Deep Learning-Based Hybrid Loss U-Net for Automated Segmentation of Common Bile Duct Stones in ERCP Fluoroscopy

Taha Pishro Dabaghiyan (MSc)[1], Mona Mohammad-Asghari (MSc)[1], Ramin Niknam (MD)[2], Hossein Parsaei (PhD)[3]*, Mohammad Mehdi Movahedi (PhD)[3], Tahereh Mahmoudi (PhD)[3], Kamran Bagheri Lankarani (MD)[2], Seyed Ali Malek-Hosseini (MD)[4], Seyed Alireza Taghavi (MD)[5], Fardad Ejtehadi (MD)[5], Ebrahim Fallahzadeh Abarghooei (MD)[5], Gholam Reza Sivandzadeh (MD)[5]

[1]Student Research Committee, Shiraz University of Medical Sciences, Shiraz, Iran
[2]Health Policy Research Center, Shiraz University of Medical Sciences, Shiraz, Iran
[3]Department of Medical Physics and Engineering, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran
[4]Shiraz Transplant Center, Abu Ali Sina Hospital, Shiraz University of Medical Sciences, Shiraz, Iran
[5]Gastroenterohepatology Research Center, Shiraz University of Medical Sciences, Shiraz, Iran

ABSTRACT

Background: Accurate segmentation of common bile duct (CBD) stones during endoscopic retrograde cholangiopancreatography (ERCP) is essential to reduce procedural complications and ensure complete stone removal. However, the high deformability of the CBD and the small size of stones make accurate identification challenging in fluoroscopic images.

Objective: To develop and validate a deep learning based model capable of segmenting CBD stones, the CBD, and the duodenoscope in ERCP fluoroscopic images.

Material and Methods: This retrospective study utilized 1,668 ERCP cholangiograms collected from a single tertiary center. A U-Net-based convolutional neural network was trained using various individual and hybrid loss functions. Model performance was evaluated using Intersection over Union (IoU), precision, and recall.

Results: The model trained with a hybrid loss function combining Dice Loss and Categorical Focal Loss achieved IoU scores of 96.93% for the duodenoscope, 89.76% for the CBD, and a mean IoU of 80.61% for CBD stones. These results reflect a 1.19% improvement in CBD segmentation and a 7.93% improvement in stone segmentation compared to existing approaches.

Conclusion: The proposed deep learning model significantly enhances segmentation accuracy in ERCP imaging and shows strong potential for supporting real-time clinical decision-making. Its integration into ERCP workflows could improve procedural safety, efficiency, and patient outcomes.

Keywords

Deep Learning; Image Segmentation; Common Bile Duct Stones; Endoscopic Retrograde Cholangiopancreatography; Computer-Assisted Diagnosis; Fluoroscopy; U-Net Architecture; Artificial Intelligence

*Corresponding author:
Hossein Parsaei
Department of Medical Physics and Engineering, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran
E-mail:
hparsaee@gmail.com

Received: 23 August 2025
Accepted: 31 January 2026

Introduction

Choledocholithiasis, the presence of stones in the common bile duct (CBD), is a prevalent cause of biliary obstruction, affecting approximately 1–15% of patients with gallstones [1]. This

condition complicates the management of cholelithiasis by increasing patient morbidity and imposing substantial financial burdens on healthcare systems due to recurrent interventions and extended hospital stays. Endoscopic retrograde cholangiopancreatography (ERCP), which combines fluoroscopic imaging with upper gastrointestinal duodenoscopy, is a widely used minimally invasive procedure that offers an effective alternative to surgery for the diagnosis and treatment of CBD stones [2-4].

The success of ERCP is critically dependent on the complete extraction of bile duct stones, as residual or recurrent stones can lead to serious complications. Despite endoscopic sphincterotomy (EST), recurrence rates remain significant, ranging from 4% to 25% [5]. Multiple factors influence the technical complexity and clinical outcomes of ERCP, including anatomical and morphological features of the stones. Notably, stones larger than 1 cm, a distal CBD diameter under 15 mm, and an angulation of the distal CBD below 135° have been identified as predictors of procedural difficulty [6-8]. Accurate assessment of these features is essential for planning and executing successful interventions. However, interpreting fluoroscopic or endoscopic images to derive these measurements remains a significant challenge, even for experienced endoscopists.

In this context, advanced image analysis techniques (particularly image segmentation) offer promising avenues for improving procedural planning and outcomes. Image segmentation involves partitioning an image into distinct regions to facilitate interpretation, enabling the precise identification and delineation of anatomical structures and pathological findings. In medical imaging, segmentation is crucial for achieving diagnostic accuracy, planning therapeutic interventions, and providing intraoperative guidance.

Recent advancements in artificial intelligence (AI), particularly deep learning, have demonstrated significant potential in automating the detection of CBD stones [9]. For instance, weakly supervised models have achieved accuracies up to 85.93% [10], and segmentation methods have reached a mean Intersection over Union of 68.35% on fluoroscopic datasets [11]. Despite these promising results, many existing approaches lack the robustness and contextual understanding required for real-time clinical decision-making during ERCP.

