows (15):

$$K^2 = V * W_V \tag{15}$$

Useful semantic information is filtered from the global by utilizing the information in the feature map to encode the feature representation at each location, which is then computed with the contextualized attention matrix W_V. Enhanced features that reflect the global contextual information are computed K^2. Finally, the resulting local contextually enhanced features K^1 and global contextually enhanced features K^2 are fused in an additive manner as the output of the Selective Contextual Transformer attention structure.

III. B. 3) Feature Enhancement and Alignment

While skip connections facilitate feature reuse by fusing multi-level features—typically by concatenating upsampled semantic features with low-level spatial features—they may also introduce feature inconsistency due to the semantic gap between early and late network stages. Specifically, the simple fusion of heterogeneous features across scales does not sufficiently capture fine-grained spatial detail nor adequately model semantic continuity, which limits the network's ability to accurately localize and delineate lesion boundaries in medical images.

To address this issue, we propose the Feature Enhancement Attention (FEA) module, which consists of two complementary components: the Feature Enhancement (FE) module and the Feature Alignment (FA) module.

The Feature Enhancement (FE) module operates by fusing the upsampled decoder feature map with the corresponding encoder feature map prior to upsampling. Through this process, the module leverages deep semantic information to reinforce low-level spatial details, enabling the network to construct feature representations that preserve both fine-grained structural elements and high-level contextual understanding—an essential capability for achieving precise segmentation boundaries.

Complementing this, the Feature Alignment (FA) module addresses semantic discrepancies between features at different levels during the decoding phase. It receives the upsampled feature map and the decoder output at the same spatial resolution, projecting them into a shared latent space to harmonize their semantic content. By aligning feature representations across varying resolutions, the FA module enhances the consistency of feature fusion, thereby improving the model's ability to delineate complex anatomical structures with greater accuracy.

IV. Design and realization of intelligent medical image processing system

The intelligent medical image processing system represents a crucial component of modern healthcare infrastructure, offering more accurate and efficient diagnostic support through automated image analysis technologies. This chapter focuses on the system-level design, functional architecture, and technical implementation of the platform.

To ensure a scientifically grounded and effective development process, both functional and non-functional requirements were thoroughly analyzed. From a functional perspective, the system establishes a comprehensive user management framework that supports multi-method registration, rigorous identity authentication, and hierarchical access control. The patient medical record module enables structured data storage, facilitating the association with imaging data and supporting fast retrieval. As the system's core, the image processing module supports multi-format image upload and preview, along with interactive features such as zooming and window width/level adjustment. The lesion detection and segmentation module leverages deep learning algorithms to

accurately identify pathological regions, while the diagnosis visualization module translates data into charts and 3D models to assist clinical decision-making.

In terms of non-functional requirements, the system is designed for high-performance responsiveness to ensure rapid image processing. A scalable architecture allows for the seamless addition of new functional modules and integration with external devices or algorithms. Data security is ensured through encryption and backup/recovery mechanisms, while fault monitoring capabilities support stable long-term system operation.

The overall system adopts a modular layered architecture based on the Model-View-Controller (MVC) pattern, which clearly separates responsibilities among the user interface, backend processing logic, and database management. This approach enhances the system's maintainability, scalability, and inter-module decoupling.

From a data management perspective, the system includes a relational database designed to manage patients' information, diagnostic records, lesion analysis reports, and system logs. The database schema consists of tables such as Patients, Users, MedicalImages, and DetectionResults, with well-defined primary-foreign key relationships to ensure data integrity and efficient querying. Indexing strategies and normalization techniques are applied to optimize storage and retrieval performance, especially for high-throughput image handling scenarios.

A schematic overview of the system architecture is shown in Figure X, highlighting the data flow between the front-end (user portal), middleware (image preprocessing and deep learning inference engine), and back-end services (database and file system). RESTful APIs facilitate communication among subsystems and enable future interoperability with external hospital information systems (HIS) or picture archiving and communication systems (PACS).

