

Deep Learning Based Classification Of Gastrointestinal Lesions Using Wireless Capsule Endoscopic Images

Shoukat Ali

Times School of Computing, Times University, Multan, Pakistan

Email: shoukat210ali@gmail.com

Humayun Salahuddin*

Times School of Computing, Times University, Multan, Pakistan

Email: humayunsalahuddin@tum.edu.pk

Sadia Jafar

Computer Science, National College of Business Administration & Economics,
Multan, Pakistan Email: sadiajafar987@gmail.com

Muhammad Rafiq Mufti

Times School of Computing, Times University, Multan, Pakistan

Email: drmuhammadrafiq@tum.edu.pk

Aasma Ayub

Times School of Computing, Times University, Multan, Pakistan

Email: asmaayub1225@gmail.com

Hafsa Malik

Times School of Computing, Times University, Multan, Pakistan

Email: digitalhafsa10@gmail.com

Abstract

Although wireless capsule endoscopy (WCE) has proven as an important minimally invasive diagnostic tool in the field of gastrointestinal (GI) diseases, it presents enormous difficulties when examining these images and the variety of lesions encountered. The deep ensemble learning model for gastrointestinal lesion segmentation and detection are involved image preprocessing, lesion segmentation using deep learning, feature learning, and ensemble classification, which are designed to enhance the localization of lesions and diagnostic accuracy. For experimental evaluation we used the Kvasir-SEG dataset comprising of 1000 gastrointestinal polyp images, each accompanied by a segmentation mask. Three different types of CNN architectures (SegNet, U-Net and DeepLab) were trained to learn complementary spatial and contextual features independently. The segmentation results were then stacked to combine them together to improve the prediction accuracy with the help of a Random Forest Classifier (RFC). The individual models exhibited competitive

segmentation performance with U-Net performing the highest validation accuracy (92.96%) and SegNet performing best with Dice coefficient (0.7210) and IoU (0.5654). The proposed ensemble framework showed a significant improvement in performance with 99.95% accuracy, 99.91% F1-score, 99.85% ROC-AUC, 99.99% precision, and 99.70% recall. Confusion Matrix analysis also substantiated good lesion identification with only a few false-positive and false-negative predictions. The

Author Details

Keywords: Wireless Capsule Endoscopy; Gastrointestinal Lesions; Deep Learning; Segmentation; Ensemble Learning; SegNet; U-Net; DeepLab; Random Forest.

Received on 15 Feb 2026

Accepted on 20 Mar 2026

Published on 31 Mar 2026

Corresponding E-mail & Author*:

Humayun Salahuddin*

Times School of Computing,
Times University, Multan,
Pakistan Email:
humayunsalahuddin@tum.edu.pk

results show the effectiveness of heterogeneous ensemble learning that merges the different strengths of various segmentation architectures and offers an effective method for automated analysis of gastrointestinal lesions. The proposed framework can be applied to decrease diagnostic burden and to aid computer-aided diagnosis systems for clinical applications of WCE.

Introduction

Gastrointestinal (GI) disorders are a major health burden worldwide, as they are becoming more prevalent, complex to diagnose, and can lead to serious health complications. Delayed diagnosis of gastrointestinal cancers, inflammatory disease, bleeding disorders, ulcers and polyps can lead to progression of disease and sub-optimal clinical outcomes and therefore early and accurate diagnosis is essential (Caroppo et al., 2023). Several anatomical areas in the GIT, such as the oesophagus, stomach, small intestine, colon/rectum and associated digestive organs, can develop abnormalities within the mucosal layer that can be challenging to detect at early stages of disease development (Mughal et al., 2025; Afzal et al., 2026). GI malignancies continue to be one of the major contributors to cancer deaths globally, and reliable diagnostic strategies capable of early detection of pathological changes are crucial (Soffer et al., 2020; Ashraf et al., 2025).

The wireless capsule endoscopy (WCE) has become a sophisticated new minimally-invasive imaging modality that allows the imaging of areas within the gastrointestinal tract that are inaccessible to conventional endoscopy. During the process, the patient is given a small capsule with a video camera inside, which then passes through the digestive tract in the normal manner, taking thousands of photos. This technology has a lot of advantages for small bowel examination, where the conventional endoscopic examination is limited. Yet, although clinically relevant, WCE produces large amounts of imaging data, which is analyzed manually by gastroenterologists, resulting in high workload, interpretation time, and large potential for diagnosis variation, (Smedsrud et al., 2021).

One of the drawbacks of WCE is the huge number of acquired frames and the majority of the clinically significant abnormalities will only be seen in only a small number of frames. The entire capsule can produce several thousand images, and the physician will have to make extensive efforts to examine each one of them, frame by frame. Further, the accurate identification of lesions is difficult because of the variations in the light intensity, motion artifacts, capsule orientation, image blurring, and the fact that the normal mucosa and the pathological areas are similar. These restrictions underscore the need for the creation of computer-aided diagnosis (CAD) systems that can notify clinicians of abnormal frames and minimize misinterpretations, especially when they are performed automatically on the computer (Muruganantham & Balakrishnan, 2022).

Most of the traditional computer vision methods used for gastrointestinal image analysis were based on manually designed features such as texture, color distribution, shape descriptors, edge information. These methods had moderate success but were limited in their utility due to the inability of manually designed features to completely capture the rich visual features of gastrointestinal lesions. Conventional machine learning methods lack in generalization ability due to the differences in lesion morphology, size, colour intensity and anatomical location (Nguyen et al., 2020).

Deep learning especially convolutional neural networks (CNNs), medical image analysis has seen a transformation of sorts, with the ability to extract hierarchical features automatically from raw images. Unlike the traditional methods, CNN-based models are able to capture discriminative patterns directly from a large number of image databases, which enables better detection and classification of complex pathological structures (Zahid et al., 2025). In the field of gastrointestinal image analysis, deep learning methods have shown potential in various tasks such as lesion

detection, segmentation, localization, and disease classification (Fatima et al., 2026). According to recent reviews, deep learning techniques are performing well in WCE applications, such as bleeding, ulcer, polyp and inflammatory lesion detection, but issues of dataset diversity, interpretability and clinical validation are yet to be addressed (Soffer et al., 2020).

The CNN-based methods for automated gastrointestinal lesion analysis have been investigated by several studies. Large-scale deep learning system to diagnose small bowel diseases, based on images obtained from capsule endoscopy. High sensitivity in patients (99.88%) and lesions (99.90%) was achieved by their model, which was trained on millions of WCE images. Overall, the study demonstrated that deep learning models can effectively decrease the workload of capsule endoscopy interpretation and achieve high diagnostic accuracy (Ding et al. 2019).

For automatic detection of erosions and ulcerations in WCE images. Their model showed that it could detect subtle but noticeable changes in the mucous membrane and had a good diagnostic performance, suggesting the feasibility of CNN architectures for automated lesion screening. Likewise, deep convolutional neural network (CNN) to detect bleeding in capsule endoscopy images with high accuracy and AUC of 0.9998, indicating the efficacy of deep learning in bleeding detection (Aoki et al. 2020).

Pretrained CNN architectures have also contributed to the enhancement of the classification of WCE using transfer learning-based approaches. The performance of various pretrained models such as VGG16, ResNet18 and GoogLeNet on gastrointestinal disease classification with wireless endoscopy images. Their findings showed that it is possible to achieve good performance in the classification of several GI abnormalities using pretrained CNN models with a reduction in the amount of training required. Single CNN models however fail to perform well in complex variations of lesion and limited generalization across datasets (Yogapriya et al. 2021).

In recent years, the research has concentrated on hybrid deep learning, attention-based deep learning, and ensemble deep learning, to address the weakness of the CNN models. A hybrid CNN model with a meta-feature learning approach for abnormality detection in WCE images. The proposed model has improved the classification performance by combining more than one feature extraction technique, which shows the benefits of using feature fusion approaches (Jain et al. 2022).

Artificial intelligence system with CNN architecture for detecting the blood content in capsule endoscopy images. With an accuracy of 98.5% and a precision, sensitivity, and specificity of 98.7%, 98.6%, and 98.9%, respectively, the model is highly effective, demonstrating the potential of AI to streamline clinical workflows in capsule analysis (Saraiva et al. 2022).

In addition, deep learning using segmentation has become significant as precise localization of lesion regions is expected to enhance the classification performance. They have been using the U-Net and SegNet and DeepLab architectures for medical image segmentation which can extract spatial and contextual information widely because of being; more closely related to real world problems, more intuitive to network structure, and more easy to implement. DeepLabv3+-based method with multi-scale feature extraction to segment gastrointestinal lesions, which resulted in better localization than traditional GI segmentation methods. In the same way, U-Net based approaches have shown good performance in the detection of small regions of the lesion, but there are issues with unclear boundaries and artifacts from motion (Lee et al. 2023).