To address these limitations, we propose a novel deep learning framework based on the U-Net architecture, enhanced by hybrid loss functions, for the automatic segmentation of the CBD and bile duct stones in ERCP fluoroscopy. Our model generates grayscale masks delineating the duodenoscope, bile duct, and stones, enabling accurate localization and measurement. This facilitates comprehensive feature extraction, supports the development of a procedural difficulty scoring system, and enhances the prediction of post-ERCP complications. Furthermore, duodenoscope segmentation enables precise spatial calibration, which is essential for translating image-based measurements into real-world dimensions.

The remainder of this paper is organized as follows: the "Materials and Methods" section describes our data and modeling approach; the "Results" section presents our findings and performance evaluation; the "Discussion" section explores the clinical implications; and the final "Conclusion" section summarizes our key contributions.

Material and Methods

A. Study Design and Participants

This retrospective study was conducted at Namazi Hospital, Shiraz, Iran. A total of 1,668 cholangiogram images from 776 patients who underwent ERCP procedures between December 2013 and November 2024 were collected for model training and evaluation. The procedures were performed by four interventional endoscopists, each with over 10 years of

experience and more than 1,000 ERCP cases, as well as one additional interventional endoscopist with over three years of experience and more than 500 ERCP procedures.

The study was approved by the Ethics Committee of Shiraz University of Medical Sciences. Inclusion criteria encompassed patients undergoing ERCP for suspected or confirmed CBD stones evident on cholangiography. Exclusion criteria included diagnoses unrelated to CBD stones, such as pancreatic cancer, cholangiocarcinoma, ampullary tumors, chronic liver diseases affecting the biliary system (e.g., primary biliary cirrhosis, primary sclerosing cholangitis), prior surgeries or trauma, and other biliary disorders. Additionally, ERCP cases solely involving stent placement, replacement, or removal were excluded, as these did not involve stone detection. Cases with incomplete cholangiography, due to failed cannulation, intubation issues, or gastric retention, were also excluded to ensure data quality.

Fluoroscopic images were acquired using a duodenoscope (TJF-Q180V; Olympus Medical Systems, Japan) and a flat-panel detector (C-arm 8000; Ziehm Imaging, Germany). Image annotation was performed by two expert endoscopists and AI engineers using the labelme annotation tool (https://github.com/wkentaro/labelme). The annotated structures included the CBD, bile duct stones, and the duodenoscope. These annotations were used as the gold standard for model training and evaluation. Representative examples of annotated images, spanning a range of image complexity, are shown in Figure 1.

B. Design of Segmentation Model

This section details the development of an AI-based segmentation model for automatic identification of CBD, CBD stone, and the duodenoscope in fluoroscopic images. The model development process involved three primary steps: image pre-processing, model development, and model evaluation.

1. Image Pre-Processing

The image pre-processing pipeline consisted of noise reduction, contrast enhancement, sharpening, and standardization. A 3-pixel radius median filter was applied to eliminate text overlays and outliers. Contrast

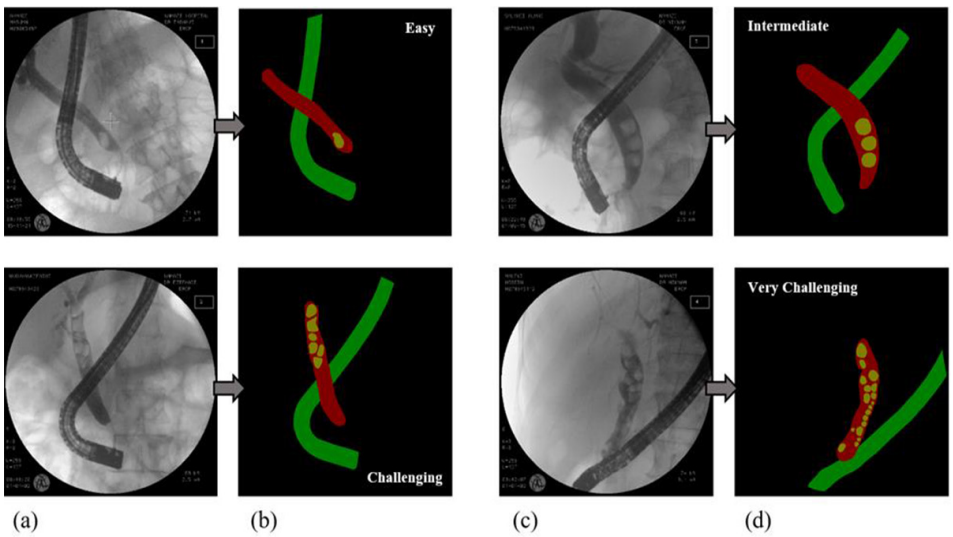

Figure 1: Examples of annotated fluoroscopic images spanning low- to high-complexity cases. Columns (**a**) and (**c**) show the original images, while columns (**b**) and (**d**) display the corresponding ground truth annotations, with the common bile duct (CBD) shown in red, bile duct stones in yellow, and the duodenoscope in green.

enhancement was performed through normalization (with a pixel saturation rate of 0.3) and histogram equalization, which improved visual contrast. For sharpening, a first-order Laplacian filter was used to enhance edge details. Finally, images were resized and standardized for training input. The original images had a resolution of 576×576 pixels (20.32×20.32 cm, 8-bit grayscale). For computational efficiency, images were resized to 256×256 pixels while maintaining the grayscale format.