IV. A. System requirements analysis

Functional requirements analysis constitutes a critical phase in the development of software and information systems, serving as the foundation for ensuring that the designed functionalities align with user expectations and practical needs. A thorough understanding and specification of functional requirements can significantly reduce the risk of implementation errors, minimize redundant development efforts, and enhance the overall system quality and user satisfaction.

In this subsection, the application context and target user group of the proposed system are clearly defined. The system is specifically designed to support hospital-based medical imaging professionals, such as radiologists and diagnostic technicians, in conducting efficient lesion detection and segmentation tasks.

Through detailed assessment of the clinical usage environment, operational workflows, and diagnostic needs, the functional scope and objectives of the system are established. These include modules for user authentication, patient information management, medical image upload and display, automated lesion detection and segmentation, and result visualization and storage.

The key functional modules of the intelligent medical image processing system are illustrated in Figure 5, which outlines the primary user interactions and system operations.

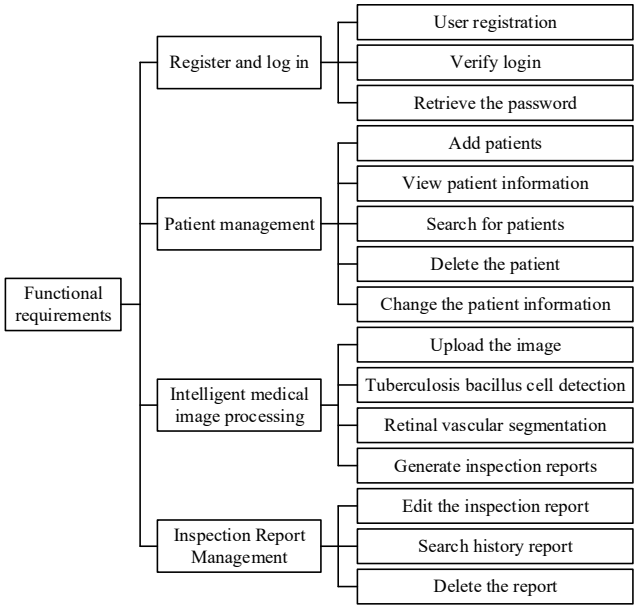

Figure 5: Functional Requirements

Non-functional requirements refer to the constraints and quality attributes a software system must fulfill beyond its core functionalities. These requirements—such as usability, security, and maintainability—are essential for ensuring the system's long-term stability, operational effectiveness, and user acceptance.

Usability: The system must offer high operational efficiency and ease of use. For instance, in the image preview interface, clinicians should be able to adjust CT window width and level in real time using the mouse scroll wheel, with a response latency of less than 200 milliseconds. If uploading a DICOM image fails, the system should display a clear error message (e.g., "File format corrupted") and provide options such as one-click retry or manual import from an alternative path.

Security: Given the sensitive nature of medical data, a robust security framework is required. During login, users must pass dual authentication via account password and dynamic SMS verification code. Patient MRI images are encrypted using the AES-256 protocol during transmission, and are automatically chunked, encrypted, and backed up to an offsite server during storage. The system also performs real-time monitoring for SQL injection attacks; when abnormal SQL queries are detected, the request is immediately blocked and logged.

Maintainability: To accommodate evolving clinical demands, the system architecture follows a modular design. For example, to add a new "AI-based Pulmonary Nodule Malignancy Prediction" feature, developers need only integrate the corresponding deep learning model into the backend algorithm module and configure a new function entry in the frontend interface—without altering the underlying data processing logic. System components (e.g., image rendering engine and medical record management module) communicate via loosely coupled APIs, ensuring that upgrades to one module do not affect the functionality of others..

IV. B. System design and implementation

IV. B. 1) Overall architectural design

The architecture of the system is chosen as Browser-Server and the architectural pattern is chosen as MVVM. This architectural pattern is a classic front-end design pattern whose name stands for three different components, i.e. Model, View and View Model.

