IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 1, February, 2025, pp. 167-176
Journal homepage: http://iieta.org/journals/ts

Deep Learning Based Object Detection in Medical Image with YOLOv4-CSP with U-Net Algorithms

Arasakumaran Umamageswari[1*], Christopher Sundarajan Anita[2], L. Sherin Beevi[3], Arasakumaran Sangari[4]

[1] Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai 600089, India
[2] Department of Artificial Intelligence and Machine Learning, R.M.D Engineering College, Kavaraipettai 601206, India
[3] Department of Computer Science and Engineering, R.M.D Engineering College, Kavaraipettai 601206, India
[4] Department of Electrical and Electronics Engineering, Rajalakshmi Engineering College, Chennai 602105, India

Corresponding Author Email: anitacs28377@gmail.com

https://doi.org/10.18280/ts.420115

Received: 5 January 2024
Revised: 7 April 2024
Accepted: 20 August 2024
Available online: 28 February 2025

Keywords:
medical image object detection, YOLOv4-CSP, U-Net, deep learning, medical image segmentation

ABSTRACT

In the realm of medical image analysis, the object detection task poses significant challenges related to classification and regression. The demand for accurate computer assisted detection and diagnosis techniques has prompted researchers to adapt existing object detection methodologies to the medical domain. However, prevailing approaches often overlook critical factors such as the low resolution inherent in medical images, the pervasive presence of noise, and the diminutive size of the objects under scrutiny. The response to these challenges, this study presents a novel algorithmic model termed the YOLO-UNET Fusion (Combination of YOLOv4-CSP and U-Net segmentation). A self-supervised learning strategy, the YOLO-UNET fusion employs a random mask applied to the input image. This process serves to reconstruct input features, fostering the acquisition of a more intricate feature vector while concurrently mitigating the impact of extraneous noise. To specifically address the detection of small objects, the paper introduces a YOLOv4 model. A sliding window incorporating a local self-attention mechanism is employed, assigning elevated attention scores to the smaller objects within the image. A streamlined single stage object detection structure is implemented. This framework predicts a sequence of sets, encompassing the position of the bounding box and the corresponding class of the objects in focus. The proposed model is put to the test on a benchmark dataset, namely NIH DeepLesion, where it outperforms existing methodologies. The suggested approach achieves a compromise between speed and accuracy by blending YOLOv4-CSP with U-Net. This lends to the proposed model appropriate for real-world medical applications. Comprehensive experiments were done on datasets to demonstrate the efficacy and generalization of the proposed methodology. This offers an in-depth investigation of the model's functionality, highlighting its advantages in terms of accuracy and efficiency through comparative assessments with cutting-edge techniques with a detection rate of 89.23%.

1. INTRODUCTION

Medical image enhancement plays a crucial role in object detection within field of medical imaging. Object detection involves identifying and locating specific structures, anomalies within medical images. The purpose of medical image enhancement in object detection includes: Improved visibility of Structures, enhanced feature discriminations, optimized preprocessing for object detection, increased specificity and sensitivity, and to reduce the noise. The purpose of medical image enhancement in object detection is to optimize the quality of input images, making it easier for detection algorithms to identify and locate specific structures or abnormalities. Enhanced images contribute to the overall performance, accuracy, and reliability of object detection models in the field of medical imaging. As deep learning technologies advance, the application of object detection techniques in medical diagnostics has become widespread, proving particularly beneficial in practical scenarios such as the identification of exudates in the retinas of diabetic patients [1, 2], before time tumor detection, and the segmentation of vascular plaques. In conventional medical diagnosis, the identification of lesions within images typically relies on manual assessment by physicians. This process is not only time-consuming but also demands significant labor, especially considering the vast number of images clinicians encounter daily. The repetitive nature of this task poses a risk of visual fatigue among doctors, potentially leading to critical misdiagnoses or missed detections—errors that can have severe consequences. Hence, the integration of deep learning techniques becomes imperative to empower machines to autonomously learn features from images and identify irregular areas, extensively contributing to the field of medical detection [3, 4]. This work presents an innovative approach

that utilizes the YOLOv4-CSP and U-Net algorithms for better object detection for medical photos. Although current technologies have demonstrated potential, intrinsic challenges including variations in size, shape, and texture frequently render it hard for them to accurately recognize things in complex medical images. By leveraging the benefits of U-Net, which is renowned for being precise in semantically segmenting tasks, and YOLOv4-CSP, which is regarded for its effectiveness and speed, this study seeks to get around these limitations.

However, the unique challenge in medical image object detection lies in the small size of the objects of interest. Effectively guiding machines to sift through conditions information and accurately pinpointing these small lesions remains a critical hurdle in the field. This emphasizes the ongoing need for advancements in object detection methodologies tailored specifically for medical applications.

2. BACKGROUND OF THE RELATED WORK

In recent strides in object detection, A work introduced, R–CNN model which employed a CNN to extract image features after obtaining candidate frames and used an SVM for object classification [5]. To enhance model accuracy, a recent work projected the Faster R–CNN model. Utilizing Region Proposal Networks (RPN), this model selected candidate frames, requiring only one feature extraction pass for the subsequent classification and regression of these frames. The R-CNN significantly improved detection speed [6].

In the latest breakthroughs, an algorithm DETR, a groundbreaking approach that integrates a Convolutional Neural Network (CNN) for extracting features [7]. They then harnessed the power of a Transformer for both encoding and decoding to predict bounding boxes. Shifting gears, introduced the Vision Transformer (ViT), a revolutionary concept that dissects input images into patches [8]. Utilizing a linear embedding layer for transformation and a self-attention mechanism, ViT excels in extracting discriminative features. Expanding on the ViT paradigm, a new algorithm introduced the Swin Transformer, a noteworthy extension that introduces the window attention mechanism [9]. This innovative addition not only reduces model complexity but also achieves remarkable performance gains in both image recognition and object detection.

Comparatively, Transformers exhibit superior scalability with large-scale datasets compared to CNNs. The intercorrelation of features extracted by Transformers provides richer semantic information for models.