2. Model Architecture

The segmentation model was based on the U-Net architecture [12], widely recognized for its performance in medical image segmentation. Models were trained using a batch size of 16 for 40 epochs with the Adam optimizer and a fixed learning rate of 0.0001. Hyperparameters were optimized through grid search. Figure 2 shows the architecture of the adopted U-Net model.

3. Loss Function Design

A key step in designing an AI-based model is the choice of the loss function. In this work, to mitigate the effect of class imbalance [13] and improve the detection of small anatomical structures such as CBD stones, we evaluated the following five individual loss functions and explored several pairwise hybrid combinations.

Dice Loss (DL) is a loss function commonly used in image segmentation tasks to evaluate the similarity between predicted and ground truth segmentations [14], that defined as:

$$DL = 1 - \frac{2\times|U \cap V|}{|U|+|V|} \qquad (1)$$

where V represents the predicted binary mask, and U is the ground truth.

Jaccard Loss (JL) is derived from the Intersection over Union (IoU) metric, commonly referred to as the Jaccard index [15]. It measures the ratio of the overlap between the predicted segmentation mask and the ground truth mask to the total area covered by their combined regions [16]:

$$JL = \frac{|U \cap V|}{|U \cup V|} \qquad (2)$$

where V and U are defined as mentioned before (Eq.1).

Categorical Focal Loss (CFL) is a modified version of cross-entropy loss designed to address class imbalance in multi-class classification tasks [17].

$$CFL(p) = -\alpha\,(1-p)^{\gamma} log(p) \qquad (3)$$

where p represents the predicted probability

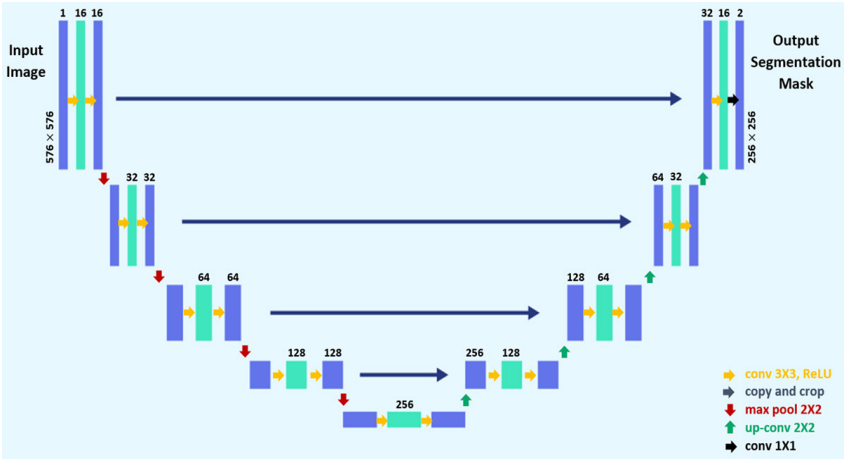

Figure 2: Schematic illustration of the adopted U-Net architecture used for segmentation. The network consists of an encoder path for hierarchical feature extraction and a decoder path for spatial resolution recovery, with skip connections that fuse low-level and high-level feature representations to improve localization accuracy.

for the target class, and α is the class weight factor (set to 2 in this study), and the parameter of γ is the amount of focusing of the loss on hard examples.

Categorical Cross-Entropy Loss (CCEL) quantifies the disparity between two probability distributions for a specific random variable [16]. In segmentation tasks, cross-entropy loss is employed to assess the accuracy of the model's predictions against the true labels [18]. The model applies the softmax function to produce pixel-wise probability maps, indicating the chance of each pixel belonging to different classes [16]. The loss is computed by taking the negative logarithm of the predicted probability corresponding to the correct class for each pixel:

$$CCFL(p) = -\alpha \gamma log(p) \quad (4)$$

where p, α and γ are defined as in Eq.3.

Weighted Pixel-wise Categorical Cross-Entropy Loss (WPCCEL) in cases of imbalanced datasets can be useful, when a common strategy for cross-entropy loss is to apply different weights to each class [16]. This helps to equalize the impact of each class on the total loss and enhances the models performance on classes that are underrepresented. One method to determine these weights is based on the inverse of class frequency, meaning that classes with fewer samples receive higher weights, while those with more samples are assigned lower weights [16].