The overall flowchart of the system is shown in Figure 6, which consists of a Web front-end, a back-end and a model reasoning end. When a user performs an operation, requests are sent to different back-end interfaces depending on the type of operation. If it is a routine operation such as logging in, changing patient information or managing examination reports, the request is sent to the normal interface of the back-end and the response is received. If it is a test operation, the request is sent to the back-end medical image processing interface. After receiving the request, the back-end image processing interface invokes a deep learning model to reason about the image uploaded by the user. Finally, the image inference results are stored and returned layer by layer to the front-end for rendering.

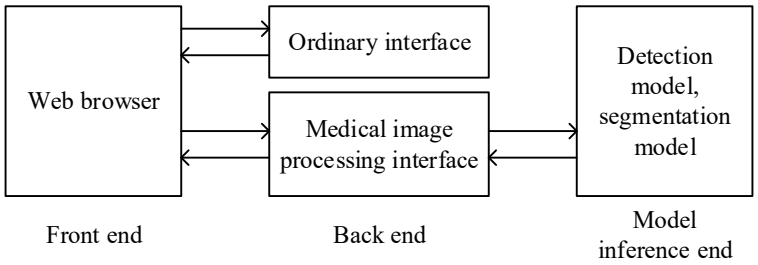

Figure 6: The overall flowchart of the system

IV. B. 2) Database design

IIn the medical image intelligent processing system, it is necessary to manage and store a large volume of data, including medical images, patient records, pathology reports, and diagnostic outputs. A structured and secure relational database system serves as an efficient and reliable solution for this purpose, ensuring data consistency, accessibility, and integrity.

Based on the functional requirements outlined in this study, the database schema is designed to include three core tables: doctor information, patient information, and examination reports. The relational logic is as follows: each patient may undergo multiple examinations, resulting in a one-to-many relationship between the patient information table and the examination report table; conversely, each examination is conducted by a specific doctor, establishing a many-to-one relationship between the examination report table and the doctor information table.

The examination report table includes attributes such as patient ID, doctor ID, report content, report date, and the system-processed medical image storage address. The report_id serves as the primary key, while both doctor_id

and patient_id are foreign keys referencing the respective tables. The patient information table stores personal and clinical details, including patient name, ID number, date of birth, and consultation method. The doctor information table may further include fields such as name, department, and professional title.

Figure 7 illustrates the Entity-Relationship (E-R) diagram of the database. It visually represents the structure and cardinality between key entities—Doctor, Patient, and Examination Report—highlighting the referential integrity maintained through foreign key constraints.

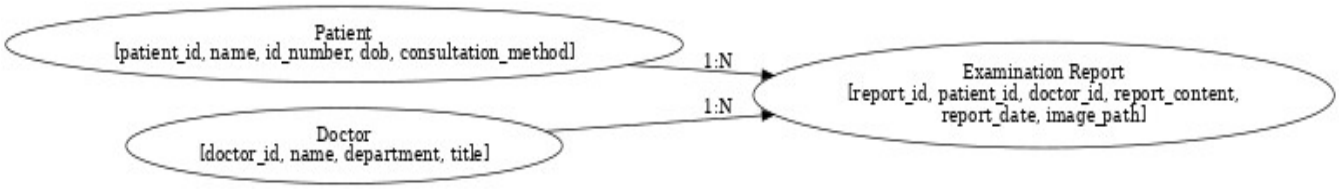

Figure 7: E-R diagram of the medical image processing system database

V. Performance testing of medical image detection and image segmentation algorithms

V. A. Performance Test of Medical Image Detection Algorithm

V. A. 1) Data sets and evaluation indicators

Developed by a team at the National Institutes of Health Clinical Center (NIHCC), the DeepLesion dataset is currently the world's largest open dataset of multi-category, lesion-level labeled CT images, and is the only large open-source dataset that can handle the task of generic lesion detection in CT images.

This section focuses on the evaluation metrics to assess the target detection performance. For the predicted detection frames, when the intersection and concurrency ratio between the predicted detection frames and the real detection frames is greater than a set threshold, the prediction is considered successful and positive, and vice versa, the prediction is considered failed and negative.