In the realm of medical advancements, a method elevated the stability of retinal lesion detection by seamlessly integrating domain adaptive capabilities with fully convolutional embedding networks [10]. Venturing into the three-dimensional landscape, an innovative method applied a 3-dimensional convolutional neural network to heighten the accuracy of lesion detection in CT scans [11]. Expanding the horizons of region proposal networks (RPN), a method called extended RPN into the 3D space, ensuring the effective extraction of intricate 3D backgrounds from CT data [12, 13]. Taking a holistic approach to semantic segmentation of 3D medical images, a new technique introduced CoTr, a cutting-edge model that amalgamates the advantages of both CNN and transformer architectures for precise segmentation results [14]. An innovator brought forth Swin-Unet, a novel approach that leverages the Swin Transformer as a background encoder for medical images, coupled with a symmetric Transformer decoder designed for spatial resolution recovery [15]. This comprehensive strategy enhances the overall accuracy and efficiency of medical image analysis. A new method presented TransBTS, leveraging the Transformer architecture for local 3D contextual history information extraction, excelling in tumor segmentation on 3D MRI scans [16].

While these methods demonstrate superior performance in general object detection, their applicability to small object detection in medical images is limited. Medical images often contain small, concentrated lesions, making them challenging for conventional algorithms due to low clarity and significant noise [17]. To address these challenges, the mask mechanism proves instrumental in filtering raw data, removing noise, and obtaining richer semantic information [18]. Integrating the mask mechanism into a hierarchical transformer model introduces a novel self-attention paradigm within a dynamic sliding window [19]. This innovation empowers the model to effectively distinguish nuanced features and concentrate its attention on diminutive objects [20]. This innovative approach holds promise for effective small object detection in the medical field [21]. This paper presents a holistic annotation approach for clinically significant findings in diverse CT images using radiology reports and label ontology, improving automated medical image analysis [22]. The study proposes a deep learning-based method for identifying and classifying leaf diseases, enhancing agricultural disease detection accuracy [23, 24].

To address resolution challenges, noise, and the detection of smaller objects, this paper introduces a U-Net to self-supervised learning in medical images. By segmenting the medical image into patches, applying random sampling and masking operations, and utilizing a YOLOv4-CSP as both encoder and decoder, the model reconstructs the image, effectively filtering out noise. The YOLOv4-CSP model, further optimized to include nonoverlapping windows with a self-attention mechanism, ensures the model's adaptability to the nuances of small object detection areas. This comprehensive approach bridges the gap between low-resolution medical images, noise, and the effective detection of small objects. This paper is structured to provide a comprehensive overview of deep learning-based object detection from various modalities of medical images. Following this introduction, subsequent sections will delve into the various components of the object detection process. These sections will include discussions on pre-processing techniques, segmentation, machine learning models, dataset considerations, performance evaluation, and potential conclusions with future enhancements.

3. DATASET

3.1 NIH DeepLesion dataset

The NIH DeepLesion Dataset stands as an extensive collection of CT images, meticulously compiled by the esteemed U.S. National Institutes of Health (NIH) Clinical Center. Within its expansive database, one can find: 32,125 Axial CT Slices, 10,599 CT Scans (Studies), 4,430 Distinct Patients, and 32,740 Lesions. Diverse in nature, the lesions encompass an array of types, including but not limited to: Lung Nodules, Liver Tumors, and Enlarged Lymph Nodes.

Common methods of preparation, such as sampling for consistent voxel separation, intensity leveling for consistent contrast, and brightness levels, are used to get the data ready for deep learning tasks. To improve data diversity and lessen class imbalance, data augmentation techniques including geometrical and intensity modifications are applied. combining these stages shared, deep learning models trained on the DeepLesion dataset grow more robust and capable of generalizing, which enhances lesion recognition in clinical settings and improves reliability as well as accuracy.

Notably, the dataset boasts over 32,000 annotated lesions, offering valuable insights into medical imaging. The annotations provide 2D diameter measurements and bounding boxes for the lesions. It's important to note that the dataset does not include lesion segmentation masks, 3D bounding boxes, or fine-grained lesion details.

Setting itself apart from many existing medical image datasets, the NIH DeepLesion Dataset possesses the potential to contribute significantly to various medical image applications. Its distinctive characteristic lies in its ability to detect and classify a multitude of lesion types, unlike conventional datasets that may focus solely on one specific lesion type.

This breadth of lesion diversity enhances its utility in medical research and applications, marking it as a valuable resource for advancing our understanding and capabilities in the realm of diagnostic imaging.

4. MATERIALS AND METHODS

The provided diagram in Figure 1 illustrates the architectural framework of the proposed methodology. This system is designed to handle input in the form of images, from SIH DeepLesion Subset. A deep learning model with a hybrid approach that utilizes the U-Net and YOLOv4-CSP architectures for identifying objects in medical pictures. YOLOv4-CSP is the core of an effective and rapid object identification technology that uses the Cross Stage Partial (CSP) connections to enhance the scalability and representation of features. At the same time, the U-Net component is used for fine-grained semantic segmentation, thereby enabling precise localization of objects inside the identified areas. This uses a multi-task method of learning during training, where the model is learned using an assortment of regression and classification losses on both detecting and segmentation assignments simultaneously. In addition, the proposed method incorporates a local self-attention system into the model design to improve overall detection accuracy by enhancing the fusion of features and preserving contextual connections at various levels.

The sequential process involves the following key steps:

(1) Input processing

The journey commences with the intake of digital images of medical findings. These images might be any medical images with various modalities.