In this study, we measured the number of pixels corresponding to the duodenoscope, CBD, stone, and background in 100 labeled images and calculated their average. Let $\overline{P}_{Backgroun}$, $\overline{P}_{Duodenoscope}$, $\overline{P}_{CBD}$, and $\overline{P}_{Stone}$ denotes the average number of pixels in the background, duodenoscope, CBD, stone, the weight for each class was calculated as:

$$W_i = \frac{\overline{P}_{Backgroun}}{\overline{P}_i} \quad (5)$$

where $i \in \{background, duodenoscope, CBD, stone\}$

Thus, the weight for the background class will be one, and for other classes, the lower the average value of their calculated number of pixels, such as a stone, which always occupies a very small space in the image; the greater the weight value will be, so that the focus of error improvement is more on that class. Ultimately, the WPCCFL value is defined as:

$$WPCCFL(p) = -\alpha \gamma p log(p) W_i \quad (6)$$

where p, α and γ are as previously described (Eq. 3), and W_i denotes the weight assigned to class i, calculated using Eq. 5.

Conventional loss functions, such as Cross-Entropy [18] and Dice loss [14], often struggle to effectively address class imbalance [13]. To overcome challenges associated with small foreground regions and significant intra-class variation in region sizes, we combined specialized pixel- and region-based loss functions: Dice, Jaccard [15], Cross-Entropy, Categorical Focal [17], and Weighted Pixel-wise Cross-Entropy loss. This approach could enhance segmentation accuracy by accounting for variations in object size and distribution.

C. Model Evaluation

The segmentation algorithms were evaluated using recall, precision, and intersection over union (IoU), as in [11]. The metrics are defined as:

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$IoU = \frac{TP}{TP + FP + FN} \quad (9)$$

Where true positives (TP) were defined as correctly segmented regions overlapping with expert annotations, false positives (FP) were predictions not found in the ground truth, and false negatives (FN) were expert-annotated regions not identified by the model. Here, precision reflects the portion of predicted regions that are relevant, and recall indicates how effectively the model detects all true objects.

To ensure unbiased performance evaluation and simulate real-world applications, the

dataset was partitioned into temporally distinct training and testing subsets. This strategy helps prevent data leakage and mimics a prospective validation setup. The training set consisted of 1,500 images from 686 patients, collected between December 2, 2013, and June 8, 2024. The testing set included 168 images from 90 patients, acquired between June 9 and November 21, 2024, ensuring that the model was evaluated on entirely unseen future data. This chronological separation enhances the robustness and generalizability of the model's performance assessment.

Results

To ensure a realistic and unbiased evaluation, the dataset was temporally separated into distinct training and testing sets, as described in the Material and Methods section. This design simulates real-world prospective validation and reduces the risk of data leakage, thus enhancing the reliability and generalizability of the model's performance assessment.

All models were implemented in TensorFlow and trained on a workstation equipped with an NVIDIA GeForce RTX 3050 GPU (16 GB memory). Each model was trained using a unique loss function or a combination of loss functions, as outlined in the Methods section. Performance was evaluated on the independent test set using three key segmentation metrics: IoU, precision, and recall. The metrics, recall, precision, and IoU, were computed in the range [0,1] and are reported as percentages for clarity.

Quantitative segmentation results for all developed models are summarized in Table 1, including performance for the duodenoscope, CBD, and CBD stones. Among the examined configurations, the model trained with a hybrid loss combining DL and CFL provided the best overall performance. Specifically, the DL+CFL model outperformed all other configurations, achieving the highest IoU for all three structures, 96.93% for the duodenoscope, 89.76% for the CBD, and 80.61% for the CBD stones.

Quantitative comparison with previously published methods is presented in Table 2. Huang et al. [11] reported mean IoU values of 86.42% for CBD segmentation and 68.35% for stone segmentation using a cascaded CNN approach. Another study [19], incorporating data augmentation strategies, achieved mean

Table 1: Quantitative segmentation performance (mean ± SD) of U-Net models trained with different loss functions for the duodenoscope, common bile duct (CBD), and CBD stones. Performance is reported in terms of intersection over union (IoU, %), precision (%), and recall (%). Bold values indicate the best performance for each metric.

Loss Function	Duodenoscope			CBD			Stone		
	IoU (%)	Precision (%)	Recall (%)	IoU (%)	Precision (%)	Recall (%)	mIoU (%)	Precision (%)	Recall (%)
DL	95.70±02.62	98.13±01.89	97.47±01.71	82.26±12.35	89.68±09.29	90.23±08.76	67.07±13.75	81.40±13.01	76.25±13.20
JL	95.81±02.71	97.45±02.25	98.26±01.42	82.39±12.54	89.43±10.28	90.71±08.29	67.18±13.93	80.80±12.72	77.27±13.76
CFL	95.32±03.18	96.90±02.46	98.31±01.83	79.44±13.47	87.23±11.25	89.30±09.05	59.51±13.27	80.48±12.64	68.22±14.35
CCEL	95.69±03.97	97.54±03.57	98.04±01.54	80.96±13.72	88.12±11.46	90.11±08.82	62.89±13.15	77.75±13.09	74.16±12.94
WPCCEL	92.00±04.44	93.29±04.12	98.02±01.85	72.54±12.43	78.01±10.88	90.71±09.12	62.60±11.74	68.52±11.45	77.61±10.12
JL + CCEL	95.79±03.23	97.55±02.23	98.14±02.17	82.14±13.06	89.18±10.42	90.48±08.51	70.26±13.31	79.06±13.17	79.04±12.57
JL + CFL	95.98±02.20	98.25±01.62	97.65±01.46	82.63±12.79	90.23±09.87	90.09±08.77	74.38±12.80	83.07±11.28	79.95±11.83
DL + CCEL	95.79±02.71	97.69±02.21	98.00±01.43	81.70±13.11	90.08±10.31	89.13±09.14	72.77±12.92	81.10±11.72	78.72±12.10
DL + CFL	**96.93±01.36**	**98.87±00.91**	**98.52±00.85**	**89.76±09.22**	**93.33±06.71**	**95.55±05.71**	**80.61±11.70**	**90.38±10.14**	**85.35±10.92**