There are four cases according to the categories of prediction detection frames as follows:

(1) True Positive (TP): the true category is positive and the predicted category is positive;

(2) False positive (FP): the true category is negative and the predicted category is positive;

(3) True negative (TN): true category is negative, predicted category is negative;

(4) False negative (FN): true category is positive, predicted category is negative

According to the predicted detection frame category, the sensitivity can be obtained, which indicates the percentage of the number of correctly predicted detection frames among all the predicted detection frames, and the higher value of the sensitivity represents the better performance of the target detection. Its calculation formula is shown in (16).

$$Sensitivity = \frac{TP}{TP + FN} \tag{16}$$

In addition, the FROC (ROC) curve is a commonly used metric for medical analysis, which is improved on the basis of the ROC curve and solves the problem that the ROC curve cannot solve the problem of evaluating multiple abnormalities on a single image.The FROC curve can evaluate any number of false-positive situations in the input image, and the vertical coordinate of the FROC curve is the sensitivity, and the horizontal axis is the false The number of positives (FPs) is plotted as follows:

1) For all the prediction detection frames, they are sorted according to their scores from largest to smallest;

2) Take out the prediction detection frames in order, calculate the intersection ratio between them and the real detection frames, which is greater than the threshold value and is recorded as true positive (TP) samples, and vice versa is recorded as false positive (FP) samples, and the cumulative number of true positives and false positives can be used to calculate the current value of sensitivity and FPs, which is the ratio of the number of false positives to the number of the total number of test pictures;

3) Every time a prediction test frame is taken out, the sensitivity and FPs will change, and the FROC curve is obtained with FPs as the horizontal coordinate and sensitivity as the vertical coordinate.

V. A. 2) Medical Image Detection Algorithm Performance Test Results

In order to objectively evaluate the detection performance of this paper's model, this paper compares the performance of the proposed method with several state-of-the-art algorithms, such as 3DCE, ULDOR, RetinaNet, Tao et al., MVP-Net, MULAN, ElixirNet, and 3D-AG, on the DeepLesion dataset.

Table 1 shows the comparative accuracies of each algorithm on the DeepLesion dataset, containing the sensitivity of the different algorithms under multiple FPs, as well as the average sensitivity of the algorithms. From the comparison results, it can be seen that the sensitivity values of the algorithms in this paper under multiple FPs, as well as the average sensitivity values, exceed the other algorithms including MULAN. At lower FPs, this paper's algorithm shows a greater advantage, compared to the optimal algorithm MULAN, this paper's algorithm increases the sensitivity by 1.56% at 1 FPs, 0.76% at 3 FPs, 0.73% at 5 FPs, 0.48% at 7 FPs, and the average sensitivity increases by 0.46% compared to MULAN. As the FP threshold increases, the relative advantage of our model diminishes, likely because all methods perform well at high FP levels, leaving limited room for improvement.

Table 1: The performance comparison of the deeplesion data set

Method	Sensitivity (%)						
	1	3	5	7	9	13	Mean
3DCE	62.48	73.37	80.70	85.65	89.09	91.06	75.55
ULDOR	52.86	64.80	74.84	84.38	87.17	91.80	69.22
RetinaNet	72.22	80.03	86.42	90.79	94.09	96.31	82.39
Tao et al.	71.38	78.52	84.07	87.60	90.19	91.35	80.42
MVP-Net	73.78	81.86	87.63	91.25	--	--	83.63
MULAN	76.12	83.74	88.74	92.35	94.71	95.72	88.56
ElixirNet	67.37	77.91	86.03	91.68	93.57	93.59	85.03
3D-AG	74.66	81.46	85.47	89.19	92.08	94.70	86.26
This method	77.31	84.38	89.39	92.79	94.95	95.02	88.97

V. B. Image Segmentation Algorithm Performance Test

V. B. 1) Data sets and evaluation indicators

Medical image segmentation, as the basis of medical image analysis, has become one of the hotspots of current research. The effect of medical image segmentation is directly related to the precision and accuracy of subsequent medical applications, so the selection and use of medical image datasets are also crucial in the research of medical image segmentation algorithms. In this study, we conduct experiments on several widely-used medical image datasets.