In recent years, ensemble learning methods have shown impressive results when using complementary deep feature representations. A hybrid stacking ensemble model for the classification of gastrointestinal abnormalities with accuracy of 98.53% and AUC of 99.95% on HyperKvasir dataset (Sivari et al. 2023). Deep learning models for multiclass classification of ulcerative colitis, polyps and dyed-lifted polyps from

WCE images. By using advanced CNN architectures along with optimized classification strategies, they achieved 99.45% accuracy, which is the best performance of their architectures (Malik et al. 2024).

Recently, attention mechanisms and feature fusion methods have further improved gastrointestinal lesion classification. A network-level fused deep learning model that combines DenseNet201, self-attention mechanisms, CNN-GRU networks and feature optimization methods. The proposed framework obtained 99.60% accuracy, showing that multi-level feature fusion can boost the robustness and classification ability.

There are some research challenges that yet remain: Previous research often relies on a specific CNN design, a restricted amount of lesions, or rigorously curated datasets. Such factors as illumination variation, camera-movement, low contrast, noise, and class imbalance produce significant variations in WCE images. Moreover, many models only address the classification without considering the location and refinement of the lesions. Imbalanced data sets are still a big challenge because the image of an abnormal lesion is less in number than the normal image and thus the models are biased towards majority classes (Khan et al., 2025). Hence, the adoption of advanced frameworks that combine segmentation, feature fusion and ensemble classification is crucial to build more reliable and clinically viable CAD system.

To overcome these shortcomings, this research suggests a framework for wireless capsule endoscopic image-based gastrointestinal (GI) lesion classification using deep learning. The proposed methodology includes image preprocessing, augmentation, segmentation of lesion region based on SegNet, U-Net and DeepLab model architectures, deep feature extraction, feature fusion and ensemble-based classification. In this paper, by means of complementarities of feature representations and optimized classification methods, an effective computer-assisted diagnostic solution for gastrointestinal disease analysis is proposed to achieve more accurate recognition of lesions and to decrease the workload of the doctors' diagnosis.

Materials and Methods

Overview of the Proposed Deep Learning Framework

In this study, an automated computer-aided diagnosis (CAD) system for the detection and classification of gastrointestinal (GI) lesions from wireless capsule endoscopy (WCE) images using deep learning was designed. The overall design combines image preprocessing, lesion segmentation, deep feature learning, and ensemble-based classification, which are all designed to enhance the accuracy of automated lesion localization and diagnosis. To reduce variations between images due to illumination, noise, and acquisition artefacts, standardisation of image quality was performed by preprocessing techniques. In order to obtain multi-level spatial, hierarchical and contextual lesion features, 3 advanced segmentation architectures were used: Seg-Net, U-Net and DeepLab. The pixel-level segmentation maps were then fused together using ensemble learning approach to improve the robustness, stability and generalization ability of the prediction for automated gastrointestinal lesion analysis.

Dataset Description and Data Preparation

The proposed deep learning framework was designed, and assessed on the publicly available gastrointestinal endoscopy dataset Kvasir-SEG with 1000 gastrointestinal polyp images and manually labeled segmentation masks. To guarantee accurate localization of lesions at the pixel level and accurate training and evaluation of the models, each image was associated with its ground truth mask. The dataset contained a wide variety of gastrointestinal diseases and varying image resolutions (332×487 to 1920×1072), which allowed for clinical imaging variability. In order to be compatible with the proposed deep learning architectures, all RGB images and masks were resized to $256 \times 256 \times 3$ before being implemented into the model.

Image Preprocessing

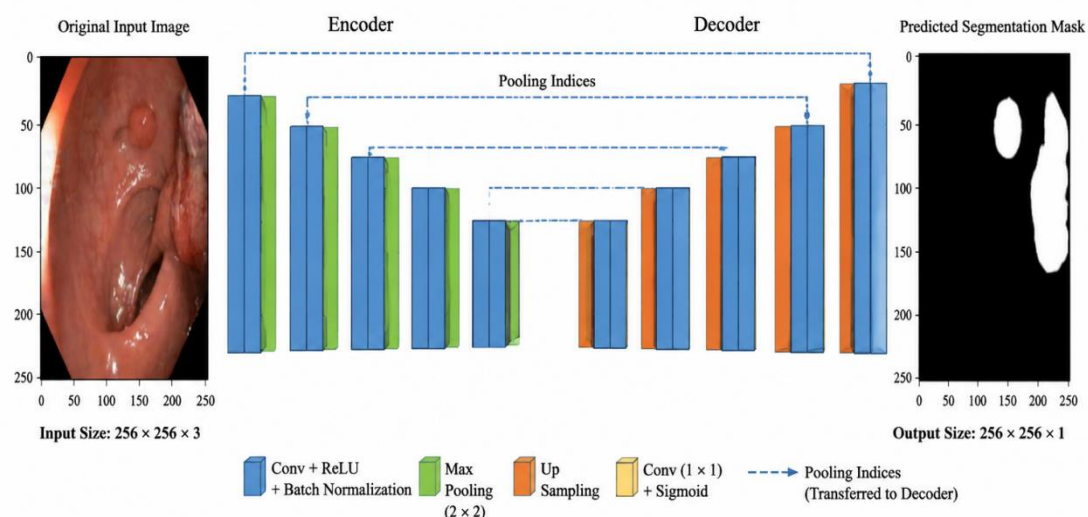
In medical image analysis, image preprocessing is a crucial stage that enhances the quality of medical images and provides uniform input for deep learning models. In typical WCE images, variations are frequently found due to the illumination variations, noise, motion artefacts, colour distortion and acquisition conditions. The preprocessing involved converting RGB images to retain the pathological color information, resizing images to $256 \times 256 \times 3$ for CNN compatibility, and standardizing the masks to binary images. To minimize between-pixel fluctuations and stabilize the model, pixel normalization was performed to focus on the features of the lesions. Other enhancement methods, such as noise reduction, contrast enhancement and illumination correction, further enhanced the visibility of the lesions and the definition of their boundaries. The processed images and masks were then fed into Seg-Net, U-Net and DeepLab segmentation models.

Deep Learning-Based Lesion Segmentation Framework

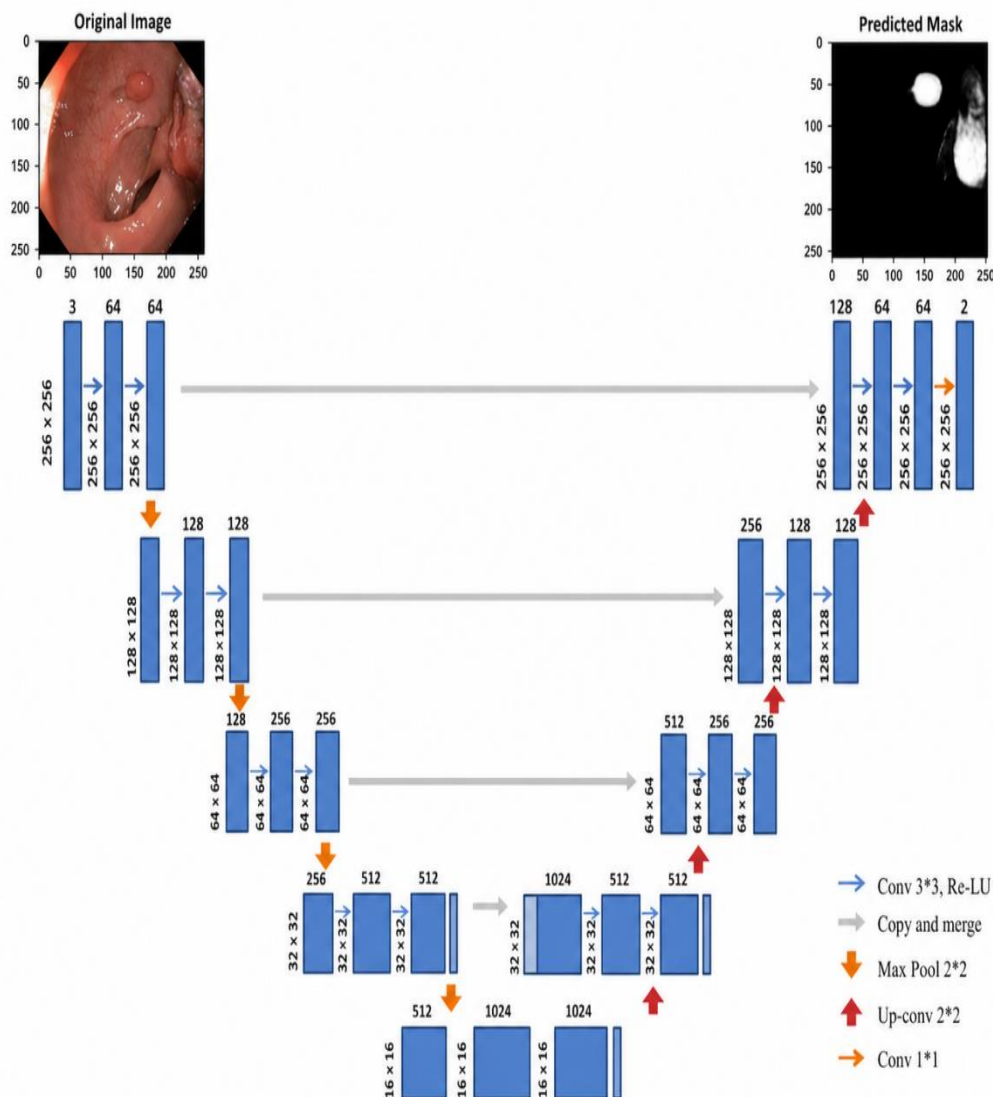
The gastrointestinal segmentation was performed using three CNN-based segmentation architectures: Seg-Net, U-Net and DeepLab, to accurately locate gastrointestinal lesions in WCE images. The models were trained independently to learn the complementary spatial, hierarchical and contextual features. Their segmentation probability maps were combined in an ensemble technique to improve the robustness, accuracy and generalization of the classification.

