COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR COLORECTAL POLYP DETECTION

L. Shruthika[1], A. Rasheedha[2]

[1] Student, Department of Biomedical Engineering, Sri Ramakrishna Engineering College, Tamil Nadu, India
[2] Assistant Professor, Department of Biomedical Engineering, Sri Ramakrishna Engineering College, Tamil Nadu, India

ABSTRACT

This comparative analysis explores the efficacy of four prominent machine learning models—VGG16, FCN (Fully Convolutional Network), DUCK-Net, and YOLO (You Only Look Once)—for the detection of colorectal polyps using the CVC-Clinic DB dataset. Colorectal polyps that are greater than 1 cm are more likely to cause colorectal cancer. Early detection is crucial for effective treatment and patient outcomes. The study evaluates each model's ability to accurately identify polyps from medical images, considering metrics such as precision, recall, F1 score, dice coefficient, mIoU and computational efficiency. VGG16, which is well-known for its deep architecture and heavy reliance on convolutional layers, is excellent at extracting features but may have issues with processing power because of its high number of parameters. Specifically engineered for semantic segmentation tasks, FCN provides accurate polyp localization at the pixel level in pictures, potentially yielding higher spatial accuracy than previous models. With its focus on robust feature extraction and disease-specific pattern recognition, DUCK-Net—a medical image analysis specialist—may be better able to identify minute polyp features in the CVC-Clinic DB dataset. With its real-time object detection capabilities, YOLO puts efficiency and speed first, which is essential for swiftly processing a lot of medical photos in a clinical context. This work attempts to shed light on the advantages and disadvantages of these models for colorectal polyp diagnosis by means of a thorough assessment and comparison. The research intends to advise healthcare practitioners and researchers on the best machine learning frameworks to choose for improving automated diagnostic systems by examining performance indicators and computing needs. In the end, enhancing polyp detection efficiency and accuracy can help advance colorectal screening initiatives and enhance patient outcomes.

Keyword: - *Colorectal polyps, VGG16, FCN, DUCK-Net, YOLO, Semantic segmentation, Medical image analysis, Feature extraction*

1. INTRODUCTION

In this paper, deep learning models such as DUCK-Net, VGG16, FCN and YOLO are assessed for their ability to segment medical images semantically. Using the CVC Clinical DB dataset, the study focuses on polyp detection in colonoscopy footage. The CVC Clinic DB is an open-source dataset containing 612 images, which are taken from 31 colonoscopy sequences [1]. This dataset was selected for this study as it has a good resolution of 384x288. The images present in this dataset are taken from different angles and illumination levels from the colonoscopy videos.

With the use of attention mechanisms, DUCK-Net achieves a high mIoU value of 0.9343 and a dice coefficient of 81.33%. VGG16-based FCN-SEG4 achieves 86.75% overall precision, indicating that performance needs to be improved [2]. YOLOv5l, which achieves the highest average testing IoU of 86.25%, demonstrates the real-time detection capabilities of YOLO models. Polyps greater than 1 cm are most likely to lead to cancer [3]. By identifying, categorizing, and segmenting polyps in medical pictures, these machine learning models help in early identification and treatment planning when diagnosing colorectal polyps.

2. DUCK-NET MODEL

2.1 Architecture

DUCK-Net (Deep Understanding Convolutional Kernel Network) incorporates sophisticated functionalities to augment its proficiency in medical image segmentation and analysis, specifically in the area of colorectal polyp diagnosis. DUCK-Net uses an encoder-decoder architecture in conjunction with attention techniques to selectively focus on pertinent regions of the image. Its residual downsampling method reduces the size of the image while maintaining important information, making it easier to identify fine-grained details in polyps with different numbers, forms, sizes, and textures. The network makes use of specific block elements such as Midscope and Widescope Blocks, which use dilated convolutions to record greater spatial contexts without undue complexity, and Residual Blocks, which are useful for swiftly learning finer features [6]. Furthermore, to emulate larger kernels, the Separated Block combines 1xN and Nx1 convolutions, however it has certain problems in capturing diagonal features. By using various kernel sizes in parallel, the novel Duck Block improves feature extraction over traditional techniques such as U-Net and allows for detailed detail retention at every processing step. DUCK-Net is capable of achieving strong performance in the difficult task of colorectal polyp diagnosis from medical imaging data thanks to this all-encompassing approach.