The selected datasets include smaller-scale datasets—GLAS, RITE, and MoNuSeg—as well as large-scale datasets such as DSB 2018 and CVC-ClinicDB.

Driven by deep learning, medical image segmentation technology has been widely applied and developed. However, performance evaluation of medical image segmentation has been a challenging problem. In this subsection, the performance metrics of medical image segmentation are described in detail and the applications in medical image segmentation are discussed.

Four common metrics are used to evaluate the segmentation model, namely Fl score, IoU score, Precision and Recall, which are tested on Matlab. Among them Precision and Recall are two important metrics to measure the performance of a classifier. They are commonly used to evaluate the performance of a classifier in a binary classification problem, where the classifier classifies the input samples into two classes: positive and negative. Both Precision and Recall can be used to evaluate the performance of a classifier in different situations. Specifically, Precision refers to how many samples predicted by the classifier to be in the positive class are truly in the positive class, whereas Recall the proportion of all truly positively classified samples that are correctly predicted as positive by the classifier. Precision can be calculated using the following formula (17):

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

Recall can be calculated using the following formula (18):

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

where FN denotes positive samples that were incorrectly predicted as negative (False Negative).Recall measures the ability of the classifier to correctly predict samples from positive categories. Specifically, the higher the Recall, the lower the underreporting rate of the classifier, i.e., the classifier is able to correctly identify more samples from positive categories.

The Fl metric is the reconciled mean of Precision and Recall. The F1 metric combines Precision and Recall for situations where the number of samples from different classes is unbalanced. When Precision and Recall are the same, the F1 score reaches a maximum value of 1. The formula (19) for the Fl metric is .

$$F1 = \frac{2*(Precison*Recall)}{Precison + Reall} \tag{19}$$

IoU metric stands for Intersection over Union and is also known as Jaccard index.IoU metric measures the ratio between the intersection and concatenation of the predicted region and the real labeled region and is a common metric of segmentation accuracy.The higher the IoU, the closer the segmentation result is to the real labels.The IoU metric is calculated by the formula (20):

$$IoU = \frac{TP}{TP + FP + FN} \tag{20}$$

F1 metrics and IoU metrics are widely used in medical image segmentation tasks.F1 metrics are able to synthesize the effects of Precision and Recall when the number of samples from different categories is unbalanced, while IoU metrics can accurately measure the similarity between predicted results and real labels.

V. B. 2) Image Segmentation Algorithm Performance Test Results

To comprehensively evaluate the effectiveness of the proposed segmentation algorithm, comparative experiments were conducted against several mainstream segmentation methods across four benchmark datasets. The evaluation focused on four key performance metrics: F1 score, Intersection over Union (IoU), Precision, and Recall. All metrics were computed using MATLAB.

Table 2 presents the quantitative results on the GLAS dataset. The proposed method achieved an mF1 score of 0.8285 and an mIoU score of 0.7246, both of which surpassed those of all compared algorithms. In particular, the method outperformed the next-best model, MedT, by 4.98% in mF1 and 8.64% in mIoU, demonstrating superior performance in accurately delineating glandular structures in histopathological images.

Table 2: Quantitative results on the GLAS data set

Network	mF1	mIoU	Recall	Precision
SCTNet	0.8285	0.7246	0.8921	0.8006
MedT	0.7892	0.6670	0.8970	0.7336
Kiu-Net	0.7695	0.6422	0.8665	0.7234
U-Net	0.7568	0.6302	0.8250	0.7604
U-Net++	0.7796	0.6569	0.7579	0.8416
Attention Unet	0.6789	0.5391	0.6834	0.7554
R2U-Net	0.6206	0.4959	0.6952	0.6656
Resunet	0.7449	0.6101	0.8038	0.7307
Channel-Unet	0.7772	0.6507	0.8111	0.7914
UCTransNet	0.7871	0.6668	0.8253	0.7863
TMUNet	0.7926	0.6704	0.7842	0.8279