Seg-Net Architecture

Seg-Net is an encoder–decoder based CNN architecture for the semantic segmentation of gastrointestinal lesions in the WCE images. The model is organized into three parts: an encoder network for extracting hierarchical features, a decoder network for reconstructing the image at the spatial level and a classification layer per pixel. Convolutional and pooling operations enable the encoder to learn discriminative features, such as boundaries of lesions, texture patterns, colour variations and structural features. The decoder reconstructs spatial information by employing up-sampling operations to ensure spatial accuracy in lesion localization. An important aspect of Seg-Net is the employment of max-pooling indices for the purpose of keeping significant spatial information during the decoding process and therefore enhancing the segmentation accuracy. A 1×1 convolution layer with sigmoid activation function was used to create a probability map of lesion and background regions for binary lesion segmentation. Adam Optimizer with learning rate (0.0001) and binary cross-entropy loss are used to train the model. The architecture had 3,925,057 parameters of which 3,922,113 were trainable.



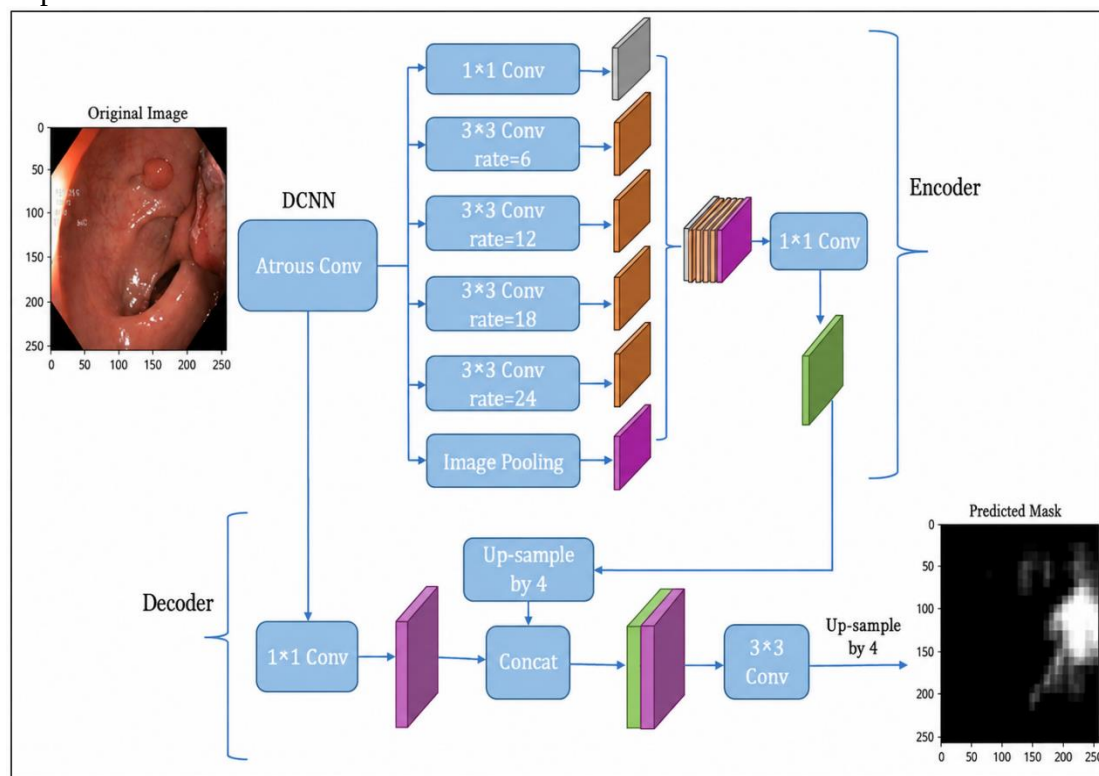
U-Net is a popular architecture for biomedical image segmentation because of its outstanding capability of retaining spatial information while providing high-level semantic information. The model is based on an encoder–decoder architecture with the encoder learning hierarchical representations by performing convolution and pooling operations, and the decoder reconstructing high-resolution segmentation maps. The early layers are used to detect low-level features like edges, color variations and textures, while the deeper layers are used to detect the complex features of the lesions. The key idea behind U-Net is the skip connections, connecting corresponding layers between the encoder and decoder, which enhance the localization of the boundaries and reduce the loss of information. A binary mask of lesions was obtained using a 1×1 convolution with a sigmoid activation. The model was trained with Adam optimization, binary cross-entropy loss and 31,043,521 parameters of which 31,037,633 parameters were trainable.



DeepLab Architecture

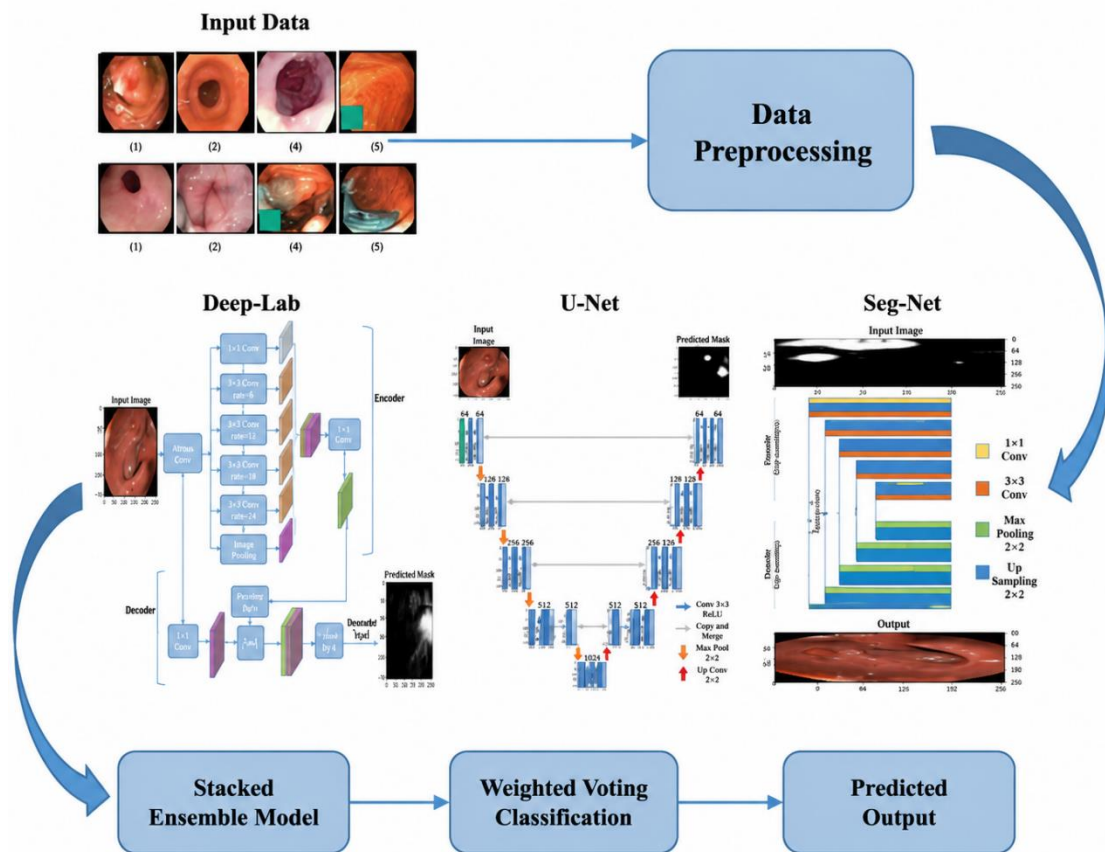
DeepLab is a CNN-based semantic segmentation architecture for the purpose of accurately localizing the lesions by capturing spatial details and multi-scale contextual information. DeepLab was used in this study because of its capability to detect gastrointestinal lesions from WCE images of various sizes, shapes, texture, and appearance through the use of atrous convolution. The architecture consisted of three main components: an encoder network, Atrous Spatial Pyramid Pooling (ASPP) module, and decoder network. The encoder was responsible for extracting hierarchical

