Accepted Manuscript online: 2023-02-13 Article published online: 2023-04-18

Artificial intelligence for diagnosing neoplasia on digital cholangioscopy: development and multicenter validation of a convolutional neural network model

OPEN ACCESS

Authors
Carlos Robles-Medranda[1], Jorge Baquerizo-Burgos[1], Juan Alcivar-Vasquez[1], Michel Kahaleh[2], Isaac Raijman[3,4], Rastislav Kunda[5], Miguel Puga-Tejada[1], Maria Egas-Izquierdo[1], Martha Arevalo-Mora[1], Juan C. Mendez[6], Amy Tyberg[2], Avik Sarkar[2], Haroon Shahid[2], Raquel del Valle-Zavala[1], Jorge Rodriguez[1], Ruxandra C. Merfea[1], Jonathan Barreto-Perez[1], Gabriela Saldaña-Pazmiño[7], Daniel Calle-Loffredo[1], Haydee Alvarado[1], Hannah P. Lukashok[1]

Institutions
1 Gastroenterology, Instituto Ecuatoriano de Enfermedades Digestivas (IECED), Guayaquil, Ecuador
2 Gastroenterology, Robert Wood Johnson Medical School Rutgers University, New Brunswick, New Jersey, United States
3 Houston Methodist Hospital, Houston, Texas, United States
4 Baylor Saint Luke's Medical Center, Houston, Texas, United States
5 Department of Advanced Interventional Endoscopy, Universitair Ziekenhuis Brussel (UZB)/Vrije Universiteit Brussel (VUB), Brussels, Belgium
6 mdconsgroup, Artificial Intelligence Department, Guayaquil, Ecuador
7 Gastroenterology, Hospital Clínico San Carlos, Madrid, Spain

submitted 10.6.2022
accepted after revision 13.2.2023
accepted manuscript online 13.2.2023
published online 18.4.2023

Bibliography
Endoscopy 2023; 55: 719–727
DOI 10.1055/a-2034-3803
ISSN 0013-726X

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Supplementary material
Supplementary material is available under https://doi.org/10.1055/a-2034-3803

Scan this QR-Code for the author commentary.

Corresponding author
Carlos Robles-Medranda, MD, Instituto Ecuatoriano de Enfermedades Digestivas, Av. Abel Romero Castillo y, Av. Juan Tanca Marengo SN. Torre Vitalis II, Office 405-406, Guayaquil 090505, Ecuador
carlosoakm@yahoo.es

ABSTRACT

Background We aimed to develop a convolutional neural network (CNN) model for detecting neoplastic lesions during real-time digital single-operator cholangioscopy (DSOC) and to clinically validate the model through comparisons with DSOC expert and nonexpert endoscopists.

Methods In this two-stage study, we first developed and validated CNN1. Then, we performed a multicenter diagnostic trial to compare four DSOC experts and nonexperts against an improved model (CNN2). Lesions were classified into neoplastic and non-neoplastic in accordance with Carlos Robles-Medranda (CRM) and Mendoza disaggregated criteria. The final diagnosis of neoplasia was based on histopathology and 12-month follow-up outcomes.

Results In stage I, CNN2 achieved a mean average precision of 0.88, an intersection over the union value of 83.24 %, and a total loss of 0.0975. For clinical validation, a total of 170 videos from newly included patients were analyzed with the CNN2. Half of cases (50 %) had neoplastic lesions. This model achieved significant accuracy values for neoplastic diagnosis, with a 90.5 % sensitivity, 68.2 % specificity, and 74.0 % and 87.8 % positive and negative predictive values, respectively. The CNN2 model outperformed nonexpert #2 (area under the receiver operating characteristic curve [AUC]-CRM 0.657 vs. AUC-CNN2 0.794, $P<0.05$; AUC-Mendoza 0.582 vs. AUC-CNN2 0.794, $P<0.05$), nonexpert #4

(AUC-CRM 0.683 vs. AUC-CNN2 0.791, $P<0.05$), and expert #4 (AUC-CRM 0.755 vs. AUC-CNN2 0.848, $P<0.05$; AUC-Mendoza 0.753 vs. AUC-CNN2 0.848, $P<0.05$).