CBD: Common Bile Duct, IoU: Intersection over Union, DL: Dice Loss, JL: Jaccard Loss, CFL: Categorical Focal Loss, CCEL: Categorical Cross-Entropy Loss, WPCCEL: Weighted Pixel-wise Categorical Cross-Entropy Loss

Table 2: Quantitative comparison of duodenoscope, common bile duct (CBD), and CBD stone segmentation performance in fluoroscopic images with previously published methods. Performance is reported in terms of intersection over union (IoU, %), precision (%), and recall (%). Bold values indicate the best performance for each metric.

	Type of Validation	Duodenoscope			CBD			Stone		
		IoU (%)	Precision (%)	Recall (%)	IoU (%)	Precision (%)	Recall (%)	mIoU (%)	Precision (%)	Recall (%)
Huang et al. 2021 [11]	Internal Validation	95.54	97.87	97.58	83.25	91.55	90.18	76.51	83.62	90.00
	External Validation	96.49	97.81	98.63	85.64	93.51	91.06	60.41	70.58	80.74
	Total Validation	95.85	97.85	97.92	86.42	94.34	91.15	68.35	80.12	82.30
Huang et al. 2023 [19]	Total Validation	**97.52**	97.89	98.15	88.57	91.78	90.14	72.68	83.04	**85.60**
Our proposed Model	Internal Validation	96.93	**98.87**	**98.52**	**89.76**	**93.33**	**95.55**	**80.61**	**90.38**	85.35

CBD: Common Bile Duct, IoU: Intersection over Union

IoU scores of 88.57% for CBD and 72.68% for stones. Using the same evaluation metric, our model achieved higher mean IoU values, exceeding prior methods by 1.19% for CBD segmentation and 7.93% for stone segmentation. A summary of these comparative results is provided in Table 2.

Figure 3 presents an example of qualitative segmentation results across cases with increasing levels of image complexity. In low-complexity images, single-loss and hybrid-loss models showed comparable segmentation performance. As image complexity increased, characterized by a higher stone burden or reduced stone-to-background contrast, differences between the models became more obvious. Under these challenging conditions, single-loss models such as CCEL and WPCCEL frequently failed to detect stones reliably, whereas hybrid-loss models, particularly DL+CFL, maintained more consistent stone delineation and improved boundary representation.

Overall, these results represent significant improvements over models trained with single loss functions or alternative hybrid combinations, particularly in the accurate segmentation of small and low-contrast structures such as CBD stones. The superior performance of the DL+CFL model highlights its effectiveness in addressing both spatial accuracy and class imbalance, validating its potential for integration into clinical ERCP imaging workflows.

Discussion

Accurate segmentation of the CBD, duodenoscope, and bile duct stones in fluoroscopic images is a critical task in ERCP. These anatomical landmarks are essential for guiding diagnostic and therapeutic interventions such as biliary cannulation, sphincterotomy, and stone extraction. Precise delineation aids in quantifying stone burden and improves localization, which is fundamental for reducing procedure time and minimizing complications such as retained stones, perforation, or misidentification of anatomy [20]. Moreover, the duodenoscope serves as a vital spatial reference for estimating sizes and distances in the fluoroscopic field, contributing to accurate device navigation and measurement calibration [21]. However, challenges such as low tissue contrast, overlapping anatomy, and instrument interference make interpretation difficult,