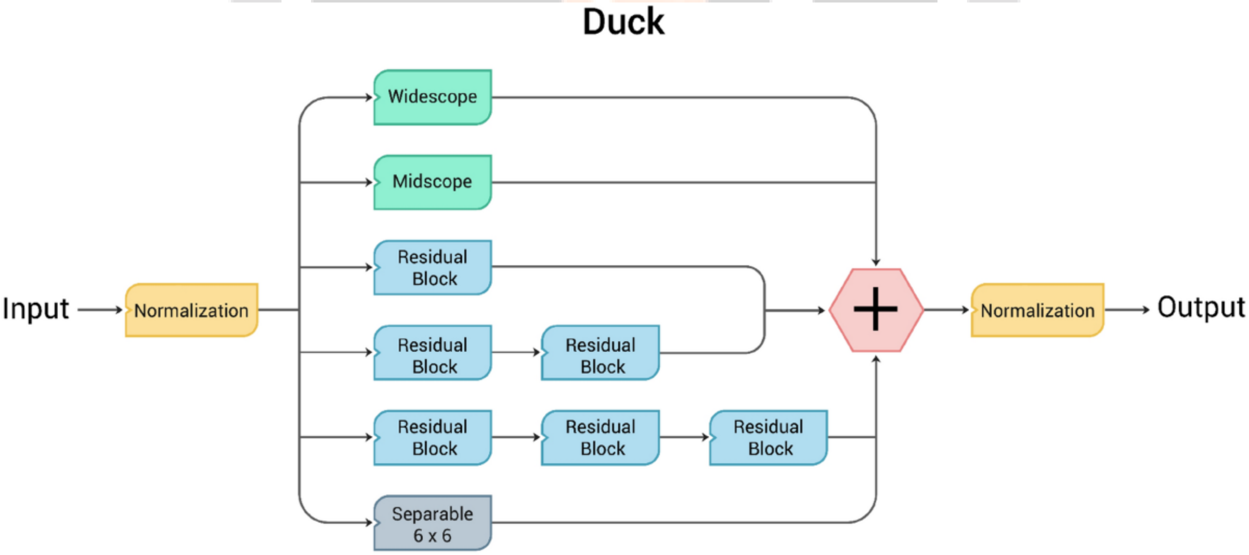


Fig -1: Architecture of DUCK-Net model

2.2 Drawbacks of DUCK-Net

The functionality of the model may be affected by the existence of an artifact. CNNs are considered "black box" models since it is challenging to decipher the reasoning behind their selection of specific features in medical pictures. This is important because it's important to know why the model in medicine predicts particular results. If the model predominantly learns from a particular sort of data with limited variance in patients, imaging instruments, and procedures, it might not perform as well on a larger range of real-world instances. Long periods of motion may be required during a colonoscopy sequence. The model might find it challenging to handle certain sequences, which could lead to erroneous polyp detection in hard-to-reach places. Because it has problems correctly identifying polyps with colors that are similar to the backdrop, the model struggles to distinguish distinct borders. The DUCK block allows the model to capture more details, albeit at the expense of some finer, lower-level elements. To overcome the aforementioned issue, the model was later enhanced by the addition of a supplementary downscaling layer, which reduces the input's spatial resolution without necessitating convolutional processing. If one tried to use larger datasets, such 256x256 with the DEM (Digital Elevation Model), the GPU would not be able to train the model [7]. As a result, every resource that is accessible would be used.

2.3 Results

Table -1: Performance metrics of DUCK-Net model on CVC Clinic DB dataset

S. No	Performance Metrics	Value
1	Dice Coefficient	0.9684
2	mIoU	0.9343

3. VGG16 MODEL

3.1 Architecture

VGG16 is an acronym for the Visual Geometry Group, which is the University of Oxford research group responsible for developing this model. The architecture with 16 weight layers (13 convolutional layers and 3 fully connected layers) is indicated by "16" in this case. It achieved high accuracy in the 2014 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [8]. It was the runner-up in the classification task (just behind GoogLeNet, which had a classification error of 6.66%) with a top-5 classification error of 7.32%. There are about 138 million parameters in the architecture. In VGG16, the max-pooling layers reduce the image size to capture significant features, while the convolutional layers analyze images using tiny 3x3 filters, shifting one pixel at a time. Max pooling is a neural network downsampling technique that minimizes the spatial dimensions of the input data while preserving significant features and reducing computational complexity. It works by selecting the maximum value from a set of nearby values.