Conclusions The proposed CNN model distinguished neoplastic bile duct lesions with good accuracy and outperformed two nonexpert and one expert endoscopist.

Introduction

The diagnosis of malignancy in indeterminate biliary strictures is challenging [1]. Bile duct strictures can be caused by neoplastic and non-neoplastic processes, and may affect diagnostic procedures and treatment [2,3]. Endoscopic retrograde cholangiopancreatography (ERCP) with brush cytology and biopsy sampling is the most used procedure for evaluating the biliary system [2,4]. However, the low sensitivity, accuracy, and imaging limitations of this method may lead to diagnostic and sampling errors [4,5]. Additionally, if the diagnosis through ERCP is inconclusive, the lesion is categorized as indeterminate, resulting in treatment delay and requiring new evaluation [6].

Digital single-operator cholangioscopy (DSOC) is a minimally invasive diagnostic and therapeutic procedure that allows direct high-resolution visualization of the pancreaticobiliary system, tissue sample acquisition, and interventional therapies, and has better accuracy than ERCP [6–9]. However, even among experts, interobserver variability in the detection of neoplastic lesions with different DSOC classification systems can occur [5,8,10,11].

In the past decade, artificial intelligence (AI) has led to the development of innovative and more precise diagnostic tools. Convolutional neural networks (CNNs), deep learning algorithms, are commonly used in the medical field to identify features in images and videos, and assist in image interpretation [12,13]. Most of the AI tools studied in gastrointestinal endoscopy have focused on luminal endoscopic procedures and cancer detection [14]; however, few studies on the diagnostic performance of DSOC CNN models have been performed [15,16]. Therefore, we aimed to develop a CNN model that recognizes macroscopic morphological characteristics of neoplastic lesions during real-time DSOC as a red flag technique, and to clinically validate the model through comparisons with DSOC used by expert and nonexpert endoscopists.

Methods

Study design and ethics

This two-stage study was designed to develop, train, and validate the performance of a CNN model that identifies neoplastic lesions in indeterminate biliary strictures in prerecorded and during real-time DSOC procedures. The study protocol was approved by the Instituto Ecuatoriano de Enfermedades Digestivas (IECED) Institutional Review Board (Guayaquil, Ecuador), and participating centers, and was conducted in accordance with the Declaration of Helsinki. Patients or their legal guardians provided written informed consent to transfer their DSOC videos to the AIworks Cloud (mdconsgroup, Guayaquil, Ecuador) for analysis and publication.

Stage I

Design and patient sample

Stage I was an observational, analytic, prospective single-center diagnostic pilot study that aimed to develop, train, and internally validate a CNN model that identifies neoplastic features in prerecorded videos and real-time DSOC procedures at IECED between January 2020 and October 2020. Based on histological findings and 12-month follow-up results (clinical and paraclinical), two cohorts were defined: patients with neoplastic lesions and patients with non-neoplastic lesions.

All patients ≥18 years referred for DSOC owing to suspected common bile duct tumor or bile duct stenosis (including common hepatic duct, hilar, and intrahepatic lesions), and who approved the recording of their DSOC procedures (regardless of the reason), were invited to participate in the study.

Patients were excluded if they met any of the following criteria: a) pre-existing clinical conditions that made DSOC unviable; b) previous DSOC procedure; c) low-quality/unrecognizable pattern in DSOC recordings, and d) inability to participate in a 12-month post-DSOC follow-up.