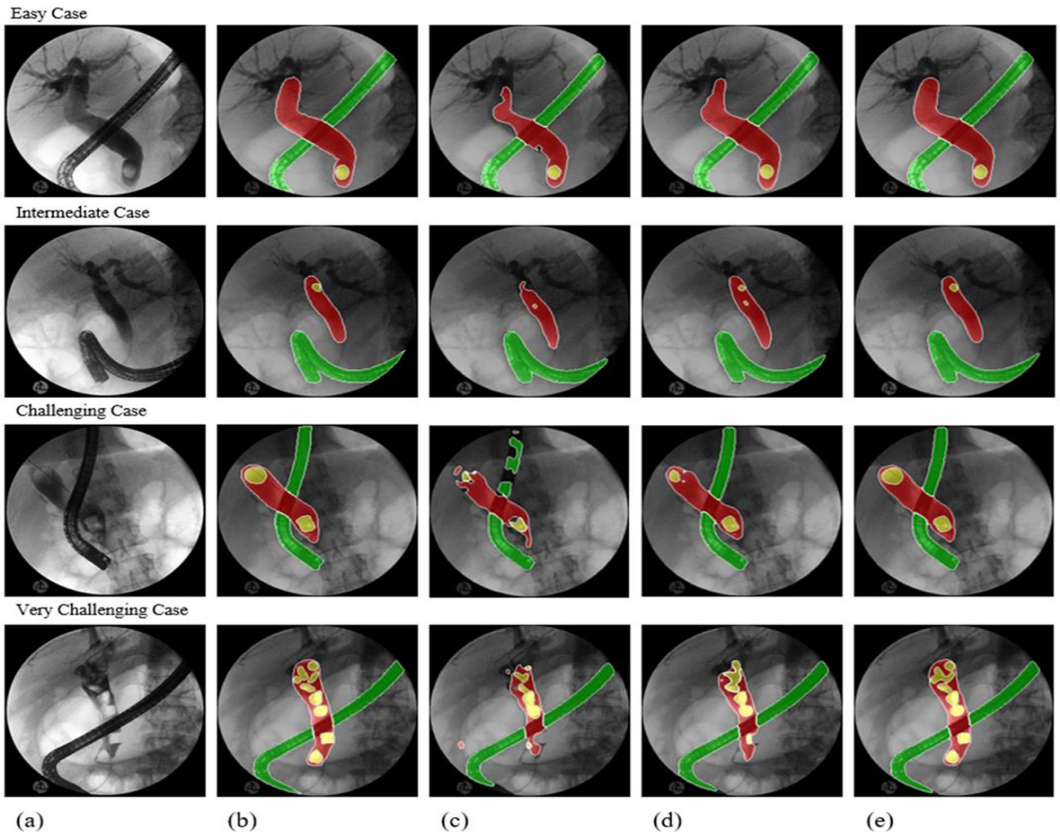

Figure 3: Qualitative segmentation results for the duodenoscope, common bile duct (CBD), and CBD stones across cases ranging from low to high visual complexity. (**a**) Original fluoroscopic image; (**b**) ground truth annotations, where red denotes the CBD, green indicates the duodenoscope, and yellow marks stones; segmentation outputs of the U-Net model trained with (**c**) CCEL, (**d**) JL+CFL, and (**e**) DL+CFL loss functions. (CCEL: Categorical Cross-Entropy Loss, JL: Jaccard Loss, CFL: Categorical Focal Loss, DL: Dice Loss)

even for experienced endoscopists [22,23]. Consequently, the development of robust, automated segmentation tools is crucial for improving procedural safety, efficiency, and outcomes. robust, automated segmentation tools to improve procedural safety, efficiency, and outcomes.

In this study, a U-Net–based deep learning framework trained with a hybrid loss function combining DL and CFL, provided high segmentation accuracy across all target structures, particularly for the complicated task of detecting CBD stones. Quantitative evaluation (Table 1) demonstrated that this configuration outperformed all other loss formulations, with IoU scores of 96.93% for the duodenoscope, 89.76% for the CBD, and mean IoU of 80.61% for CBD stones. These results reflect the benefits of combining regional overlap sensitivity (i.e., DL) with a mechanism that addresses class imbalance and hard-to-detect structures (i.e., CFL).

Several factors contribute to this outstanding performance. Effective image preprocessing techniques, including median filtering, contrast enhancement, and sharpening, obviously improve the quality of input images and the clarity of boundaries. These enhancements are particularly beneficial in differentiating the CBD from the duodenoscope, structures that frequently share similar fluoroscopic textures and indistinct margins. Although CBD

stones typically exhibit higher brightness and contrast, which should theoretically facilitate detection, their segmentation remains particularly challenging due to their small size and variable brightness. These issues persist even after preprocessing, especially for models that do not effectively learn fine-grained details. The DL+CFL model effectively addresses this issue by integrating region-based and pixel-level supervision, enabling the network to capture broad structural context while preserving intricate anatomical features.

A key challenge in the CBD segmentation is the class imbalance resulting from the small relative size and varying shapes of stones. Our hybrid loss function approach addresses this issue: DL emphasizes overall regional accuracy, while CFL enhances sensitivity to subtle and low-contrast features, such as faint stone edges. This synergy improves both recall and segmentation accuracy for small, low-contrast, difficult-to-detect structures. When DL is applied independently, it may produce smooth boundaries that lack precision, potentially missing subtle anatomical contours. In contrast, CFL might amplify details excessively, leading to irregular and anatomically implausible segmentations. By combining the two, our model balances these tendencies, DL reduces excessive irregularity, while CFL preserves critical edge details, leading to anatomically coherent and accurate segmentations across all target structures.

The qualitative assessment of the segmentation results (e.g., Figure 3) further supports these findings. In complex scenarios characterized by increased stone burden or reduced contrast, hybrid-loss models outperformed single-loss formulations. In particular, the DL+CFL configuration consistently preserved structural continuity and boundary reliability in small and low-contrast stone regions, whereas single-loss models frequently failed to detect or fully define these targets. This qualitative behavior aligns with the quantitative performance trends and supports the suitability of the hybrid loss approach for ERCP fluoroscopic imaging.