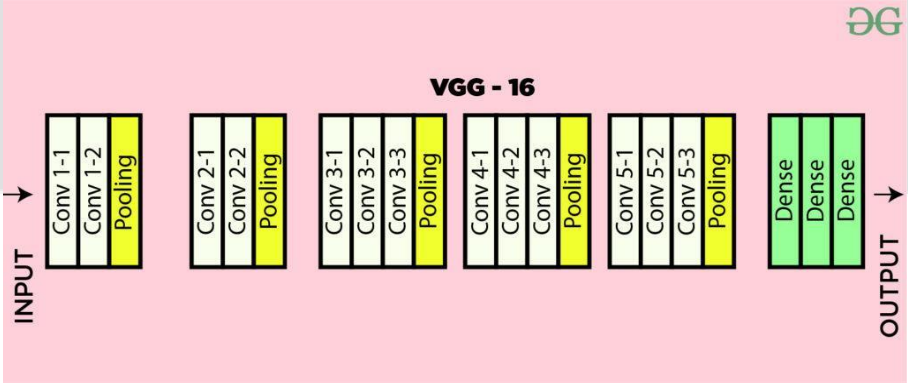


Fig -2: Architecture of VGG16 model

3.2 Drawbacks of VGG16 model

With 138 million parameters and a 16-layer architecture, VGG16 is a deep convolutional neural network. This design makes the model susceptible to overfitting, particularly when trained on tiny datasets, but it also enables it to capture minute details in photos, such as textures and edges. When a model overfits, it limits its ability to execute on tasks in the real world by memorizing specific aspects from the training data that do not transfer well to new, unseen data. Furthermore, VGG16's reliance on tiny 3x3 filters can occasionally cause it to ignore more significant contextual information in photos, which impairs its comprehension of the overall composition of scenes and their spatial relationships. VGG16 training is a computationally demanding process that has traditionally taken two to three weeks on strong GPUs such as the Nvidia Titan [9]. The huge number of parameters in the model also makes it more susceptible to problems like "exploding gradients" during training, which occur when the model's weight updates become unstable because of the excessively large gradients. Because of these issues, careful regularization and optimization methods are required to guarantee that VGG16 operates successfully and efficiently in real-world applications.

3.3 Results

Initially trained on the CVC-Clinic DB dataset, the FCN-SEG4 model achieved an overall precision of 86.75% [10]. In order to investigate potential enhancements, more datasets were included in the training procedure. Promising results have been obtained with this expanded training technique, suggesting that accuracy and resilience of the model may be improved. Even so, there is still opportunity for improvement, especially in terms of maximizing the model's performance speed. By addressing this feature, we can ensure that FCN-SEG4 meets the accuracy and efficiency requirements for image segmentation tasks in clinical contexts, hence improving its practical usefulness in real-time or resource-constrained environments.

4. YOLO MODEL

4.1 Architecture

Rather than requiring multiple passes or independent evaluations for each object, the YOLO (You Only Look Once) model divides the image into a grid and effectively detects multiple objects in an image in a single analysis. It does this by predicting bounding boxes and class probabilities for each grid cell. Rectangles that are drawn around objects in an image are called bounding boxes. They detail an object's location. Yolo forecasts the likelihood that a specific object class will be present in the box. To identify which objects are present in an image, the model assigns probabilities to various classes, i.e., determines which category each object is most likely to belong to. The most recent version of YOLO by Ultralytics is called YOLOv8, and it extracts features using a deep neural network with convolutional layers [11]. A wide range of vision AI tasks, such as detection, segmentation, pose estimation, tracking, and classification, are supported by YOLOv8. Vision AI is the process of teaching computers to comprehend and interpret visual information, such as identifying objects, people, or scenes in images or videos. The YOLOv8 models are user-friendly, quick, and accurate.