DSOC procedure

The DSOC procedures were conducted by two expert endoscopists (C.R.M., J.A.V) who each performed >150 DSOCs per year. All patients were placed in a supine position under general anesthesia and received antibiotic prophylaxis. Patients were first assessed using a standard duodenoscope (Pentax ED 3670TK; Pentax Medical, Hoya Corp., Tokyo, Japan), the Pentax video processors EPK-I and EPK-i5010, and a second-generation SpyGlass DS digital system (Boston Scientific, Marlborough, Massachusetts, USA). The SpyScope DS II catheter (SpyGlass DS Digital System) was passed proximally, suction was used to remove bile and contrast material, sterile saline solution was infused to optimize imaging, and the cholangioscope was slowly withdrawn, providing a systematic inspection of the ductal mucosa. Videos were recorded using a high definition image capture system (AIWorks Cloud, mdconsgroup, Guayaquil, Ecuador).

Development and model validation of the CNN

A first version of the model (CNN1) was developed using the DSOC videos of 23 patients with definitive neoplastic lesion diagnoses based on histological findings and 12-month follow-ups. First, based on the visual examinations performed by the experts, macroscopic neoplastic features were classified in accordance with the Carlos Robles Medranda (CRM) and Mendoza classifications. Lesions were recorded and labeled within a bounding box using the AIworks Cloud [4,8] (▶ **Table 1**, ▶ **Fig. 1a**). This platform was designed and developed to cap-

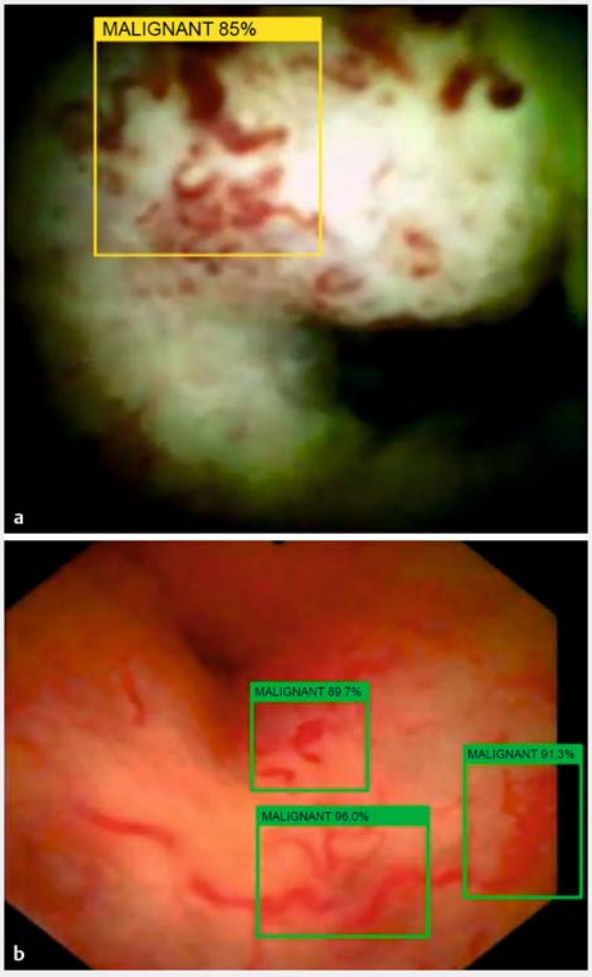
▶ **Table 1** Neoplastic lesion criteria based on the Carlos Robles-Medranda and Mendoza classifications and disaggregated neoplasia criteria.

CRM classification		Mendoza classification
Type I	Flat/smooth or irregular surface and Irregular/spider vascularity	Tortuous and dilated vessels, or Irregular nodulations, or Raised intraductal lesions, or Irregular surface with or without ulcerations, or Friability
Type II	Polypoid and Irregular/spider vascularity	
Type III	Ulcerated and infiltrative pattern and Irregular/spider vascularity	
Type IV	Honeycomb pattern with/without Irregular/spider vascularity	

Disaggregated neoplasia criteria

a) Presence or absence of tortuous and dilated vessels
b) Presence or absence of irregular mucosal surfaces
c) Presence or absence of polypoid lesions
d) Presence or absence of irregular nodulations
e) Presence or absence of raised intraductal lesions
f) Presence or absence of ulcerations
g) Presence or absence of honeycomb pattern
h) Presence or absence of friability