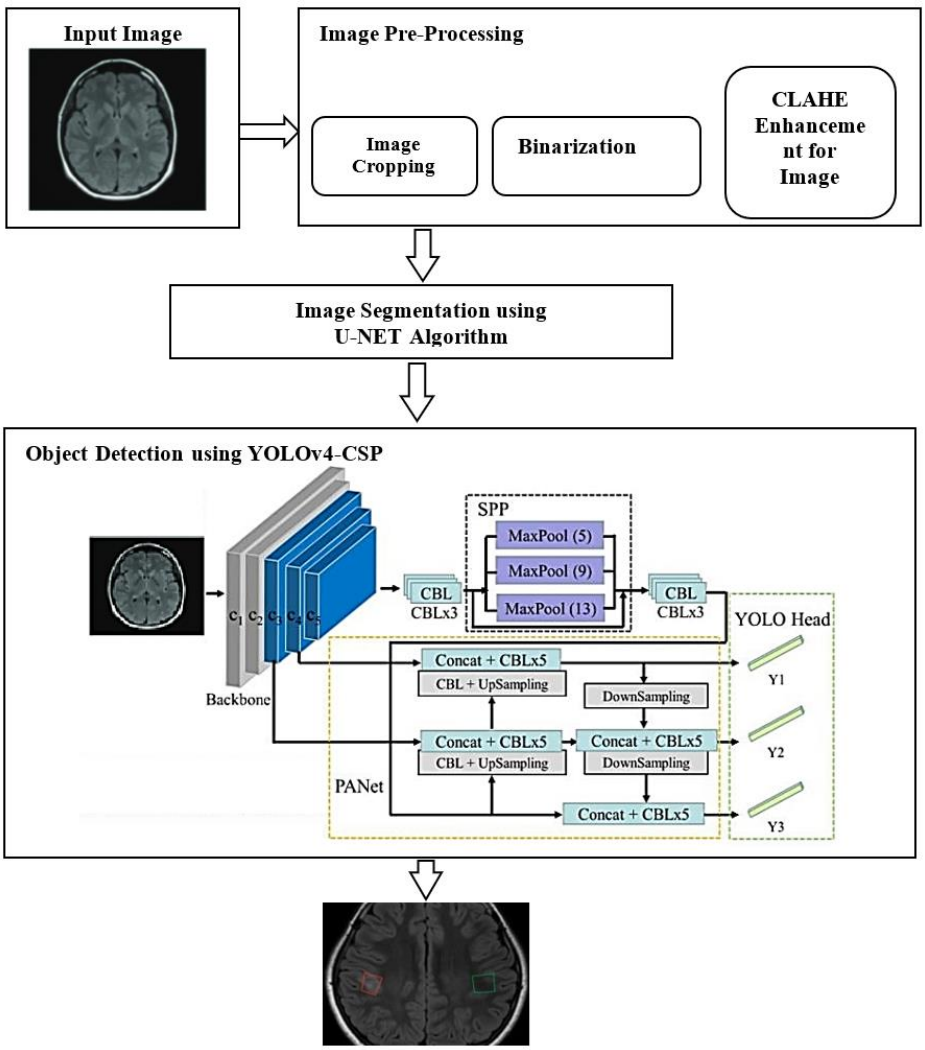

Figure 1. Architecture diagram of the proposed method

(2) Pre-processing

To optimize the input image for subsequent experimentation, this work engages in pre-processing techniques. This includes steps such as cropping, binarization using the Otsu method, and employing the CLAHE algorithm for image enhancement. These processes collectively serve to enhance image quality, eliminate salt and pepper noise, and smoothen the image, ensuring optimal conditions for object detection.

(3) Segmentation

The pre-processed image, now refined and prepared, undergoes segmentation. This is accomplished through the application of a U-NET. The primary objective here is to identify the lesions from a medical image.

(4) Object Detection using YOLOv4-CSP

The segmented medical image is fed into a YOLOv4-CSP algorithm. This sophisticated neural network is adept at object detection, leveraging its ability to learn hierarchical representations. This enables accurate object identification from a medical image.

4.1 Image pre-processing

Cropping:

Creating a cropped image entails isolating a specific rectangular Region of Interest (ROI) from an original image. Here's an alternative explanation for the process:

Procedure for Selecting and Cropping a Region of Interest (ROI) in an Image:

Step 1: Import the Original Image Begin by loading the original image into your preferred image processing software. This serves as the starting point for the cropping process.

Step 2: Define the Region of Interest (ROI) specify the ROI within the image by providing coordinates that outline a rectangular area. These coordinates determine the starting point (top-left corner) and dimensions (width and height) of the desired ROI.

Binarization using the Otsu Method:

Otsu's thresholding method is a widely used technique for automatically converting grayscale images into binary images. It operates by determining an optimal threshold that maximizes variance linking two classes of pixel intensities, effectively distinguishing foreground and background regions. Let's delve into the mathematical foundation of Otsu's thresholding algorithm. Consider a grayscale image represented by pixel intensities in the range [0, L-1], where L is the total number of possible intensity levels. The goal of Otsu's algorithm is to discover a threshold value, denoted as T, which partitions the image into two classes:

Background (B): Pixels with intensities less than T.

Foreground (F): Pixels with intensities greater than or equal to T.

The algorithm maximizes the separation between these two classes by leveraging the concept of intra-class variance and inter-class variance.

Step 1: Calculate the histogram of pixel intensities in the image, counting the frequency of each intensity value.

Step 2: Normalize the histogram so that the sum of all frequencies is 1.

Step 3: Compute the Cumulative Distribution Function (CDF) and the cumulative mean intensity up to each intensity level.

Step 4: Calculate the global mean intensity of an entire image

Step 5: For each intensity level k from 0 to L-1:

•Calculate the probabilities of background and foreground classes up to intensity level k.

•Calculate the mean intensities of the background and foreground classes up to intensity level.

•Calculate the between-class variance ($sigma_b^2$) using the formula:

$$sigma_b^2 = wb * wf * (mean_b - mean_f)^2 \quad (1)$$

where, w_b and w_f are the probabilities of the background and foreground classes, and $mean_b$ and $mean_f$ are the mean intensities of the background and foreground classes.

Step 6: Find the optimal threshold, The threshold that maximizes $sigma_b^2$ corresponds to the optimal threshold T.