semantic features, ASPP module for capturing contextual information at multiple scales, and the decoder for reconstructing the spatial information that is required for high-resolution segmentation. The model was fed with RGB images with a dimension of $256 \times 256 \times 3$. A fusion layer intermediate and high-level encoding is installed to improve spatial and semantic representation. The ASPP module was used to apply a series of atrous convolutions with dilation factors of 6, 12, 18 and 24, which allowed for multi-scale features to be extracted without losing resolution. The decoder performed convolution and up-sampling on the segmentation outputs and generated binary masks of $256 \times 256 \times 1$ with the size of the convolution 1×1 and the activation function being Sigmoid. The model was trained with 15,331,619 trainable parameters and Adam optimization (learning rate: 0.0001) and binary cross-entropy loss achieved effective learning. To complement Seg-Net and U-Net, DeepLab improved the contextual feature extraction in the ensemble framework.



Ensemble Learning Framework

While segmentation performance might be good for individual deep learning models, the diversity in gastrointestinal lesion morphology, such as size, shape, texture, and boundaries, can decrease the versatility of single architectures. Thus, to realize the complementary feature-learning capabilities of Seg-Net, U-Net and DeepLab, an ensemble learning framework was developed. The data of the preprocessed WCE image was used to train each model independently, and Seg-Net was found to be efficient in the reconstruction of the pixel level, U-Net was found effective in the spatial localization through the skip connections, and DeepLab was effective for multi-scale contextual features through the atrous convolution. A meta-model was trained to determine the optimal combination of the outputs of the probability maps and segmentation outputs, which were then combined in a stacking ensemble approach. This strategy enhanced lesion localization, classification accuracy, robustness, and generalization while considering the additional computational requirements of multiple model training and evaluation.



Model Training and Optimization

Standard optimization methods were used for training the proposed deep learning models for consistency and adequate performance evaluation. Adam optimizer was used for each segmentation architecture, with a learning rate of 0.0001. This task was a binary segmentation of lesion and background regions, so binary cross-entropy loss was chosen to minimize the difference between the predicted segmentation mask and the ground truth annotation. The models were automatically trained to continuously fine tune the parameters of the network, so that discriminative feature representations could be learned from gastrointestinal images in the process. This helped to detect the edges of the lesions, classify the pixels correctly and improve segmentation results. The results from Seg-Net, U-Net and DeepLab were ensemble in the proposed framework after the model training process. The integrated prediction approach leveraged the multi-modal feature representation of different structures and enhanced the diagnostic accuracy, robustness and generalization of the gastrointestinal lesion detection system.

Performance Evaluation Metrics

The performance of the proposed framework was evaluated by commonly used evaluation metrics in medical image segmentation and classification. The measures were chosen to achieve a complete assessment of lesion detection accuracy, sensitivity and reliability.

The following performance measures were used:

Accuracy

Accuracy represents the overall proportion of correctly classified pixels or images:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives.

Sensitivity (Recall)

Sensitivity measures the ability of the model to correctly identify lesion regions:

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

Specificity

Specificity evaluates the capability of the model to correctly identify non-lesion regions:

$$Specificity = \frac{TN}{TN + FP}$$

Precision

Precision indicates the proportion of correctly detected lesions among all predicted lesion regions:

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

F1-score

The F1-score provides a balance between precision and sensitivity:

$$F1 = \frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity}$$

These evaluation parameters enabled quantitative comparison between individual segmentation models and the proposed ensemble framework.

Computational Environment and Implementation Details

The proposed deep learning framework was implemented using a Python-based computational environment supported by advanced machine learning libraries for medical image analysis. Python was utilized for data preprocessing, model development, training, optimization, and performance evaluation. NumPy and Pandas were employed for numerical computation and data management, while Matplotlib was used for visualization and performance analysis. The segmentation models, including Seg-Net, U-Net, and DeepLab, were developed and trained using the TensorFlow framework, enabling efficient CNN implementation and GPU-accelerated computation. Experimental execution was performed using Google Colaboratory (Google Colab), providing GPU/TPU resources and a reproducible cloud-based environment for computationally intensive deep learning experiments.

Results

The results obtained from the proposed gastrointestinal lesion segmenter based on deep learning. The experimental setup consisted of three different semantic segmentation architectures: SegNet, U-Net and DeepLab, and then a stacking model with a Random Forest Classifier (RFC). The purpose of the experiments was to evaluate the ability of the single segmentation architectures to detect the gastrointestinal polyp regions, and whether the predictions of the different architectures can be combined (ensemble learning) to improve the segmentation achieved.

The dataset used for the experimental evaluation was the Kvasir-SEG consisting of 1000 gastrointestinal polyp images and their ground truth segmentation masks. Before training the models, the images were resized to 256 x 256 pixels. Before using the images to train the models, they were resized to 256x256 pixels. The data set has been divided to 800 training images, 100 validation images and 100 test images. The validation subset (compared to the current notebook) is used for the computation of comparative metrics and the results of the ensemble by the Random Forest.

The models, SegNet, U-Net and DeepLab were trained independently for 30 epochs using Adam optimizer and binary cross-entropy loss function. The accuracy, F1-score, Dice coefficient and Intersection over Union (IoU) were recorded during the model training stage. The three segmentation networks were trained and the output of the three networks were merged and passed to a meta-classification network based on Random Forest. The accuracy, F1 score, ROC-AUC, precision and recall were then computed for each individual model, as well as the ensemble model.

Dataset Distribution and Preprocessed Samples

The Kvasir-SEG dataset consists of 1000 pairs of gastrointestinal images and polyp masks of pixels. This was 80:10:10 of training, validation and test samples. The model parameters were optimized on the larger training set and during training and ensemble calculations performed in the present experimental notebook a validation set was used.

Table 3.1: Distribution of the Kvasir-SEG Dataset

Dataset subset	Number of images	Percentage
Training	800	80%
Validation	100	10%
Testing	100	10%
Total	1,000	100%

The paired image and annotation format of Kvasir-SEG enabled comparison of masks produced by Kvasir-SEG with the manually annotated masks of lesions, performed on a pixel-by-pixel basis. The resulting image and the resulting mask from the notebook are shown in Figure 3.1.

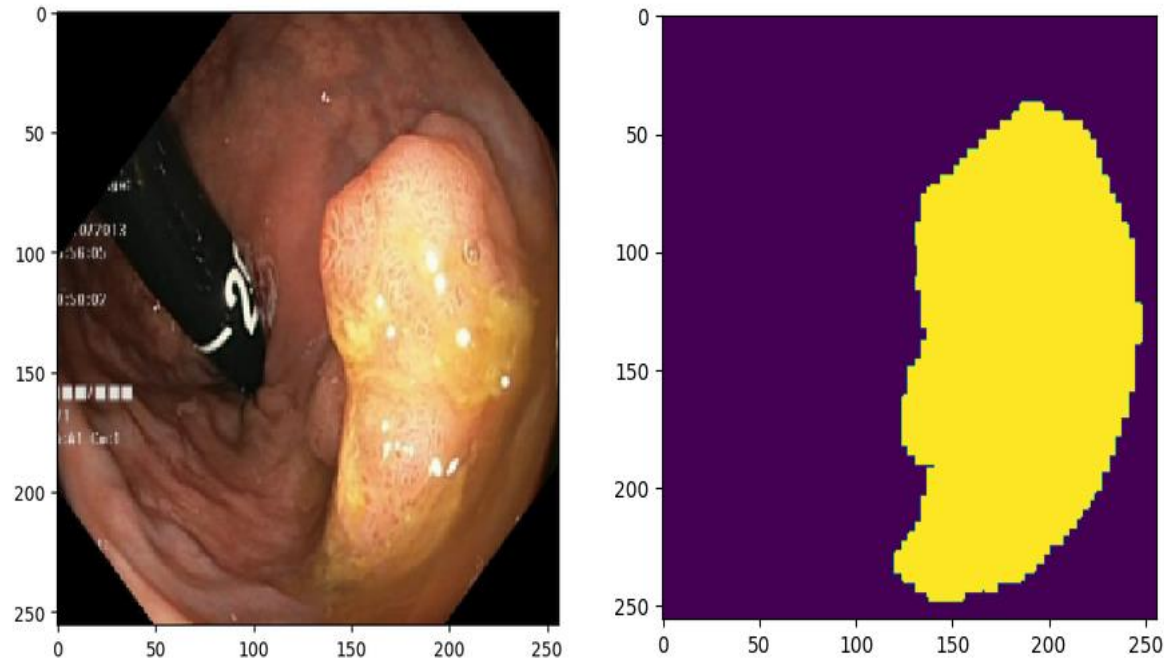


Figure 3.1: Representative gastrointestinal image and corresponding ground truth mask from the experimental dataset.