CRM, Carlos Robles-Medranda



▶ **Fig. 1** Still images from the real-time characterization of a biliary stricture during a digital cholangioscopy evaluation using the developed artificial intelligence model. The bounding box highlights the area in the biliary lesion suggestive of neoplasia. The color of the bounding box represents the convolutional neural network (CNN) model used. **a** A yellow bounding box was used for CNN1. This model was used in conjunction with a second-generation SpyGlass Digital System DS (Boston Scientific, Marlborough, Massachusetts, USA) in a neoplastic lesion. **b** CNN2 detection of a neoplastic lesion was highlighted within a green bounding box using Eye-Max microendoscopic system (Micro-Tech, Nanjing, China).

ture frames of the uploaded videos in which CNN1 was applied. After a definitive diagnosis was confirmed at the 12-month follow-up, the collected frames were fed to the AIworks software. To prevent overfitting of the models, we applied early stopping regularization. This regulation process recognizes the best metrics among training processes, and automatically notifies the training point with the best metric results and the point at which the model starts overfitting.

CNN1 was developed using YOLOv3 (You Only Look Once version 3; Seattle, Washington, USA) and trained on 90 % of the frames using an Nvidia RTX 2080ti GPU (Micro-Star International, Zhonghe District, China). The remaining 10 % of frames were used for model validation to assess the performance of CNN1. Additionally, clinical validation was performed with 25 cases that were not previously involved, which included 20 fine-tuned recorded videos and five real-time cholangioscopy procedures of patients with indeterminate biliary lesions.

A second version of the model (CNN2) was developed for the second stage of the study using 116 additional DSOC videos with identified neoplastic lesions utilizing the aforementioned criteria. This improved version of CNN1 was developed using YOLOv4, with a 90 % training and 10 % validation dataset distribution.

Four metrics were obtained from the model validation process for both CNNs. The mean average precision (mAP) represents the ability of the model to detect a trained feature; total loss denotes the difference between the prediction values and the expected results; the F1 score is the association between the model's recall and precision; and the intersection over the union (IoU) represents the model's precision to identify an object within the detection box.

Stage II

Study design

Stage II was an observational, analytic, nested case–control, multicenter diagnostic trial that aimed to clinically validate CNN2 on prerecorded DSOC videos of treatment-naïve patients from different centers between October 2020 and December 2021. In addition to our own center, three out of the four highly experienced international DSOC centers that were invited participated in this stage as part of the expert endoscopists group. The study cohorts, selection criteria, and DSOC procedure protocols for stage II were the same as those for stage I. During stage II, a second-generation SpyGlass DS digital system (96/170) and an Eye-Max microendoscopic system (74/170) (Micro-Tech, Nanjing, China) were used according to availability at the endoscopy units (▶ **Fig. 1b**).

Expert and nonexpert DSOC video assessment

Four expert endoscopists (>150 DSOCs per year), one from each center (J.A.V., I.R., R.K., and M.K.), uploaded the DSOC videos from their respective endoscopy units to the AIworks Cloud. Within the cloud, the same four expert endoscopists and four nonexpert general practitioners (J.B.-B., M.E.-I., M.A.-M., and G.S.-P.) who were blinded to clinical records, classified the uploaded DSOC videos as neoplastic or non-neoplastic based on the criteria of the CRM and Mendoza classifications [5, 8] in a disaggregated manner (▶ **Table 1**). None of the experts assessed patient videos from their own center. CNN2 was applied to the uploaded DSOC videos and also classified them as neoplastic or non-neoplastic based on the CRM and Mendoza neoplasia criteria [4, 8]. Both groups (experts and nonexperts) completed an online dataset and marked the presence or absence of macroscopic features (▶ **Table 1**). The number of videos observed by the participants depended on the number of cases they provided to the study; the assessment was split into four sessions by dividing the video set into two equal sets. For this investigation, the definition of experts and nonexperts is not based on a standard definition. There was a high probability of tiredness in experts and nonexperts during the assessment of such a high volume of DSOC videos, especially as they were blinded to digital records. Both represented the main potential biases in this study.