Image Enhancement using CLAHE:

Contrast-Limited Adaptive Histogram Equalization (CLAHE) is a widely used algorithm for enhancing the contrast in medical images. It's particularly effective in scenarios where traditional histogram equalization may lead to overamplification of noise. Here's a step-by-step explanation of how the CLAHE algorithm is applied to enhance medical images:

$I(x,y)$: Original input image.

$M \times N$: Size of the image.

L: Intensity levels (e.g., 256 for an 8-bit image).

$S \times S$: Size of the contextual region (tile).

Step 1: Divide the Image into Non-Overlapping Tiles:

Divide the image $I(x,y)$ into non-overlapping tiles of size $S \times S$.

Step 2: Calculate the Histogram for Each Tile:

For each tile, calculate the histogram $H_k(m,n)$ for intensity levels 00 to L−1.

$$H_K(m,n) = \sum_{x^1=m}^{m+S-1} \times \sum_{y^1=n}^{n+S-1} I(x^1, y^1) \quad (2)$$

Step 3: Apply Histogram Equalization to Each Tile:

For each tile, apply histogram equalization to obtain the cumulative distribution function (CDF)

$$C_k(m,n) = cumsum(H_K(m,n)) / S^2 \quad (3)$$

Step 4: Apply Contrast Limiting:

For each tile, limit the contrast by clipping the values of the CDF$C_k(m,n)$ above a specified limit (clipLimit).

$$Ck(m,n) = \begin{cases} cliplimit \ if \ Ck(m,n) > cliplimit \\ Ck(m,n) \qquad\qquad otherwise \end{cases} \quad (4)$$

Step 5: Interpolate Overlapping Tiles:

Interpolate the contrast-limited histograms of overlapping tiles.

Step 6: Map Original Intensities to Enhanced Intensities:

For each pixel (x,y) in the image, find the corresponding histogram equalized intensity from the interpolated histograms.

$$I_{\text{enhanced}}(x,y) = C_k(m,n) \cdot (I(x,y) + 1) \quad (5)$$

It enhances the contrast in local regions of an image while

avoiding the overamplification of noise by limiting the contrast in each tile.

Figure 2 shows the original input image taken for pre-processing and Figure 3 shows the output after applying pre-processing steps cropping and Figure 4 shows the image after applying Binarization, Figure 5 shows the output image after applying median and Gaussian filtering process.

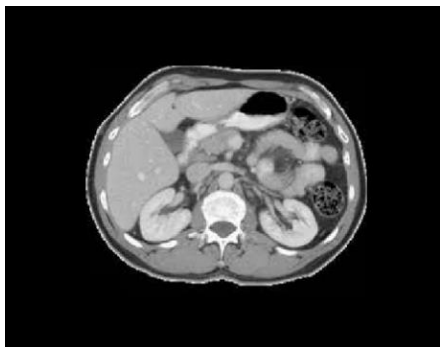

Figure 2. Original input image

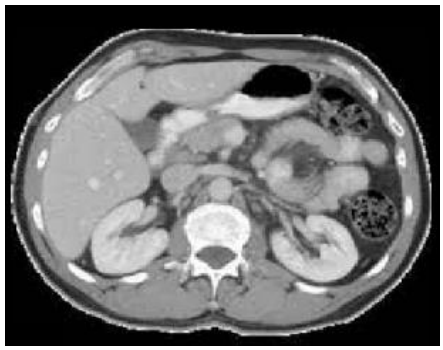

Figure 3. Pre-processed image (cropping)

Figure 4. Pre-processed image (binarization)

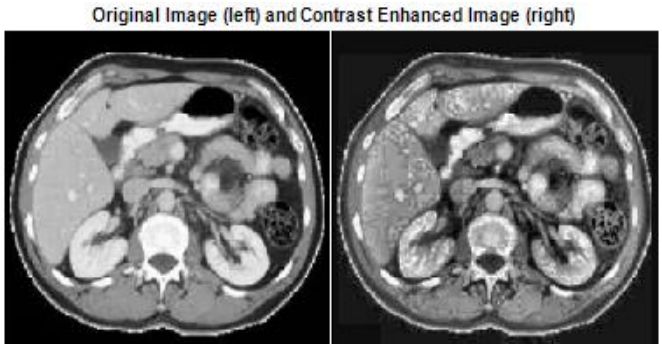

Figure 5. Pre-processing (CLAHE)

4.2 Image segmentation using U-Net algorithm

The U-Net algorithm, a convolutional neural network (CNN) architecture, is widely employed for image segmentation tasks, particularly in the domain of medical image analysis. Originally crafted for biomedical image segmentation, its primary objective is to discern and categorize distinct structures within an image, such as organs or tumors. Below is a detailed mathematical walkthrough of the U-Net algorithm:

The U-Net architecture is composed of two fundamental components: the contracting path (encoder) and the expansive path (decoder).

1. Contracting Path (Encoder):

Convolutional Block:

Let X be the input image.

Convolution: $Z_1=Conv2D(X, W_1)+b_1$, where W_1 is the filter, and b_1 is the bias.

Activation: $A_1=ReLU(Z_1)$.

Down-sampling (Max-pooling):

$P_1=MaxPooling2D(A_1)$.

Repeat Convolutional Block and Down-sampling:

$Z_2=Conv2D(P_1, W_2)+b_2$, $A_2=ReLU(Z_2)$.

$P_2=MaxPooling2D(A_2)$.

Repeat this process to create a contracting path with multiple convolutional blocks and down-sampling layers.