Comparison with previously published methods (Table 2) indicates that the proposed approach achieves superior segmentation performance, particularly for stone detection. These improvements are likely attributable to the hybrid loss formulation and to the use of rigorous preprocessing and temporally separated training and testing sets, which reduce the risk of data leakage and provide more reliable estimates of generalization performance.

From a clinical perspective, the proposed AI model has direct implications for ERCP-guided stone removal. Accurate segmentation of the CBD and stones facilitates precise determination of their number, size, and location, critical information that can shorten procedure time, minimize the risk of retained stones, and enhance patient outcomes. Additionally, the segmentation of the duodenoscope as a reference object improves measurement accuracy, allowing for more precise calibration during procedures. This capability supports real-time clinical decision-making, potentially increasing both safety and procedural efficiency. Consequently, integrating this model into ERCP workflows has the potential to enhance procedural efficiency and patient outcomes.

Despite these promising results, this study does have limitations. While the dataset is relatively substantial, comprising 1,668 images from 776 patients, it was derived from a single center, which may restrict the generalizability of the findings. Future validation utilizing multi-center datasets and diverse imaging protocols is essential. Moreover, the detection of small stones with low contrast remains a challenge. Advanced feature extraction strategies, such as attention mechanisms or transformer-based architectures, may further enhance model performance in these circumstances. Lastly, while our model significantly improves segmentation accuracy, its real-time deployment in ERCP suites requires further optimization for latency and clinician usability.

Conclusion

This study presents a deep learning-based segmentation model that demonstrates high performance in segmenting the duodenoscope, CBD, and CBD stones in ERCP fluoroscopic images. The use of a U-Net architecture trained with a hybrid Dice and Categorical Focal Loss function enabled the model to effectively capture both broad anatomical structures and subtle, low-contrast features. The model achieved mean IoU scores of 96.93%, 89.76%, and 80.61% for the duodenoscope, CBD, and stones, respectively, surpassing previously reported benchmarks, particularly in the segmentation of stones and the CBD. These results have direct clinical relevance. Improved segmentation accuracy facilitates better localization and quantification of CBD stones, enhances procedural planning, and reduces the risk of complications such as retained stones or ductal injury. The duodenoscope's segmentation further supports precise calibration and navigation during ERCP. While further validation and real-time deployment optimization are needed, this model has strong potential to enhance clinical decision-making, streamline ERCP workflows, and ultimately improve outcomes in the management of biliary disease.

Acknowledgment

The authors used ChatGPT for English language editing and improvement of manuscript clarity, but all revisions were reviewed and approved by the authors.

Authors' Contribution

T. Pishro Dabaghiyan was responsible for methodology, algorithm development and validation, investigation, data collection, and had full access to all study data with ultimate responsibility for data integrity and accuracy of the analysis. He also contributed to writing the original draft and reviewing and editing the manuscript. M. Mohammad-Asghari contributed to the data collection and algorithm development. H. Parsaei and MM. Movahedi contributed to the conceptualization, methodology, supervision, formal analysis, resource management, and writing and editing of the manuscript. T. Mahmoudi contributed to the conceptualization, methodology, advisement, formal analysis, resource management, and writing and editing of the manuscript. R. Niknam, KB Lankarani and SA. Malek-Hosseini contributed to the conceptualization, supervision, conceptualization, data collection, management, and reviewing and editing of the manuscript. SA. Taghavi, F. Ejtehadi, E. Fallahzadeh Abarghooei, and GR. Sivandzadeh contributed to the management, reviewing, and editing of the data collection process. All authors contributed to the revision and finalization of the manuscript.

Ethical Approval

The Research Ethics Committee of the Shiraz University of Medical Sciences approved this study (IR.SUMS.REC.1403.032).

Funding

This study was funded by Shiraz University of Medical Sciences (Grant #28977).

Conflict of Interest

H. Parsaei, as the Editorial Board Member, was not involved in the peer-review and decision-making processes for this manuscript.

References

1. McNicoll CF, Pastorino A, Farooq U, Froehlich MJ, St Hill CR. Choledocholithiasis. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023.
2. Galeazzi M, Mazzola P, Valcarcel B, Bellelli G, Dinelli M, Pasinetti GM, Annoni G. Endoscopic retrograde cholangiopancreatography in the elderly: results of a retrospective study and a geriatricians' point of view. *BMC Gastroenterol.* 2018;**18**(1):38. doi: 10.1186/s12876-018-0764-4. PubMed PMID: 29540171. PubMed PMCID: PMC5853060.
3. Evans N, Buxbaum JL. Endoscopic treatment of ERCP-related duodenal perforation. *Techniques*

in Gastrointestinal Endoscopy. 2019;**21**(2):83-90. doi: 10.1016/j.tgie.2019.04.002.