Statistical analysis

For both stages, statistical analyses were performed using R version 4.0.3 (R Foundation for Statistical Computing, Vienna, Austria) by our biostatistician (M.P.-T.). The diagnostic accuracy, which was defined as the sensitivity, specificity, positive and negative predictive values (PPV and NPV), and observed agreement, of both CNN models was calculated based on both patients and frames. The frame-based diagnostic accuracy considered frames of interest in the patient's prerecorded videos. Histological findings and 12-month follow-up results were considered the gold standard. A P value of <0.05 was considered statistically significant.

For stage II, the number of neoplastic or non-neoplastic cases required was calculated using the power.diagnostic.test function in the MKmisc package (version 1.6) [17]. Assumption for the proportion of discordant pairs considered a 94.7% sensitivity for a CNN model, as estimated by Saraiva M et al. [16]. A probabilistic sample was designed considering a 10% delta and a 50% prevalence (1:1 case vs. controls ratio, to recreate the same probability of neoplastic or non-neoplastic cases during DSOC videos assessment, as in the Bernoulli trial). A 5% significance level was considered, along with a defined 5% and 20% alpha and beta error, respectively. Based on the above, a sample size of 66 neoplastic or non-neoplastic cases was estimated with an 80% statistical power (an initial sample of 122 cases was expected). Moreover, considering the five invited centers, the final number included a total of 170 patients in stage II, with approximately 33–34 patients at each participating center (see the online-only **Supplementary material)**.

The numerical variables are described as the mean (SD) or median (interquartile range [IQR]), depending on their statistical distribution assessed with the Kolmogorov–Smirnov test. The corresponding categorical variables are described as frequencies (%) with 95%CI.

The baseline characteristics of the neoplastic and non-neoplastic cases were compared through statistical hypothesis testing (Wilcoxson test, chi-squared, or Fisher's exact test). Any missing baseline or 12-month follow-up data were described if necessary. The cases in which experts or nonexperts omitted any assessment were excluded.

We calculated the diagnostic accuracy and the area under the receiver operating characteristic (ROC) curve (AUC) of the experts, nonexperts, and patient-based CNN2 model. The diagnostic accuracy of the expert and nonexpert estimations was calculated separately for the CRM and Mendoza classifications. The comparison between CNN2 and diagnostic accuracies through the CRM or Mendoza classifications were defined through DeLong's test for two ROC curves, from dependent samples. Histological findings and 12-month follow-ups were considered the gold standard.

Results

Stage I

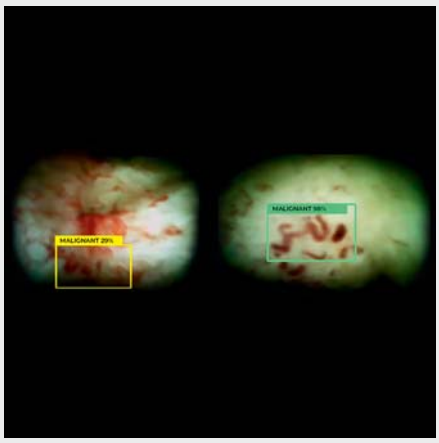
CNN1 model development and validation

A total of 81 080 frames from 23 patients distributed into training and testing datasets were used to develop CNN1. This model achieved mAP of 0.29, IoU of 32.2, F1 score of 0.280, and total loss of 0.1034 (**Table 1 s**).

CNN1 clinical validation

During the clinical validation, an additional 25 consecutive treatment-naïve patients were included. CNN1 accurately detected tumor vessels in 10/10 histology-confirmed cholangiocarcinoma cases and 5/5 real-time endoscopic procedures (▶ **Video 1**).