Table 3 presents the evaluation results on the RITE dataset. This dataset contains only 40 images with intricate structural details, which poses a challenge for most segmentation networks, leading to suboptimal performance. Nevertheless, the proposed method demonstrates competitive performance. This can be attributed to the incorporation of the Selective Contextual Transformer (SCT) module, which enhances the model's ability to capture fine-grained image features and effectively model global contextual relationships—thereby improving the reconstruction of retinal vessel contours. The proposed method achieves an mF1 score of 0.7718 and an mIoU of 0.6291, representing improvements of 2.25% and 3.50%, respectively, over the next-best-performing KiU-Net. Additionally, it achieves a Recall of 0.7670 and a Precision of 0.7813, outperforming most competing approaches.

Table 3: Quantitative results on the RITE data set

Network	*mF1*	*mIoU*	*Recall*	*Precision*
SCTNet	0.7718	0.6291	0.7670	0.7813
MedT	0.6593	0.4926	0.6197	0.7092
Kiu-Net	0.7548	0.6078	0.7418	0.7740
U-Net	0.6475	0.4792	0.5102	0.8929
U-Net++	0.6723	0.5073	0.6818	0.6698
Attention Unet	0.7408	0.5882	0.7123	0.7775
R2U-Net	0.7239	0.5670	0.6418	0.8347
Resunet	0.6655	0.4995	0.5506	0.8467
Channel-Unet	0.7395	0.5872	0.8011	0.6889
UCTransNet	0.7396	0.5873	0.7438	0.7413
TMUNet	0.7398	0.5868	0.7428	0.7412

Table 4 shows the quantitative results of the model on the DSB dataset. The algorithm in this paper corresponds to an mF1 score of 0.9279, an mIoU score of 0.8630, a Recall of 0.9241, and a Precision of 0.9353. Its mF1 and mIoU outperform the other methods, and it is 1.92% and 2.52% higher than the MedT of the next-best performing method, respectively.

Table 4: Quantitative results on the DSB data set

Network	*mF1*	*mIoU*	*Recall*	*Precision*
SCTNet	0.9279	0.8630	0.9241	0.9353
MedT	0.9104	0.8418	0.9324	0.8981
Kiu-Net	0.9038	0.8283	0.9107	0.9031
U-Net	0.8986	0.8207	0.8837	0.9251
U-Net++	0.9017	0.8281	0.9233	0.8920
Attention Unet	0.8963	0.8197	0.9059	0.9021
R2U-Net	0.8719	0.7820	0.8857	0.8848
Resunet	0.8940	0.8123	0.8997	0.8971
Channel-Unet	0.9048	0.8336	0.9139	0.9077
UCTransNet	0.9063	0.8316	0.9004	0.9196
TMUNet	0.9071	0.8342	0.9148	0.9071

The quantitative results of CVC-ClinicDB are shown in Table 5. The algorithm in this paper achieves 0.9071 mF1, 0.8917 mIoU, 0.9531 Recall, and 0.9042 Precision, which is 2.46% and 0.70% higher than the next best performing MedT in terms of mF1 scores and mIoU scores, respectively, with a significant improvement in accuracy, while Precision is also the highest. Recall is also superior to most methods.

Table 5: Quantitative results on the CVC-ClinicDB data set

Network	*mF1*	*mIoU*	*Recall*	*Precision*
SCTNet	0.9071	0.8917	0.9531	0.9042
MedT	0.8853	0.8855	0.9870	0.8854
Kiu-Net	0.6255	0.4983	0.5880	0.8015
U-Net	0.7616	0.6710	0.8823	0.7573
U-Net++	0.7902	0.6991	0.8566	0.8072
Attention Unet	0.6856	0.5745	0.7316	0.7517
R2U-Net	0.4715	0.3764	0.4958	0.5845
Resunet	0.6854	0.5920	0.7147	0.7798
Channel-Unet	0.7684	0.6728	0.8029	0.8135
UCTransNet	0.8816	0.8248	0.8959	0.9193
TMUNet	0.8775	0.8116	0.8752	0.9113

VI. Conclusion

This study verifies the effectiveness and clinical applicability of the proposed method through an in-depth exploration of the innovative application of interpretable artificial intelligence (AI) based on image processing algorithms in medical image analysis.