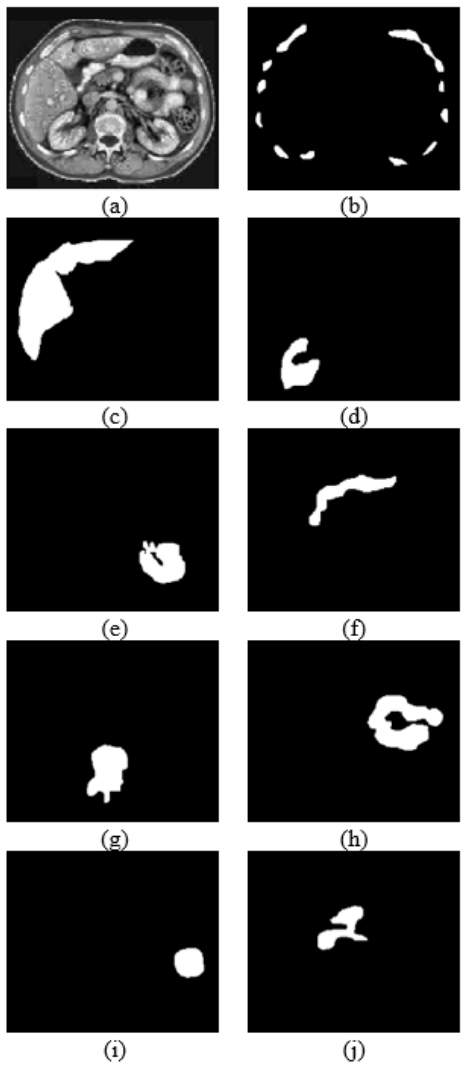

Figure 6. (a) Original image (b) to (j) Segmentation using U-Net algorithm

2. Bottleneck:

Central Convolutional Block:

$Z_n=Conv2D(P_n, W_n)+b_n$, $A_n=ReLU(Z_n)$.

3. Expansive Path (Decoder):

Up-sampling (Transposed Convolution):

$U_1=Conv2DTranspose(A_n, W_{up})+b_{up}$, where W_{up} is the transposed convolution filter, and b_{up} is the bias.

Concatenation and Convolutional Block:

Concatenate feature maps from the corresponding contracting path:

$C_1=Concatenate(U_1,A_2)$.

$Z_{n+1}=Conv2D(C_1,W_{n+1})+b_{n+1}$, $A_{n+1}=ReLU(Z_{n+1})$.

Repeat Up-sampling, Concatenation, and Convolutional Block:

Repeat this process to create an expansive path with multiple up-sampling, concatenation, and convolutional block layers.

4. Output Layer:

Convolutional Layer with Sigmoid Activation:

$Z_{output}=Conv2D(A_{expansive\_path},W_{output})+b_{output}$.

Apply a sigmoid activation function to obtain the final output:

$Y_{output}=Sigmoid(Z_{output})$.

Figure 6(a) shows the original image and Figures 6(b)-6(j) shows the output of the u-net segmentation algorithm which segments all objects from an input image that will be given as input to the next stage called object detection. Apart from that Figure 7 shows the output of contour-based segmentation which will help the experts to identify all the objects in a single image.

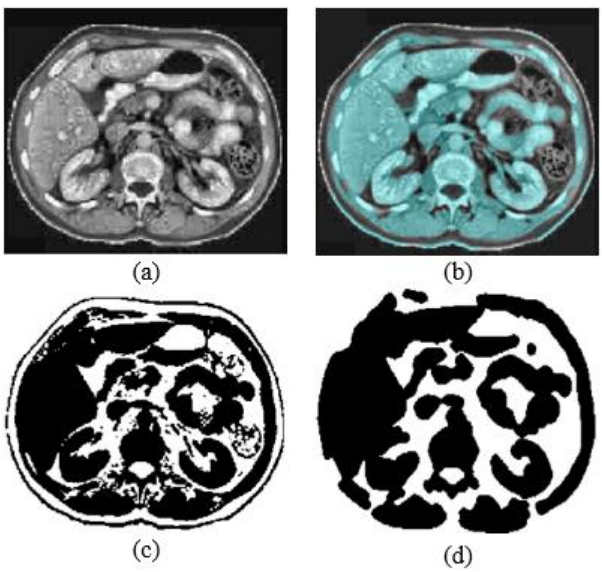

Figure 7. (a) Original image (b) Contour-based segmentation using U-Net algorithm (c) Binarized image of contour-based (d) Inverted image of contour-based for object identification analysis

Figure 7(a) shows the original image Figure 7(b) shows the contour-based segmentation using the U-Net algorithm, and Figures 7(c) and (d) show the binarized and inverted binarized image of the contour-based segmented image which will be used to identify the objects in the medical images accurately.

4.3 Object detection using YOLOv4-CSP algorithm

Renowned for its real-time object detection prowess, YOLO (You Only Look Once) excels by partitioning the input image into a grid and making predictions for bounding boxes and class probabilities within each grid cell. This unique approach allows for swift and accurate identification of objects in a single pass, distinguishing YOLO as a leading algorithm in real-time computer vision applications. CSPNet (Cross Stage Partial Network) is a network module that aims to improve the information flow across different stages of a neural network. YOLOv4-CSP incorporates CSPNet into the YOLOv4 architecture to enhance its performance.

Step 1: Input Image:

Let I be the input image with dimensions $H\times W\times C$, where H is the height, W is the width, and C is the number of channels.

Step 2: Backbone Network (*CSPDarknet53*):

-The backbone network processes the input image and produces feature maps at different scales.

-Let $F_{backbone}$ represent the feature maps generated by the backbone.

Step 3: **Feature Pyramid Network (*FPN*):**

-*FPN* processes the feature maps from the backbone to create a pyramid of features at different scales.

-Let F_{FPN} represent the feature maps from the *FPN*.

Step 4: Head Network:

-The head network processes the feature maps from FPN to make predictions.

-For each grid cell in the feature map, the head predicts bounding box coordinates *(x,y,w,h)*, confidence scores, and class probabilities.

-Let B represent the number of bounding boxes predicted per grid cell.

-Predictions for each grid cell:

$$P=(x,y,w,h,confidence,class1,class2,...,classN)$$

where, N is the number of classes.

Step 5: Anchor Boxes:

-YOLOv4-CSP uses anchor boxes to improve bounding box prediction accuracy.

-Let A represent the number of anchor boxes. Each anchor box has a width and height.

Step 6: Loss Function:

The loss function is defined as a combination of localization loss, confidence loss, and class loss.