Training Performance of the Individual Models

The values of the last epoch (epoch 30) from the training procedure used (30 epochs) shown in table 4.2. The validation accuracy and Dice coefficient, IoU for the individual models were different: U-Net had the highest validation accuracy, SegNet had the highest validation Dice coefficient and IoU. For DeepLab, the scores obtained with the implemented configuration were lower with the method based on overlap.

Table 3.2: Final-Epoch Training and Validation Performance

Model	Training Accuracy	Validation Accuracy	Validation Loss	Validation Dice	Validation IoU
SegNet	96.89%	92.63%	0.2167	0.7210	0.5654
U-Net	95.74%	92.96%	0.2095	0.6951	0.5340
DeepLab	90.73%	90.43%	0.2394	0.5337	0.3661

SegNet Performance

The accuracy of the model after training and validation for SegNet were 96.89% and 92.63% respectively. The final validation loss was 0.2167 and the Dice coefficient and IoU were 0.7210 and 0.5654 respectively. The values reflect the better performance of SegNet in terms of overlap with the base architectures in the last epoch recorded.

U-Net Performance

The accuracy that U-Net achieved at the end of training was 95.74% and the highest individual-models accuracy for the validation was 92.96%. Its validation loss was 0.2095, with a Dice coefficient of 0.6951 and IoU of 0.5340. The results demonstrate a high level of classification accuracy pixel-wise, and the advantage of the skip connections added in the U-Net to preserve the spatial information in lesion-segmentation.

DeepLab Performance

The final accuracy of the training set was 90.73% and for the validation set it was 90.43% with DeepLab. The Dice coefficient and IoU were 0.5337, and 0.3661, respectively. DeepLab is more than 90% accurate for its overall accuracy (overlap metrics are not as high as SegNet and U-Net) on the recorded experiment.

Comparative Performance of Individual and Ensemble Models

The individual training was followed by upsampling of the predictions of SegNet, U-Net and DeepLab to generate features for the Random Forest meta-model. The metric comparison from the notebook, which is completed, is shown in Tab. 4.3. The ensemble clearly outperformed all individual models on all the evaluation measures reported.

Table 3.3: Comparative Performance of SegNet, U-Net, DeepLab and Ensemble RFC

Model	Accuracy	F1-score	ROC-AUC	Precision	Recall
SegNet	92.63%	86.49%	87.96%	73.90%	81.19%
U-Net	92.96%	84.88%	81.11%	87.33%	63.92%
DeepLab	90.43%	77.44%	72.65%	84.63%	46.85%
Ensemble RFC	99.95%	99.91%	99.85%	99.99%	99.70%

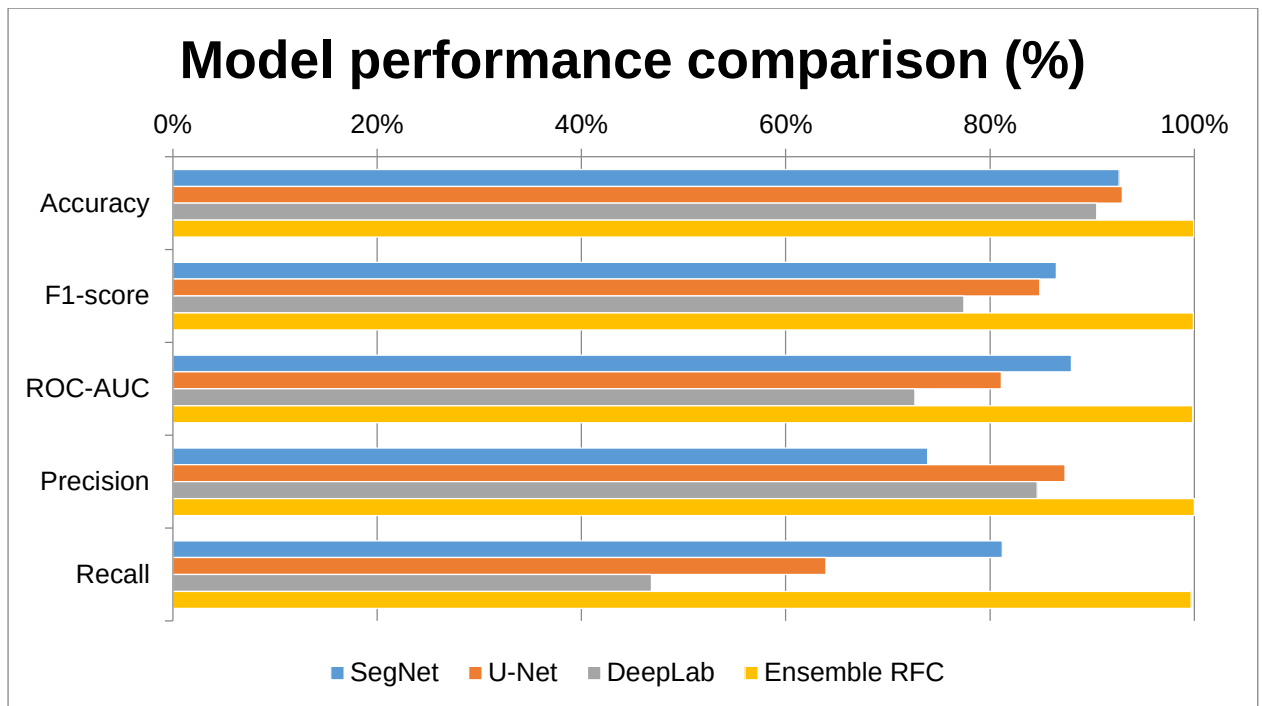


Figure 3.2: Comparative performance of the three individual segmentation models and the Ensemble RFC.

The overall accuracy of the three models was highest for the U-Net model with approximately 92.96% which was close to the SegNet model with 92.63%. SegNet achieves higher scores than the other models in the individual-model F1 score, ROC-AUC and recall, while U-Net achieves higher scores than the other models in precision. The base models on average had the lowest recall, likely due to the lowest recall of DeepLab which had more of the true area of lesions missed in the reported validation experiment.

Performance of the Proposed Ensemble Model

The accuracy, f1 score, ROC-AUC, precision and recall were 99.95%, 99.91%, 99.85%, 99.99% and 99.70% respectively, on the stacking based Random Forest ensemble model for the validation pixels used in the experiment. The high precision suggests that there are only a few false positive predictions for lesions, and the high recall suggests that most annotated pixels for lesions were correctly predicted.

One of the qualitative outputs of the ensemble is shown in Fig. 4.3. The mask predicted is displayed with the shape and position of the mask coinciding with the ground truth mask for the validation image that is selected as blue, indicating the ensemble gives a good agreement for this sample.

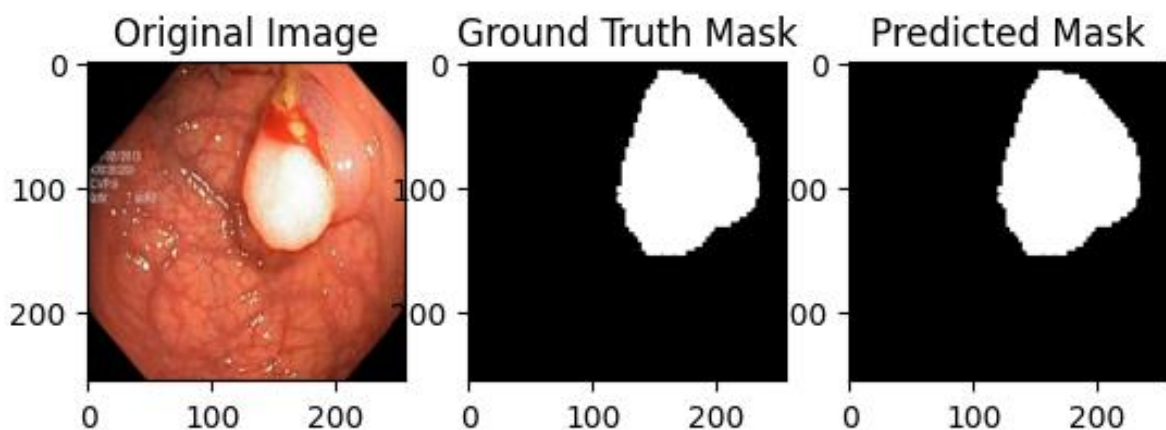


Figure 3.3: Original image, ground truth mask, and Ensemble RFC predicted mask for a representative validation sample.