The enhanced YOLOv12 detection network, incorporating the AGs-ECSA hybrid attention module and the loopback residual structure, achieves a 1FPs sensitivity of 77.31% on the DeepLesion dataset—representing a 1.56% improvement over the benchmark MULAN algorithm—and reaches an average sensitivity of 88.97%. Meanwhile, the proposed SCTNet segmentation network, equipped with a selective contextual transformer module, achieves an mF1 score of 0.8285 on the GLAS dataset, significantly outperforming several state-of-the-art segmentation methods. The model also demonstrates strong generalization ability, achieving an mF1 score of 0.9279 on the DSB dataset.

Practical clinical case studies further validate the system's effectiveness in real-world scenarios. In cerebral hemorrhage CT images, the system accurately identifies lesion regions, performs localization, measures lesion radius, and supports quantitative diagnostic analysis. The integration of attention mechanisms and residual feedback structures not only enhances detection performance but also improves interpretability, allowing clinicians to better understand the model's decision-making logic.

The complete medical image intelligent processing system developed in this work offers a robust technical foundation for clinical deployment. Its high accuracy, interpretability, and scalability underscore its potential for widespread application in computer-aided diagnosis, contributing to improved clinical efficiency and diagnostic consistency.

References

[1] Kumar, S. N., Lenin Fred, A., Padmanabhan, P., Gulyas, B., Ajay Kumar, H., & Jonisha Miriam, L. R. (2021). Deep learning algorithms in medical image processing for cancer diagnosis: Overview, challenges and future. Deep learning for cancer diagnosis, 37-66.

[2] Jiang, H., Diao, Z., Shi, T., Zhou, Y., Wang, F., Hu, W., ... & Yao, Y. D. (2023). A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation. Computers in Biology and Medicine, 157, 106726.

[3] Stoitsis, J., Valavanis, I., Mougiakakou, S. G., Golemati, S., Nikita, A., & Nikita, K. S. (2006). Computer aided diagnosis based on medical image processing and artificial intelligence methods. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 569(2), 591-595.

[4] Anas, R., Elhadi, H. A., & Ali, E. S. (2019). Impact of edge detection algorithms in medical image processing. World Scientific News, 118, 129-143.

[5] Jiang, J., Trundle, P., & Ren, J. (2010). Medical image analysis with artificial neural networks. Computerized Medical Imaging and Graphics, 34(8), 617-631.

[6] Ramesh, K. K. D., Kumar, G. K., Swapna, K., Datta, D., & Rajest, S. S. (2021). A review of medical image segmentation algorithms. EAI Endorsed Transactions on Pervasive Health & Technology, 7(27).

[7] Norouzi, A., Rahim, M. S. M., Altameem, A., Saba, T., Rad, A. E., Rehman, A., & Uddin, M. (2014). Medical image segmentation methods, algorithms, and applications. IETE Technical Review, 31(3), 199-213.

[8] Latif, J., Xiao, C., Imran, A., & Tu, S. (2019, January). Medical imaging using machine learning and deep learning algorithms: a review. In 2019 2nd International conference on computing, mathematics and engineering technologies (iCoMET) (pp. 1-5). IEEE.

[9] Chandy, A. (2019). A review on iot based medical imaging technology for healthcare applications. Journal of Innovative Image Processing (JIIP), 1(01), 51-60.

[10] Pinto-Coelho, L. (2023). How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications. Bioengineering, 10(12), 1435.

[11] Li, M., Jiang, Y., Zhang, Y., & Zhu, H. (2023). Medical image analysis using deep learning algorithms. Frontiers in public health, 11, 1273253.