4. Pavlidis ET, Pavlidis TE. Current management of concomitant cholelithiasis and common bile duct stones. *World J Gastrointest Surg*. 2023;**15**(2):169-76. doi: 10.4240/wjgs.v15.i2.169. PubMed PMID: 36896310. PubMed PMCID: PMC9988640.
5. Nzenza TC, Al-Habbal Y, Guerra GR, Manolas S, Yong T, McQuillan T. Recurrent common bile duct stones as a late complication of endoscopic sphincterotomy. *BMC Gastroenterol.* 2018;**18**(1):39. doi: 10.1186/s12876-018-0765-3. PubMed PMID: 29544453. PubMed PMCID: PMC5856321.
6. Schutz SM, Abbott RM. Grading ERCPs by degree of difficulty: a new concept to produce more meaningful outcome data. *Gastrointest Endosc*. 2000;**51**(5):535-9. doi: 10.1016/s0016-5107(00)70285-9. PubMed PMID: 10805837.
7. McHenry L, Lehman G. Difficult bile duct stones. *Curr Treat Options Gastroenterol.* 2006;**9**(2):123-32. doi: 10.1007/s11938-006-0031-6. PubMed PMID: 16539873.
8. Kim HJ, Choi HS, Park JH, Park DI, Cho YK, Sohn CI, et al. Factors influencing the technical difficulty of endoscopic clearance of bile duct stones. *Gastrointest Endosc.* 2007;**66**(6):1154-60. doi: 10.1016/j.gie.2007.04.033. PubMed PMID: 17945223.
9. Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans Pattern Anal Mach Intell.* 2017;**39**(4):640-51. doi: 10.1109/TPAMI.2016.2572683. PubMed PMID: 27244717.
10. Chang YH, Lin MY, Hsieh MT, Ou MC, Huang CR, Sheu BS. Multiple Field-of-View Based Attention Driven Network for Weakly Supervised Common Bile Duct Stone Detection. *IEEE J Transl Eng Health Med*. 2023;**11**:394-404. doi: 10.1109/JTEHM.2023.3286423. PubMed PMID: 37465459. PubMed PMCID: PMC10351611.
11. Huang L, Lu X, Huang X, Zou X, Wu L, Zhou Z, et al. Intelligent difficulty scoring and assistance system for endoscopic extraction of common bile duct stones based on deep learning: multicenter study. *Endoscopy.* 2021;**53**(5):491-8. doi: 10.1055/a-1244-5698. PubMed PMID: 32838430.
12. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention; Munich, Germany: Springer International Publishing; 2015. p. 234-41.
13. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw*. 2018;**106**:249-59. doi: 10.1016/j.neunet.2018.07.011. PubMed PMID: 30092410.
14. Milletari F, Navab N, Ahmadi SA. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In fourth international conference on 3D vision (3DV); Stanford, CA, USA: IEEE; 2016. p. 565-71.
15. Rahman MA, Wang Y. Optimizing intersection-over-union in deep neural networks for image segmentation. In International symposium on visual computing; Cham: Springer International Publishing; 2016. p. 234-44.
16. Azad R, Heidary M, Yilmaz K, Hüttemann M, Karimijafarbigloo S, Wu Y, et al. Loss functions in the era of semantic segmentation: A survey and outlook [Internet]. arXiv [Preprint]. 2023 [cited 2023 December 28]. Available from: https://arxiv.org/abs/2312.05391.
17. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal Loss for Dense Object Detection. *IEEE Trans Pattern Anal Mach Intell*. 2020;**42**(2):318-27. doi: 10.1109/TPAMI.2018.2858826. PubMed PMID: 30040631.
18. Shannon CE. A mathematical theory of communication. *The Bell System Technical Journal.* 1948;**27**(3):379-423. doi: 10.1002/j.1538-7305.1948.tb01338.x.
19. Huang L, Xu Y, Chen J, Liu F, Wu D, Zhou W, et al. An artificial intelligence difficulty scoring system for stone removal during ERCP: a prospective validation. *Endoscopy*. 2023;**55**(1):4-11. doi: 10.1055/a-1850-6717. PubMed PMID: 35554877.
20. Chandrasekhara V, Khashab MA, Muthusamy VR, Acosta RD, Agrawal D, Bruining DH, et al. Adverse events associated with ERCP. *Gastrointest Endosc.* 2017;**85**(1):32-47. doi: 10.1016/j.gie.2016.06.051. PubMed PMID: 27546389.

21. Lewey S, Adler DG. Fundamentals of ERCP image interpretation. Pract Gastroenterol. 2023;47(8):21.
22. Dietrich CF, Bekkali NL, Burmeister S, Dong Y, Everett SM, Hocke M, et al. Controversies in ERCP: Technical aspects. *Endosc Ultrasound*. 2022;**11**(1):27-37. doi: 10.4103/EUS-D-21-00102. PubMed PMID: 34677144. PubMed PMCID: PMC8887038.
23. Gulliver DJ, Cotton PB, Baillie J. Anatomic variants and artifacts in ERCP interpretation. *AJR Am J Roentgenol.* 1991;**156**(5):975-80. doi: 10.2214/ajr.156.5.2017963. PubMed PMID: 2017963.