Image-Level F1-Score Analysis

To validate images, we randomly selected them from the notebook and calculated image level F1-Score for SegNet, U-Net, DeepLab and Ensemble RFC. The mean scores of the results are shown in Table 4.4. This analysis gives another perspective of the performance as the score of each mask was derived separately and then averaged.

Table 3.4: Average F1-Score Across 50 Random Validation Images

Model	Average F1-score
SegNet	77.31%
U-Net	67.79%
DeepLab	55.55%
Ensemble RFC	99.83%

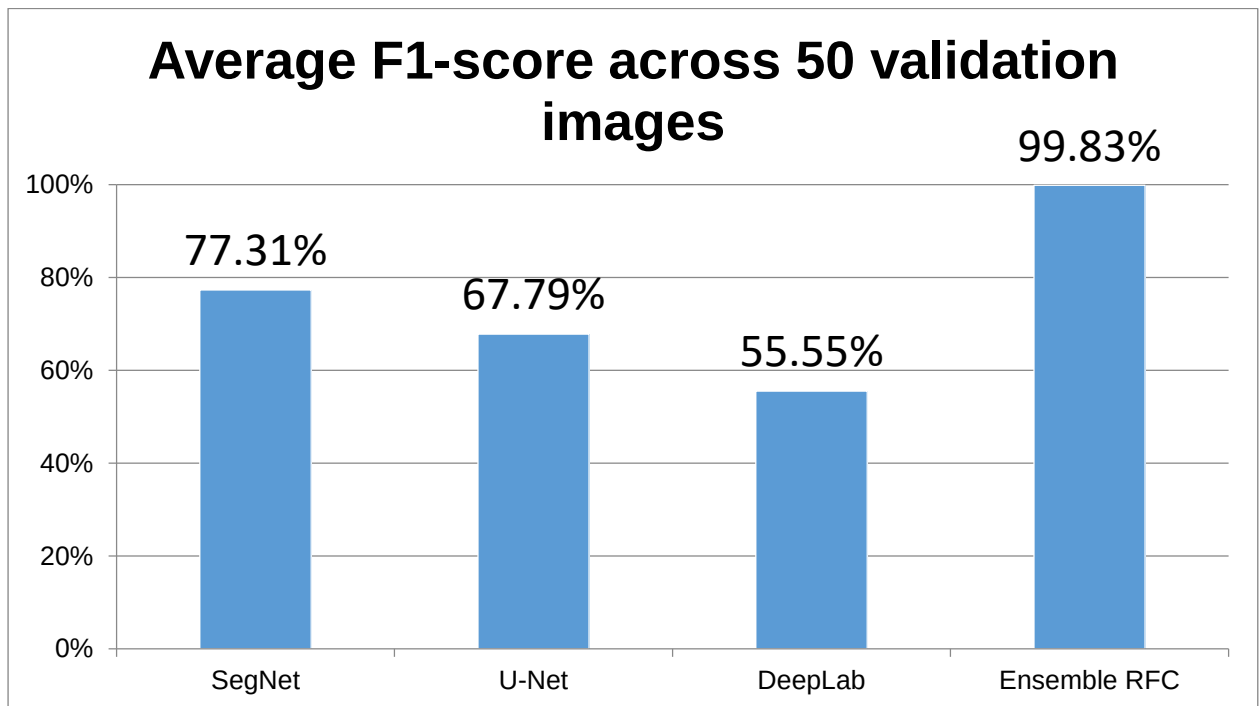


Figure 3.4: Average F1-score across 50 randomly selected validation images.

SegNet's average F1 score on the images level is 77.31%, the best of all the individual networks, followed by U-Net (67.79%) and DeepLab (55.55%). The Ensemble RFC's success rate was 99.83%. This was analysed using 50 images, and a typical 6 panel comparison is shown in Figure 3.5.

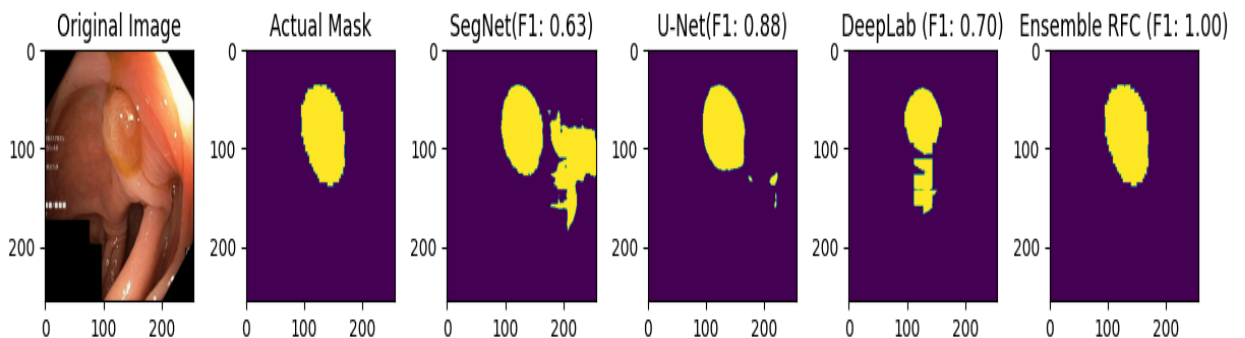


Figure 3.5: Representative comparison of original image, actual mask, SegNet, U-Net, DeepLab, and Ensemble RFC predictions with sample-level F1-scores.

Confusion Matrix Analysis

There were 6,553,600 validated pixels (100 images x 256 x 256 pixels) in the validation subset. The confusion matrix generated by the notebook had 5536275 true negative, 152 false positive, 3027 false negative and 1014146 true positive pixels. The values will give an understanding of why the ensemble is working well in the recorded experiment when it has a high precision and recall.

Table 3.5: Pixel-Level Confusion Matrix Values for Ensemble RFC

Outcome	Number of pixels
True Negative	5,536,275
False Positive	152
False Negative	3,027
True Positive	1,014,146

A lot of diagonal elements are 1 in the confusion matrix. The number of false positives is small (compared to the number of background pixels) and the number of false negatives is also small (compared to the number of lesion pixels). In terms of precision and recall, it is in line with the reported 99.99% and 99.70% respectively.

ROC-AUC Analysis

ROC-AUC were chosen for summarising the discrimination between lesion and background pixels. The individual ROC-AUC of SegNet was 87.96% which is the best result among the three models, followed by U-Net with 81.11% and DeepLab with 72.65%. For the validation data, the score for the Ensemble RFC was 99.85%, which is close to the perfect separation of the two classes of pixels, for the meta-model experiment.