Additionally, this model was tested with 10 patient videos in which cholangioscopy was performed to confirm stone removal following laser lithotripsy; 2/10 videos were incorrectly classified as positive for neoplasia by CNN1, most likely due to the



Video 1 Artificial intelligence-aided digital cholangioscopy procedure in a patient with a neoplastic biliary lesion using convolutional neural network model version 1 (CNN1) followed by the upgraded CNN2; the bounding boxes indicate areas suggestive of neoplasia.
Online content viewable at:
https://doi.org/10.1055/a-2034-3803

detection of a stone-related inflammatory pattern. In the per-patient analysis, the CNN1 model reached a 92.0% observed agreement, with a sensitivity, specificity, PPV, and NPV of 100%, 80.0%, 88.0%, and 100%, respectively. In the per-frame analysis, CNN1 reached a 97.0% observed agreement, with a sensitivity, specificity, PPV, and NPV of 98.0%, 95.0%, 98.0%, and 94.0%, respectively (**Table 1 s**, **Fig. 1 s**).

CNN2 model development and validation

After CNN1 was clinically validated, the model was improved. An additional 116 patients with a definitive neoplastic diagnosis were enrolled, and their corresponding videos were uploaded to AIworks Clouds for training and model validation. A total of 198 941 frames were used to develop CNN2 (159 153 frames for training and 39 788 frames for validation). CNN2 had a reading rate of 30–60 frames per second, with a 5-second validation data reading. The developed model achieved mAP of 0.88, total loss of 0.0975, F1 score of 0.738, and IoU average of 83.24% (**Table 1 s**).

Stage II

Baseline characteristics

A total of 170 treatment-naïve patients from the participating centers were included and equally distributed into neoplastic and non-neoplastic groups to clinically validate CNN2. The median patient age was 62.5 years (IQR 57.0–68.8), and 46.5% of the patients were female. The most common DSOC indication observed in our study was tumor suspicion (34.1%), followed by indeterminate stenosis (27.1%) and indeterminate dilations (18.2%). The strictures were located at the common bile duct (28.2%), hilum (28.2%), common hepatic duct (41.2%), and intrahepatic duct (2.4%); they were evenly distributed among the cases of neoplastic and non-neoplastic DSOC findings (▶ **Table 2**).

CNN2 clinical validation

The diagnostic accuracy of CNN2 was obtained for both frames and patients (**Table 1 s**). The diagnostic accuracy of the model in detecting lesions per frame was 98%, with a 98.6% sensitivity, 98.0% specificity, 89.2% PPV, and 99.2% NPV. The diagnostic accuracy of the model per patient reached an observed agreement of 80.0%, a sensitivity of 90.5%, a specificity of 68.2%, a PPV of 74.0%, and an NPV of 87.8%.

Comparison of CNN2 with the expert and nonexpert groups

We evaluated the overall diagnostic accuracies among expert and nonexpert groups with both DSOC classification systems and compared the results with those of CNN2 and the follow-ups. In the expert group, CNN2 achieved statistical significance with both classifications compared with expert #4. This expert obtained a 92.7% sensitivity, 48.5% specificity, 64.3% PPV, and 86.8% NPV with the CRM classification criteria, and a 100% sensitivity, 2.9% specificity, 50.8% PPV, and 100% NPV with the Mendoza classification criteria. The observed agreements for the above classifications were 70.6% and 51.5%, respectively. The AUC for expert #4 using the CRM and Mendoza classifications were 0.755 ($P=0.005$) and 0.753 ($P=0.04$), respectively, when compared with CNN2. The other experts' results are summarized in ▶ **Table 3**. It should be noted that expert #1 assessed a smaller sample because their center provided additional patients in the absence of one of the other four invited centers, as explained in **Table 2 s**.