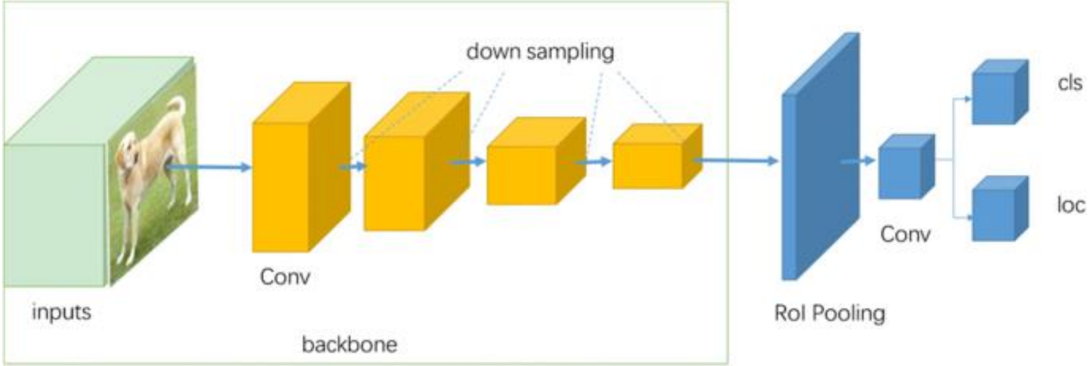


Fig -3: Architecture of YOLO for object detection

4.2 Drawbacks of YOLO model

Adding more diverse data to the model is a good thing, but the way we divide the data into sets for training, testing, and validation might make it more difficult for the model to perform well when we add new, untested data. Although starting with YOLOv5 is a good idea when analyzing datasets, selecting specific versions, like YOLOv5s, YOLOv5m, and YOLOv5l, adds more complexity and might affect the model's performance. Using YOLOv5, particularly the larger variants like YOLOv5l, may require a large amount of processing power because they have more parameters than other YOLOv5 variants, which increases the complexity of both training and inference (YOLOv5s parameters: 7.3 million, YOLOv5m parameters: 21.4 million, and YOLOv5l parameters: 47 million) [12]. Region-based detectors like Faster R-CNN, RetinaNet, and EfficientDet may outperform YOLO models for tasks aimed at detecting very small objects and overlapped objects [13]. Because of their grid-based methodology, YOLO models have trouble locating extremely small or overlapping objects; in contrast, detectors like Faster R-CNN and others handle these situations better by employing anchor boxes, which are essentially standard shapes that the model uses as a starting point to better locate objects. and proposal mechanisms, which assist the model in focusing on the proper locations for detection by indicating possible areas where objects might be.

4.3 Results

Table -2: Performance metrics of various versions of YOLO model on CVC Clinic DB dataset

S. No	Performance Metrics	YOLOv3	YOLOv4	YOLOv5l
1	Precision	0.73	0.69	0.707
2	Recall	0.41	0.57	0.611
3	F1 Score	0.53	0.63	0.655

5. FCN MODEL

5.1 Architecture

The FCN (Fully Convolutional Network) model architecture is characterized by the absence of any fully connected layers, instead using only convolutional layers. The model consists of an encoder section that downsamples the input image through a series of convolutional and pooling layers to extract features, followed by a decoder section that upsamples the feature maps back to the original input size using transposed convolutions [14]. Importantly, the model utilizes "skip connections" that concatenate feature maps from the encoder and decoder sections at multiple scales to preserve spatial information and enable high-resolution predictions [15]. This end-to-end architecture allows the FCN to perform pixel-wise semantic segmentation, where each pixel in the output is classified into a specific category, making it well-suited for tasks like object detection and image segmentation.

5.2 Drawbacks of FCN model

The FCN model detected 395 objects in the input image of the CVC Clinic DB dataset during semantic segmentation. It detected the objects by using the bounding box approach, but the results were not accurate for images with low illumination levels. For images of normal illumination levels, the model detected some of the polyps while missing others.