[12] Houssein, E. H., Abdelkareem, D. A., Emam, M. M., Hameed, M. A., & Younan, M. (2022). An efficient image segmentation method for skin cancer imaging using improved golden jackal optimization algorithm. Computers in Biology and Medicine, 149, 106075.

[13] Nadimi-Shahraki, M. H., Zamani, H., & Mirjalili, S. (2022). Enhanced whale optimization algorithm for medical feature selection: A COVID-19 case study. Computers in biology and medicine, 148, 105858.

[14] Jasti, V. D. P., Zamani, A. S., Arumugam, K., Naved, M., Pallathadka, H., Sammy, F., & Kaliyaperumal, K. (2022). Computational technique based on machine learning and image processing for medical image analysis of breast cancer diagnosis. Security and communication networks, 2022(1), 1918379.

[15] Rammurthy, D., & Mahesh, P. K. (2022). Whale Harris hawks optimization based deep learning classifier for brain tumor detection using MRI images. Journal of King Saud University-Computer and Information Sciences, 34(6), 3259-3272.

[16] Yazhinian, S., Rao, V. S., Sekhar, J. C., Duraisamy, S., & Thenmozhi, E. (2023). Whale Optimization-Driven Generative Convolutional Neural Network Framework for Anaemia Detection from Blood Smear Images. International Journal of Advanced Computer Science and Applications, 14(7).

[17] Anand, D., Khalaf, O. I., Abdulsahib, G. M., & Chandra, G. R. (2024). Original Research Article Identification of meningioma tumor using recurrent neural networks. Journal of Autonomous Intelligence, 7(2).

[18] Salehin, I., Islam, M. S., Amin, N., Baten, M. A., Noman, S. M., Saifuzzaman, M., & Yazmyradov, S. (2023). Real‐Time Medical Image Classification with ML Framework and Dedicated CNN–LSTM Architecture. Journal of Sensors, 2023(1), 3717035.

[19] Liu, W., Liu, X., Luo, X., Wang, M., Han, G., Zhao, X., & Zhu, Z. (2023). A pyramid input augmented multi-scale CNN for GGO detection in 3D lung CT images. Pattern Recognition, 136, 109261.

[20] Puttagunta, M., & Ravi, S. (2021). Medical image analysis based on deep learning approach. Multimedia tools and applications, 80(16), 24365-24398.
[21] Tian, Y., Ye, Q., & Doermann, D. (2025). YOLOv12: Attention-centric real-time object detectors. arXiv. https://doi.org/10.48550/arXiv.2502.12524
[22] Wang, C.-Y., Yeh, I.-H., Liao, H.-Y. M., & Chen, P.-Y. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. CVPR, 7464–7475.
[23] Chen, Y., Kalra, S., Gilmore, H., Tomaszewski, J. E., & Madabhushi, A. (2021). Spatially aware transformer networks for histopathological image analysis. Nature Machine Intelligence, 3(10), 773–786.
[24] Tang, Y., Yang, Z., Zhang, Y., & Li, Z. (2022). Dual attention residual U-Net for medical image segmentation. IEEE Transactions on Instrumentation and Measurement, 71, 1–10.
[25] Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. ECCV, 3–19.
[26] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. CVPR, 770–778.
[27] Chunning Na,Yudi Zhou,Feng Li & Huan Pan. (2025). MAMBA2NILM: a non-intrusive load monitoring approach combined the SENet attention mechanism and Mamba2. Electric Power Systems Research,247,111805-111805.
[28] Ren Yawei,Zhou Rui & Li Jun. (2025). UFEFusion: a new visible-infrared image fusion model combined with CBAM. International Journal of Intelligent Computing and Cybernetics,18(2),326-352.
[29] Dawei Guan,Xinping Lin,Haoyi Zhang & Hang Zhou. (2025). VariGAN: Enhancing Image Style Transfer via UNet Generator, Depthwise Discriminator, and LPIPS Loss in Adversarial Learning Framework. Sensors,25(9),2671-2671.