Localization Loss:

$$L_{conf} = y_{coord}\sum_{i=0}^{B} 1_{obj_i} \left[(a_i-\hat{a}_i)^2+(b_i-\hat{b}_i)^2+\left(\sqrt{w_i}-\sqrt{\widehat{w}_i}\right)^2+\left(\sqrt{h_i}-\sqrt{\widehat{h}_i}\right)^2\right] \quad (6)$$

Confidence Loss:

$$L_{conf} = y_{coord}\sum_{i=0}^{B} 1_{obj_i} \left[(c_i-\hat{c}_i)^2+\left(\mu_{noobj}-\hat{\mu}_{inoobji}\right)^2+(c_i-\hat{c}_i)^2\right] \quad (7)$$

Class Loss:

$$Lclass = \lambda class\sum_{i=0}^{B} 1_{obj_i} \quad (8)$$

$$\sum_{c=1}^{N} 1_{obj_i} (p_i(c) - \hat{p}_i(c))^2$$

Total Loss:

$$L=L_{loc}+L_{conf}+L_{class} \quad (9)$$

where, obj_i is an indicator function that is 1 if object i is present in the grid cell and 0 otherwise.

Step 7: Non-Maximum Suppression (NMS):

Following the generation of predictions, a crucial refinement step is employed through non-maximum suppression, strategically implemented to discard redundant bounding boxes. This post-processing technique ensures the final selection of the most accurate and relevant bounding boxes, enhancing the precision and efficiency of the object detection results.

Step 8: Post-processing:

-Bounding box coordinates are converted from the grid cell space to the image space.

-Predictions below a certain confidence threshold are filtered out.

Figure 8 shows the output after applying the YOLOv4-CSP object detection algorithm. The output image clearly shows that the proposed algorithm after the U-Net segmentation process, identifies all the objects in medical images. This will help the medical experts and also patients to identify the parts. If measurements are taken then if any lesions are there that also will be identified very easily by the medical experts.

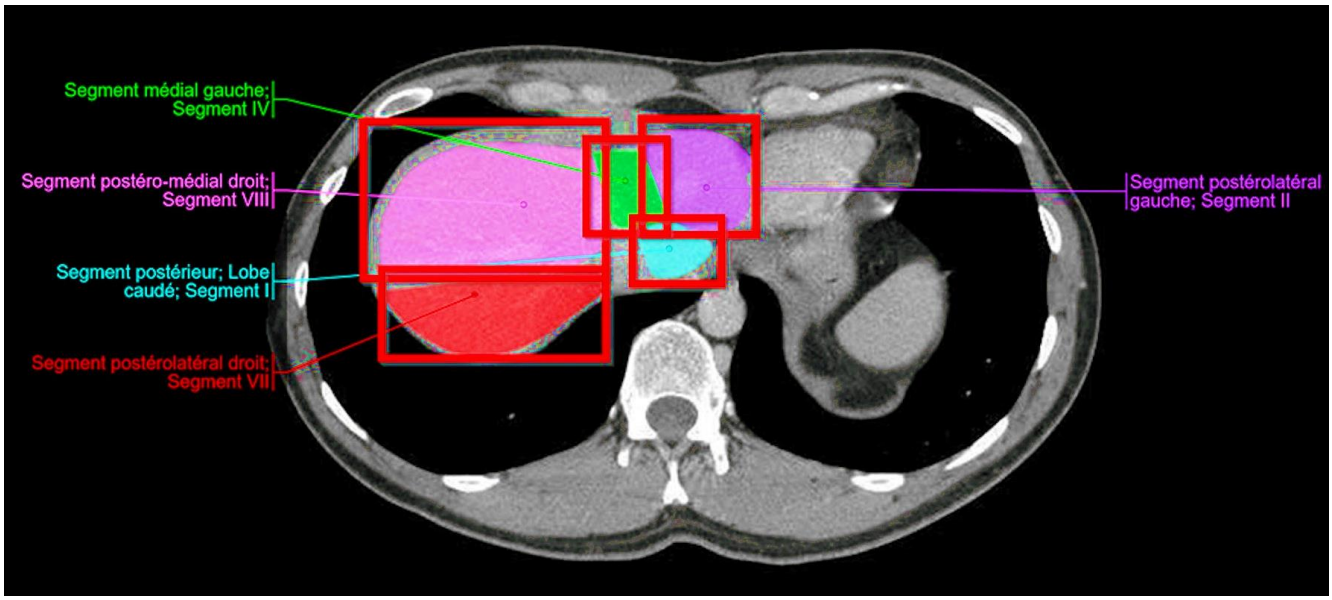

Figure 8. Output image of YOLOv4-CSP algorithm with identified objects

5. RESULTS AND DISCUSSION

The proposed methodology was realized using MATLAB, integrating the YOLOv4-CSP and U-Net segmentation algorithms. Utilizing Stochastic Gradient Descent (SGD), our system underwent an 8-epoch training regimen, initializing with a base learning rate of 0.005. Notably, this rate underwent a tenfold reduction after the 4th and 6th epochs. The inference time for predicting a sample stands at an impressive 30 milliseconds.

The experimentation hinged on the NIH DeepLesion dataset, encompassing 32,740 lesions. This meticulously partitioned the dataset into training (75%), validation (15%), and test (10%) sets at the patient level. Throughout the training phase, the proposed method incorporated data augmentation techniques into each image. These techniques comprised random resizing with a ratio of 0.75 to 1.0, random translation of -8 to 8 pixels in both the *x* and *y* axes, and 3D augmentation. Leveraging the unique annotation characteristics of NIH DeepLesion, where lesions span multiple slices, our 3D augmentation involved randomly shifting the slice index within half of the lesion's short diameter, resulting in a robust enhancement.

Each augmentation method, whether in resizing, translation, or 3D manipulation, contributed significantly to an augmented detection accuracy ranging from 0.25% to 0.45%. To assess detection performance, the proposed approach calculates sensitivities at 0.5, 1, 2, and 4 false positives (FPs) per image, averaging these metrics for a comprehensive evaluation of the system's efficacy.