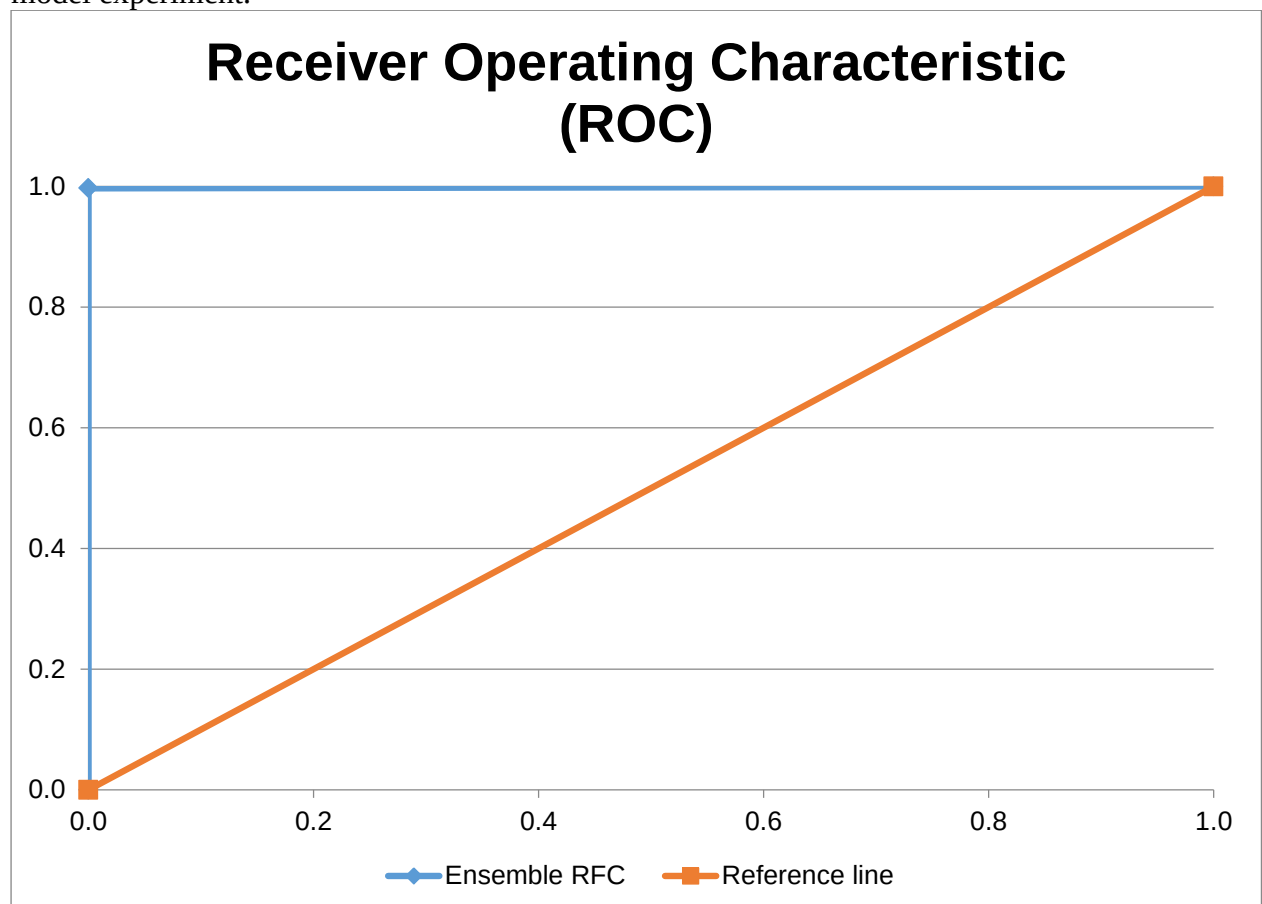


Figure 3.7: Receiver Operating Characteristic curve produced for the Ensemble RFC.

Precision-Recall Analysis

The precision recall analysis is helpful in lesion segmentation as there is not a uniform distribution of the number of lesions and background pixels. The model U-Net got the maximum value of 87.33% for precision and the SegNet model got the maximum value of 81.19% for recall for individual models. The precision of DeepLab was 84.63% and the recall was 46.85%. The highest precision and recall were for the Ensemble RFC with a score of 99.99% and 99.70% respectively.

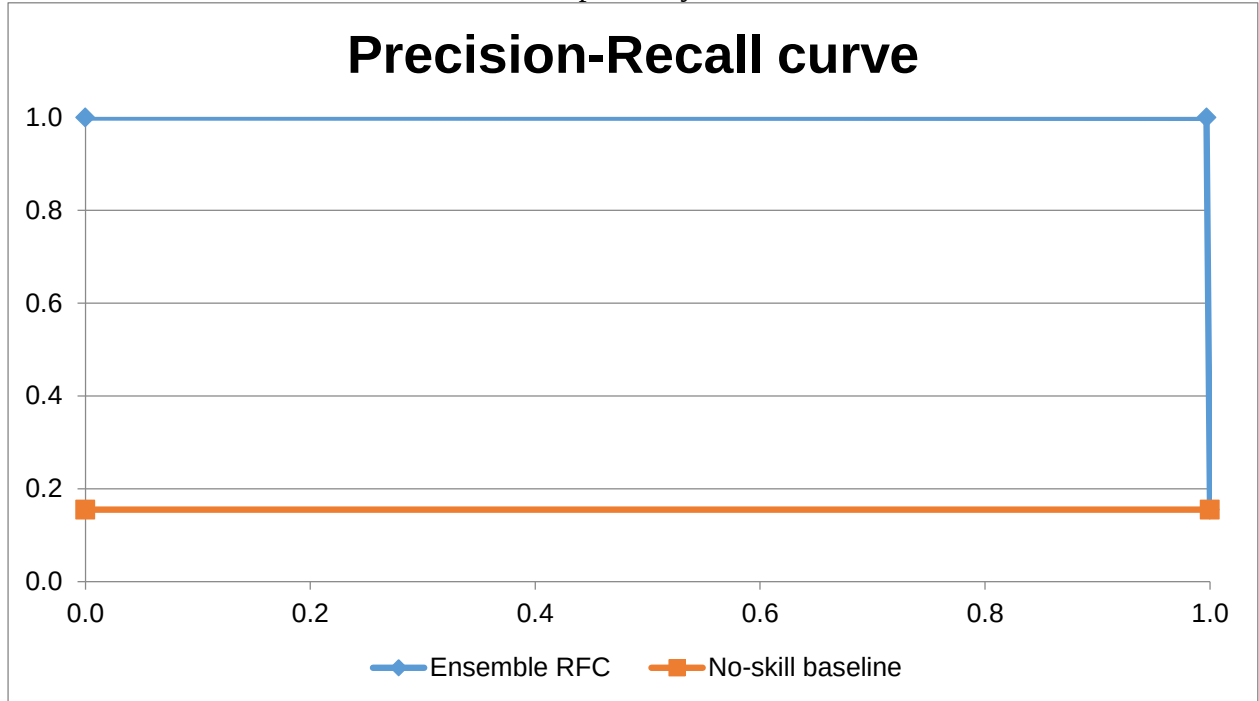


Figure 3.8: Precision-recall curve produced for the Ensemble RFC.

Qualitative Analysis of Segmentation Outputs

The various error types for the three base networks are also evident with the qualitative comparisons in the notebook. In general, the area of the lesion was bigger for SegNet as it had better recall. Overall, U-Net performed well on the accuracy of the lesion regions obtained, even if it was relatively small in size, and it matched the lesion areas on most of the samples, though in some samples, it missed the area of the lesion pixels, having a relatively high precision. DeepLab has not achieved the highest recall and image level F1 score, and has a lesser lesion coverage in a few selected samples.

The ensemble predictions were closer in the provided samples of the validation set (represented in the figure) to the corresponding annotations. The three base-models outputs were used as the inputs to the meta-classifier which was capable of making use of the agreement and disagreement of the networks to form the final binary mask.

Discussion

The present study investigated the performance of three deep learning segmentation architectures (SegNet, U-Net, and DeepLab) and evaluated whether their integration through a stacking-based Random Forest Classifier (RFC) could improve gastrointestinal lesion segmentation. Using the Kvasir-SEG dataset, the individual models showed different strengths, while the proposed ensemble framework achieved superior segmentation performance.

Among individual models, U-Net achieved the highest validation accuracy (92.96%), whereas SegNet demonstrated better lesion overlap performance with the highest Dice coefficient (0.7210) and IoU (0.5654). These differences indicate that accuracy alone is insufficient for evaluating medical image segmentation because background pixels

can dominate performance measures. Therefore, overlap-based metrics such as Dice and IoU provide more meaningful evaluation of lesion localization (Müller et al., 2022).

The strong performance of U-Net is consistent with its widespread application in biomedical image segmentation. The use of encoder–decoder architecture with skip connections enables effective preservation of spatial information and improved boundary localization (Ronneberger et al., 2015). Previous studies have also reported competitive performance of U-Net-based models on the Kvasir-SEG dataset, although advanced architectures incorporating attention mechanisms and transformer-based feature fusion have achieved higher segmentation accuracy (Das et al., 2024).

SegNet achieved superior Dice and IoU values among the individual models, which may be attributed to its ability to preserve spatial details through pooling indices during decoding (Badrinarayanan et al., 2017). The higher recall of SegNet suggests improved lesion coverage; however, this advantage may introduce additional false-positive regions. Similar observations have been reported in previous ensemble-based polyp segmentation studies, where combining heterogeneous segmentation networks improved lesion delineation compared with individual models (Guo et al., 2019).

Although DeepLab provides strong multi-scale feature extraction through atrous convolution, its performance was lower under the current experimental configuration, particularly in recall and overlap-based metrics. This suggests that DeepLab performance is highly dependent on backbone selection, feature extraction strategy, and optimization parameters (Chen et al., 2018). Recent improved DeepLab-based models incorporating attention and multi-scale feature fusion have demonstrated better performance on gastrointestinal polyp segmentation (Quindós et al., 2023).

The most important result of this study was the significant improvement attained using stack based RFC ensemble. The ensemble model achieved 99.95% accuracy, 99.91% F1-score, 99.85% ROC-AUC, 99.99% precision and 99.70% recall, outperforming all individual segmentation networks. This improvement shows that complementary lesion features are learned by different architectures and that a meta-classifier can combine the predictions of these classifiers to minimize the segmentation errors.