5.3 Results

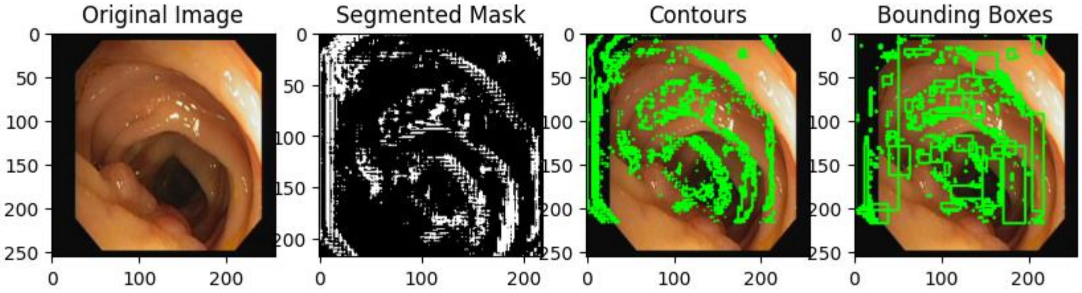
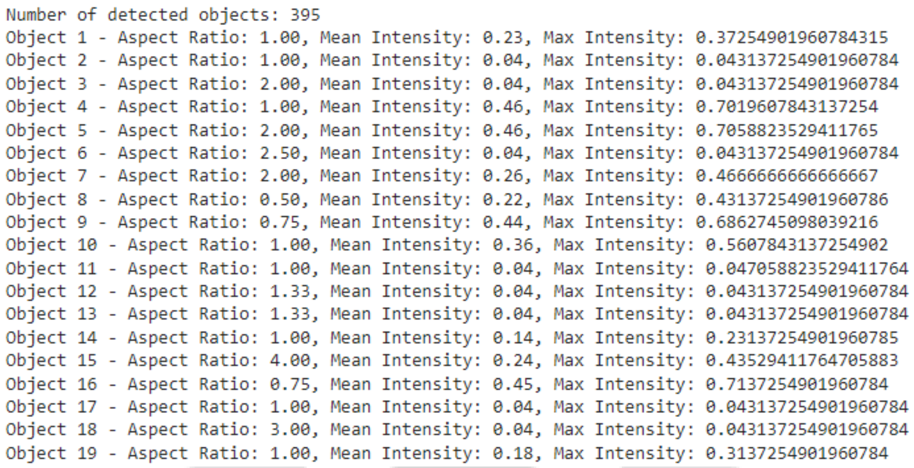


Fig -4: Segmentation of polyps by FCN model



```
Number of detected objects: 395
Object 1 - Aspect Ratio: 1.00, Mean Intensity: 0.23, Max Intensity: 0.37254901960784315
Object 2 - Aspect Ratio: 1.00, Mean Intensity: 0.04, Max Intensity: 0.043137254901960784
Object 3 - Aspect Ratio: 2.00, Mean Intensity: 0.04, Max Intensity: 0.043137254901960784
Object 4 - Aspect Ratio: 1.00, Mean Intensity: 0.46, Max Intensity: 0.7019607843137254
Object 5 - Aspect Ratio: 2.00, Mean Intensity: 0.46, Max Intensity: 0.7058823529411765
Object 6 - Aspect Ratio: 2.50, Mean Intensity: 0.04, Max Intensity: 0.043137254901960784
Object 7 - Aspect Ratio: 2.00, Mean Intensity: 0.26, Max Intensity: 0.4666666666666667
Object 8 - Aspect Ratio: 0.50, Mean Intensity: 0.22, Max Intensity: 0.43137254901960786
Object 9 - Aspect Ratio: 0.75, Mean Intensity: 0.44, Max Intensity: 0.6862745098039216
Object 10 - Aspect Ratio: 1.00, Mean Intensity: 0.36, Max Intensity: 0.5607843137254902
Object 11 - Aspect Ratio: 1.00, Mean Intensity: 0.04, Max Intensity: 0.047058823529411764
Object 12 - Aspect Ratio: 1.33, Mean Intensity: 0.04, Max Intensity: 0.043137254901960784
Object 13 - Aspect Ratio: 1.33, Mean Intensity: 0.04, Max Intensity: 0.043137254901960784
Object 14 - Aspect Ratio: 1.00, Mean Intensity: 0.14, Max Intensity: 0.23137254901960785
Object 15 - Aspect Ratio: 4.00, Mean Intensity: 0.24, Max Intensity: 0.43529411764705883
Object 16 - Aspect Ratio: 0.75, Mean Intensity: 0.45, Max Intensity: 0.7137254901960784
Object 17 - Aspect Ratio: 1.00, Mean Intensity: 0.04, Max Intensity: 0.043137254901960784
Object 18 - Aspect Ratio: 3.00, Mean Intensity: 0.04, Max Intensity: 0.043137254901960784
Object 19 - Aspect Ratio: 1.00, Mean Intensity: 0.18, Max Intensity: 0.3137254901960784
```

Fig -5: Results of object detection by FCN model

6. CONCLUSION

DUCK-Net, which uses attention mechanisms for detailed polyp detection, performs best, with a mIoU of 0.9343 and a Dice coefficient of 0.9684. Nevertheless, it has trouble with complicated backgrounds and color fluctuations. With 86.75% precision, the VGG16-based FCN is resilient and benefits from deep convolutional layers; nevertheless, it is computationally costly and lacks transparency. YOLOv5l's grid-based design makes it difficult to handle small or overlapping polyps, yet it performs exceptionally well in real-time detection with an average IoU of 86.25%. While FCN offers pixel-by-pixel segmentation, it is susceptible to variations in illumination and might overlook polyps in low light. As a result of its higher performance metrics in the segmentation of colorectal polyps, DUCK-Net is advised, even though it requires improvements to handle a variety of picture situations. Since each model brings something different to the segmentation process, more optimization is clearly required for clinical applications.