This comprehensive evaluation strategy provides a robust assessment of the system's capabilities across different levels of false positives, ensuring a thorough understanding of its performance characteristics. To assess the performance of our proposed method, both qualitatively and quantitatively, we present the results in Figures 8 and 9 and Table 1, respectively.

Table 1. Detection % based on average sensitivity and segmentation distance of proposed methodology

Method Used	Detection % (Average Sensitivity)	Segmentation (Distance in mm)
ULDor [20]	69.22	1.47
MULAN [24]	86.12	1.41
AutoRECIST [21]	79.73	1.71
3DCE [22]	75.55	1.97
Proposed YOLOv4-CSP with U-Net	89.23	1.24

Table 1 emphasizes that segmentation accuracy was determined by predicting masks based on GT bounding boxes.

This approach ensures that segmentation assessments are conducted under the same settings and remain independent of detection accuracy, offering a nuanced understanding of the model's segmentation capabilities.

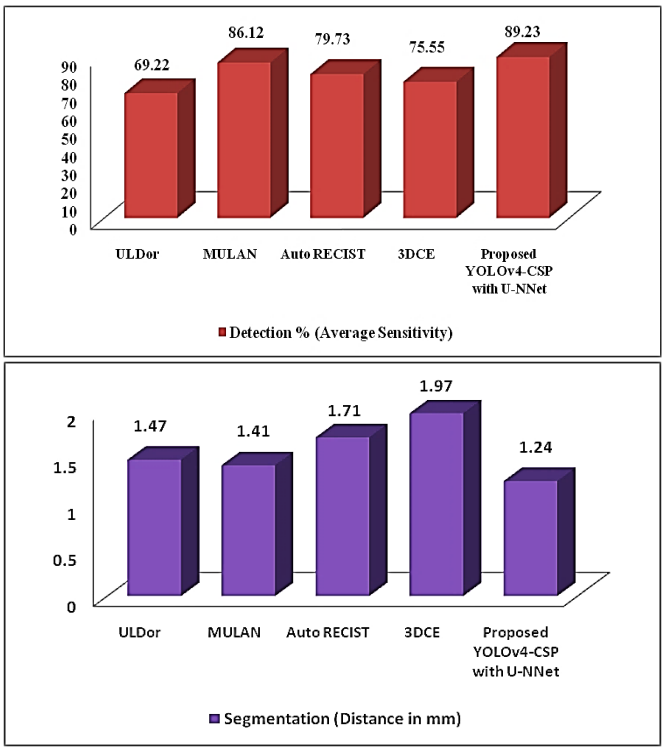

Figure 9. Detection % based on average sensitivity and segmentation distance of proposed methodology

Table 1 demonstrates that the proposed methodology outperforms existing approaches in universal lesion detection by a significant margin, exceeding them by more than 5% in average sensitivity. The conducted ablation study reveals that the incorporation of U-Net contributes the most to the enhancement in detection accuracy, resulting in an impressive 89.23%. Moreover, we ensured the robustness of our findings by conducting a thorough investigation. By randomly re-splitting the training and validation sets of the NIH DeepLesion dataset five times, our methodology consistently outperformed, showcasing the reliability and stability of the YOLOv4-CSP and U-Net combination.

Predicted diameters exhibit a minimal average error when compared to the ground truth (GT) diameters. Examining Figure 10, it becomes evident that the proposed method excels in accurately delineating lesions with well-defined borders. However, challenges arise when dealing with lesions featuring indistinct or irregular borders, presenting an area for improvement in segmentation accuracy. Notably, the detection task's impact on segmentation performance is noteworthy, potentially serving as a key factor in why the proposed method outperforms Auto RECIST, a dedicated framework for lesion measurement. This observation suggests that employing a U-Net may yield superior segmentation results, hinting at the potential for further advancements in our approach by refining the segmentation component.

For identifying objects in medical images, the YOLOv4-CSP with U-Net methods shows the possibility, but there are still a few limitations. These include possible complexity caused by merging two designs, challenges with precisely recognizing small or odd-shaped objects, and variable efficacy of the localized self-attention system in capturing contextual relationships. Important challenges also include limited interpretability, dependence on annotated data, and the high computational costs of training such models. It will take ongoing research to improve comprehension, optimize methods for training, and refine model architecture in order to get around these limitations and ensure the dependable and clinically useful use of deep learning-powered object recognition techniques in medical imaging.

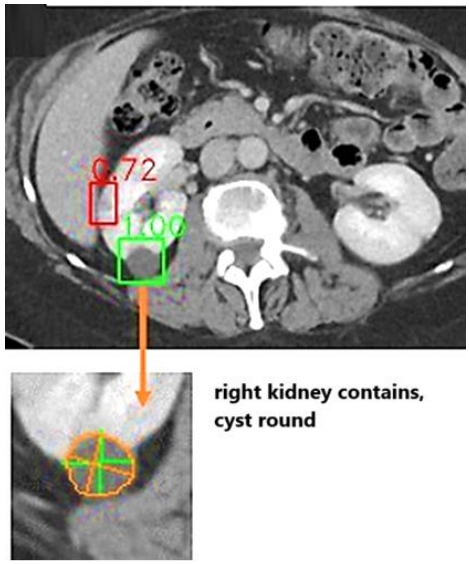

Figure 10. Output image - Object identification with lesion details

6. CONCLUSION

The integration of YOLOv4-CSP for object detection and U-Net (YOLO-UNET Fusion) for segmentation in medical image analysis has proven to be a promising and effective approach. The proposed method demonstrated notable success in lesion detection and segmentation, outperforming existing frameworks and showcasing competitive results in accuracy, particularly for lesions with well-defined borders. The proposed method achieved 89.23% of detection with average sensitivity. The minimal average error in predicted diameters compared to ground truth and the robustness of the methodology across various lesion types highlight its potential for practical clinical applications. However, challenges persist in accurately delineating lesions with indistinct or irregular borders, indicating a focus area for future improvements. Future work could involve the refinement of the segmentation component, potentially incorporating advanced architectures or fine-tuning existing ones, such as exploring variations of U-Net. Additionally, leveraging larger datasets or diverse datasets can enhance the model's generalization capabilities. Investigating the impact of different hyper-parameters and optimization strategies on model performance may also contribute to further advancements. Moreover, the integration of explainability features, such as attention mechanisms, could enhance the interpretability of the model's predictions, facilitating its acceptance and trust in clinical settings. Collaborations with medical experts for feedback and validation can provide valuable insights for refining the model's performance and ensuring its relevance in real-world medical scenarios.