The results corroborate the previous studies that demonstrated that heterogeneous ensemble learning leads to better gastrointestinal polyp segmentation by integrating complementary predictions from multiple networks (Guo et al., 2019). The advantage of the Random Forest meta-classifier over simple averaging is that it is able to learn nonlinear relationships between the models and selectively takes advantage of the strengths of each individual architecture.

The reported ensemble performance should also be further validated with an independent test data set, and external gastrointestinal image databases, to verify the generalizability. The cross-dataset evaluation and clinical validation are worth future studies and should be deployed in real-world computer-aided diagnosis systems.

Conclusion

The aim of this study was to identify and isolate gastrointestinal lesions in wireless capsule endoscopy (WCE) images by developing a deep ensemble learning framework, combining SegNet, U-Net and DeepLab with a stacking-based Random Forest Classifier. It is observed that the individual models showed complementary capabilities; the highest validation accuracy is obtained from U-Net while SegNet gives better lesion overlap performance based on Dice coefficient and IoU. The proposed ensemble framework significantly boosted the diagnostic performance; with the highest accuracy of 99.95%, F1 score of 99.91%, ROC-AUC of 99.85%, precision of 99.99%, and recall of 99.70%. The results prove that a heterogeneous ensemble of deep learning models can successfully mitigate the weaknesses of individual deep learning models, by leveraging complementary spatial and contextual features for

lesion localization with greater accuracy. While the proposed framework showed good performance, further study is needed to validate the model externally with independent gastrointestinal endoscopy datasets, to assess the generalizability of the system. Other potential enhancements involve incorporating transformer-based structures, attention mechanisms, explainable AI techniques, and lightweight models suitable for real-time clinical applications. Further extension of the framework to multi-class gastrointestinal disease detection, and prospective clinical validation will also be significant steps towards the practical implementation in computer-aided diagnostic systems.

References

- Afzal, M. A., Noor, M., Ukkasha, S. M., Akram, H. I., Farooq, M., Anees, N., Nazakat, J., Zahra, H., Ali, H., & Batool, H. (2026). Nutritional interventions in the treatment of upper and lower gastrointestinal disorders. In A. Khan, K. Saeed, & M. Rehman (Eds.), *Natural Products in Modern Medicine: Nutraceuticals, Phytochemicals and Herbal Therapeutics* (pp. 295–304). Unique Scientific Publishers.
- Aoki, T., Yamada, A., Aoyama, K., Saito, H., Tsuboi, A., Nakada, A., Niikura, R., Fujishiro, M., Oka, S., Ishihara, S., Matsuda, T., Tanaka, S., Koike, K., & Tada, T. (2019). Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network. *Gastrointestinal Endoscopy*, 89(2), 357–363.e2.
- Aoki, T., Yamada, A., Kato, Y., Saito, H., Tsuboi, A., Nakada, A., Niikura, R., Fujishiro, M., Oka, S., Ishihara, S., Matsuda, T., Nakahori, M., Tanaka, S., Koike, K., & Tada, T. (2020). Automatic detection of blood content in capsule endoscopy images based on a deep convolutional neural network. *Journal of Gastroenterology and Hepatology*, 35(7), 1196–1200.
- Ashraf, M. A., Afzal, M. A., Ijaz, M., Arshad, A., Zafar, M. N., Ukkasha, S. M., & Zahid, I. (2025). Financial impact of nutrition on non-communicable diseases in hospital or other healthcare settings. *Biological Times*, 4(9), 20–21.
- Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2481–2495.
- Caroppo, A., Siciliano, P., & Leone, A. (2023). An expert system for lesion detection in wireless capsule endoscopy using transfer learning. *Procedia Computer Science*, 219, 1136–1144.
- Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European Conference on Computer Vision*, 801–818.
- Das, A., Singh, A., Nishant, & Prakash, S. (2024). CapsuleNet: A deep learning model to classify GI diseases using EfficientNet-B7. *arXiv*.
- Ding, Z., Shi, H., Zhang, H., Meng, L., Fan, M., Han, C., Zhang, X., & Li, Y. (2019). Gastroenterologist-level identification of small-bowel diseases and normal variants by capsule endoscopy using a deep-learning model. *Gastroenterology*, 157(4), 1044–1054.
- Fatima, T., Sajjad, I., Qamar, M., Afzal, M. A., Noor, N., Tariq, N., Yawar, A., Abbasi, A. J., Rahim, F., Farooq, M. U., Noor, M., & Zahid, I. (2026). The expanding role of artificial intelligence in health, life, and applied sciences: A multidisciplinary review. *Social Science Review Archives*, 4(2), 237–257.
- Guo, Y., Bernal, J., & Liao, X. (2019). A multi-scale and multi-level ensemble network for gastrointestinal polyp segmentation. *IEEE Access*, 7, 152615–152625.
- Jain, S., Seal, A., & Ojha, A. (2022). A hybrid convolutional neural network with meta-feature learning for abnormality detection in wireless capsule endoscopy

- images. *arXiv*.
- Khan, M. A., Shafiq, U., Hamza, A., Mirza, A. M., Baili, J., AlHammadi, D. A., Cho, H. C., & Chang, B. (2025). A novel network-level fused deep learning architecture with a shallow neural network classifier for gastrointestinal cancer classification from wireless capsule endoscopy images. *BMC Medical Informatics and Decision Making*, 25, Article 150.
- Lee, H. S., Kim, J., & Park, M. (2023). DeepLabv3+ framework for multi-scale segmentation of gastrointestinal lesions using endoscopic images. *IEEE Access*, 11, 87654–87667.
- Malik, H., Naeem, A., Sadeghi-Niaraki, A., & others. (2024). Multi-classification deep learning models for detection of ulcerative colitis, polyps, and dyed-lifted polyps using wireless capsule endoscopy images. *Complex & Intelligent Systems*, 10, 2477–2497.
- Mughal, F. A., Jahangir, A., Dawood, M., Khalid, A., Saadat, S., Ali, M. A. W., Khan, S. S., Afzal, M. A., & Shahzadi, N. (2025). The Efficacy of Biologics in Inflammatory Bowel Disease (IBD): A Review. *Journal of Medical & Health Sciences Review*, 2(3).
- Müller, D., Soto-Rey, I., & Kramer, F. (2022). Towards the systematic evaluation of deep learning models in medical image segmentation. *Artificial Intelligence in Medicine*, 127, 102285.
- Muruganantham, P., & Balakrishnan, S. M. (2022). Attention aware deep learning model for wireless capsule endoscopy lesion classification and localization. *Journal of Medical and Biological Engineering*, 42, 157–168.
- Nguyen, D. T., Lee, M. B., Pham, T. D., Batchuluun, G., Arsalan, M., & Park, K. R. (2020). Enhanced image-based endoscopic pathological site classification using an ensemble of deep learning models. *Sensors*, 20(21), Article 5982.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241.
- Saraiva, M. M., Ribeiro, T., Afonso, J., et al. (2022). Artificial intelligence and capsule endoscopy: Automatic detection of small bowel blood content using a convolutional neural network. *GE Portuguese Journal of Gastroenterology*, 29(5), 331–338.
- Sivari, E., Bostanci, E., Guzel, M. S., Acici, K., Asuroglu, T., & Ayyildiz, T. E. (2023). A new approach for gastrointestinal tract findings detection and classification: Deep learning-based hybrid stacking ensemble models. *Diagnostics*, 13(4), Article 720.
- Smedsrud, P. H., Thambawita, V., Hicks, S. A., Gjestang, H., Nedrejord, O. O., Næss, E., Borgli, H., Jha, D., Berstad, T. J. D., Eskeland, S. L., Lux, M., Espeland, H., Petlund, A., Nguyen, D. T., Garcia-Ceja, E., Johansen, D., Schmidt, P. T., Toth, E., Hammer, H. L., ... Halvorsen, P. (2021). Kvasir-Capsule, a video capsule endoscopy dataset. *Scientific Data*, 8, Article 142.
- Soffer, S., Klang, E., Shimon, O., Nachmias, N., Eliakim, R., Ben-Horin, S., Kopylov, U., & Barash, Y. (2020). Deep learning for wireless capsule endoscopy: A systematic review and meta-analysis. *Gastrointestinal Endoscopy*, 92(4), 831–839.e8.
- Yogapriya, J., Chandran, V., Sumithra, M. G., & others. (2021). Gastrointestinal tract disease classification from wireless endoscopy images using pretrained deep learning model. *Computational and Mathematical Methods in Medicine*, 2021, Article 3862030.
- Zahid, I., Afzal, M. A., Tariq, N., Hassan, M., Ijaz, M., & Ashraf, M. A. (2025). Application of artificial intelligence and marketing in animal nutrition and feed. *Biological Times*, 4(8), 62–63.