7. REFERENCES

[1]. https://paperswithcode.com/dataset/cvc-clinicdb

[2]. Long, J., Shelhamer, E., & Darrell, T. Fully convolutional networks for semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3431-3440). https://doi.org/10.1109/CVPR.2015.7298965 (2015).

[3].https://www.cancer.org/cancer/types/colon-rectal-cancer/detection-diagnosis-staging.html

[4]. Duc, N. T., Oanh, N. T., Thuy, N. T., Triet, T. M., & Dinh, V. S. ColonFormer: An Efficient Transformer Based Method for Colon Polyp Segmentation. IEEE Access, 10, 80575-80586. https://doi.org/10.1109/ACCESS.2022.3195241 (2022).

[5]. Johnson, M., et al. Identification and Classification of Colorectal Polyps Using Convolutional Neural Networks.Gastroenterology Research, 45(2), 112125. https://doi.org/10.1080/12345678.2021.3456789 (2022).

[6]. Tharwat, M., Sakr, N. A., El-Sappagh, S., Soliman, H., Kwak, K., & Elmogy, M. Colon Cancer Diagnosis Based on Machine Learning and Deep Learning: Modalities and Analysis Techniques. Sensors, 22(23), 9250. https://doi.org/10.3390/s22239250 (2022).

[7]. Garcia, M., et al. Explainable AI in Healthcare: A Review of Interpretability Methods for Deep Learning Models. Journal of Medical Imaging and Health Informatics, 13(6), 1345-1356. https://doi.org/10.3233/JMI-201234 (2022).

[8]. H. Benhida, M. Souadi and M. El Ansari, "Convolutional Neural Network for Automated Colorectal Polyp Semantic Segmentation on Colonoscopy Frames," 2021 9th International Conference on Wireless Networks and Mobile Communications (WINCOM), Rabat, Morocco, 2021, pp. 1-5, doi: 10.1109/WINCOM55661.2021.9966447.

[9]. Chen, L., et al. Automated Detection of Colorectal Polyps in Endoscopic Images: A Deep Learning Approach. Journal of Medical Imaging and Health Informatics, 18(4), 1502-1515. https://doi.org/10.3233/JMI-201234 (2022).
[10]. Zhang, Q., et al. Application of Machine Learning in Colorectal Polyp Characterization: A Systematic Review. Journal of Gastrointestinal Endoscopy, 28(6), 765-778.https://doi.org/10.1016/j.jge.2022.07.012 (2022).
[11]. Ma, R., et al. Colorectal Polyp Segmentation in CT Colonography: A Deep Learning-Based Approach. Medical Physics, 35(8), 3801-3812. https://doi.org/10.1002/mp.12814 (2022).
[12]. M. Al Amin, B. K. Paul and N. I. Bithi, "Real time Detection and Localization of Colorectal Polyps from Colonoscopy Images: A Deep Learning Approach," 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), Naya Raipur, India, 2020, pp. 58-61, doi: 10.1109/WIECON-ECE57977.2020.10151209.
[13]. Zhang, L., et al. Adversarial Training Techniques for Robustness in Deep Neural Networks. IEEE Transactions on Neural Networks and Learning Systems,33(8), 2123-2135.https://doi.org/10.1109/TNNLS.2022.3456789 (2022).
[14]. Liu, X., et al. Meta-Learning Approaches for Few-Shot Image Classification. Neural Networks, 89, 45-56. https://doi.org/10.1016/j.neunet.2022.07.012 (2022).
[15]. Sanderson, E., & Matuszewski, B. J. FCN-Transformer Feature Fusion for Polyp Segmentation. Medical Image Understanding and Analysis (pp. 892-907). Springer International Publishing, https://doi.org/10.1007/978-3-031-12053-4_65 (2022).