REFERENCES

[1] Cao, W., Czarnek, N., Shan, J., Li, L. (2018). Microaneurysm detection using principal component analysis and machine learning methods. IEEE Transactions on Nanobioscience, 17(3): 191-198. https://doi.org/10.1109/tnb.2018.2840084

[2] Li, X., Xiang, J., Wang, J., Li, J., Wu, F.X., Li, M. (2020). FUNMarker: Fusion network-based method to identify prognostic and heterogeneous breast cancer biomarkers. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 18(6): 2483-2491. https://doi.org/10.1109/tcbb.2020.2973148

[3] Liao, B., Jiang, Y., Liang, W., Peng, L., et al. (2015). On efficient feature ranking methods for high-throughput data analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 12(6): 1374-1384. https://doi.org/10.1109/tcbb.2015.2415790

[4] Yi, J., Xiao, W., Li, G., Wu, P., He, Y.Y., Chen, C.M., He, Y.F., Ding, P., Kai, T.H. (2020). The research of aptamer biosensor technologies for detection of microorganism. Applied Microbiology and Biotechnology, 104: 9877-9890. https://doi.org/10.1007/s00253-020-10940-1

[5] Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, pp. 580-587. https://doi.org/10.1109/CVPR.2014.81

[6] Ren, S., He, K., Girshick, R., Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[7] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S. (2020). End-to-end object detection with transformers. In Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, pp. 213-229. https://doi.org/10.1007/978-3-030-58452-8_13

[8] Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929

[9] Liu, Z., Lin, Y., Cao, Y., Hu, H., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 9992-10002. https://doi.org/10.1109/ICCV48922.2021.00986

[10] Li, Z., Dong, M., Wen, S., Hu, X., Zhou, P., Zeng, Z. (2019). CLU-CNNs: Object detection for medical images. Neurocomputing, 350: 53-59. https://doi.org/10.1016/j.neucom.2019.04.028

[11] Yan, K., Bagheri, M., Summers, R.M. (2018). 3D context enhanced region-based convolutional neural network for end-to-end lesion detection. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, pp. 511-519. https://doi.org/10.1007/978-3-030-00928-1_58

[12] Liao, F., Liang, M., Li, Z., Hu, X., Song, S. (2019). Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky noisy-or network. IEEE Transactions on Neural Networks and Learning Systems, 30(11): 3484-3495. https://doi.org/10.1109/tnnls.2019.2892409

[13] Wang, S.H., Phillips, P., Sui, Y., Liu, B., Yang, M., Cheng, H. (2018). Classification of Alzheimer's disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling. Journal of Medical Systems, 42: 85. https://doi.org/10.1007/s10916-018-0932-7

[14] Xie, Y., Zhang, J., Shen, C., Xia, Y. (2021). COTR: Efficiently bridging CNN and transformer for 3D medical image segmentation. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, pp. 171-180. https://doi.org/10.1007/978-3-030-87199-4_16

[15] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M. (2022, October). Swin-UNet: UNet-like pure transformer for medical image segmentation. In Computer Vision – ECCV 2022 Workshops, Tel Aviv, Israel, pp. 205-218. https://doi.org/10.1007/978-3-031-25066-8_9

[16] Wang, W., Chen, C., Ding, M., Yu, H., Zha, S. (2021). TransBTS: multimodal brain tumor segmentation using transformer. arXiv preprint arXiv:2103.04430. https://doi.org/10.48550/arXiv.2103.04430

[17] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R. (2022). Masked autoencoders are scalable vision learners. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 15979-15988. https://doi.org/10.1109/CVPR52688.2022.01553

[18] Girshick, R. (2015). Fast R-CNN. In 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, pp. 1440-1448. https://doi.org/10.1109/ICCV.2015.169

[19] Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer C. (2021). Masked feature prediction for self-supervised visual pre-training. arXiv preprint arXiv:2112.09133. https://doi.org/10.48550/arXiv.2112.09133

[20] Tang, Y., Harrison, A.P., Bagheri, M., Xiao, J., Summers, R.M. (2018). Semi-automatic RECIST labeling on CT scans with cascaded convolutional neural networks. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, pp. 405-413. https://doi.org/10.1007/978-3-030-00937-3_47

[21] Tang, Y.B., Yan, K., Tang, Y.X., Liu, J., Xiao, J., Summers, R.M. (2019). ULDor: A universal lesion detector for CT scans with pseudo masks and hard negative example mining. In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, pp. 833-836. https://doi.org/10.1109/ISBI.2019.8759478

[22] Yan, K., Peng, Y., Sandfort, V., Bagheri, M., Lu, Z., Summers, R.M. (2019). Holistic and comprehensive annotation of clinically significant findings on diverse CT images: learning from radiology reports and label ontology. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 8515-8524. https://doi.org/10.1109/CVPR.2019.00872

[23] Umamageswari, A., Deepa, S., Raja, K. (2022). An

enhanced approach for leaf disease identification and classification using deep learning techniques. Measurement: Sensors, 24: 100568. https://doi.org/10.1016/j.measen.2022.100568

[24] Arasakumaran, U., Johnson, S.D., Sara, D., Kothandaraman, R. (2022). An enhanced identification and classification algorithm for plant leaf diseases based on deep learning. Traitement du Signal, 39(3): 1013-1018. https://doi.org/10.18280/ts.390328