In the nonexpert group, nonexpert #2 achieved significant difference between the two classifications and CNN2. The observed agreement was 65.7% for the CRM classification and 55.9% for the Mendoza classification. The AUCs of the CRM and Mendoza classifications were 0.657 ($P=0.01$) and 0.582 ($P<0.001$), respectively, which were significantly lower than the corresponding value for CNN2 (AUC 0.794). Additionally, CNN2 performed significantly better than nonexpert #4 with the CRM classification criteria ($P<0.05$), with a sensitivity of 80.9% and an NPV of 73.5%. The observed agreement was 66.9%. The AUCs of the nonexpert using the CRM and Mendoza classifications were 0.683 ($P=0.04$) and 0.737 ($P=0.31$), individually, compared with 0.791 for CNN2. The data of the remaining nonexperts are summarized in **Table 3 s**.

A pooled analysis of experts and nonexperts, using the CRM and Mendoza criteria, was calculated and is summarized in ▶ **Table 4**. According to the results, the CNN2 was superior to both groups.

Discussion

To date, despite the numerous advantages of DSOC, there is an ongoing discrepancy between operators' visual impressions using current classifications for indeterminate biliary lesions. To overcome this limitation, the application of new technolo-

▶ **Table 2** Baseline characteristics of patients in stage II.

	Total (n = 170)	Neoplasia (n = 85)	Non-neoplasia (n = 85)	*P* value
Age, median (IQR), years	62.5 (57.0–68.8)	64.0 (59.0–71.0)	59.0 (52.0–65.0)	<0.001[1]
Female sex, n (%)	79 (46.5)	45 (52.9)	34 (40.0)	0.12[2]
DSOC indication, n (%)				<0.001[2]
▪ Suspicion of tumor	58 (34.1)	49 (57.6)	9 (10.6)	
▪ Indeterminate stenosis	46 (27.1)	15 (17.6)	31 (36.5)	
▪ Indeterminate dilation	31 (18.2)	21 (24.7)	10 (11.8)	
▪ Filling defect	35 (20.6)	–	35 (41.2)	
Clinical presentation[3], n (%)				
▪ Jaundice	127 (74.7)	77 (90.6)	50 (58.8)	<0.001[2]
▪ Pruritus	59 (34.7)	34 (40.0)	25 (29.4)	0.20[2]
▪ Abdominal pain	76 (44.7)	56 (65.9)	20 (23.5)	<0.001[2]
▪ Weight loss	77 (45.3)	73 (85.9)	4 (4.7)	<0.001[2]
Total bilirubin, median (IQR)	3.89 (2.50–9.00)	9.00 (4.50–22.60)	3.00 (0.90–3.50)	<0.001[1]
Stricture location, n (%)				<0.001[4]
▪ Common bile duct	48 (28.2)	13 (15.3)	35 (41.2)	
▪ Hilum	48 (28.2)	39 (45.9)	9 (10.6)	
▪ Common hepatic duct	70 (41.2)	33 (38.8)	37 (43.5)	
▪ Intrahepatic	4 (2.4)	–	4 (4.7)	
▪ Cystic duct	–	–	–	
Previous ERCP, n (%)	54 (31.8)	19 (22.4)	35 (41.2)	0.01[2]
Previous stent placement, n (%)	44 (25.9)	15 (17.6)	29 (34.1)	0.02[2]
DSOC diagnosis, n (%)				<0.001[2]
▪ Non-neoplasia	85 (50.0)	–	85 (100)	
▪ Neoplasia	85 (50.0)	85 (100)	–	
Histological diagnosis, n (%)				<0.001[4]
▪ Adenocarcinoma	11 (6.5)	11 (12.9)	–	
▪ Atypical	6 (3.5)	6 (7.1)	–	
▪ Cholangiocarcinoma	67 (39.4)	67 (78.8)	–	
▪ Inflammatory	69 (40.6)	–	69 (81.2)	
▪ IPMN of the bile duct	1 (0.6)	1 (1.2)	–	
▪ Normal biliary tissue	2 (1.2)	–	2 (2.4)	
▪ Primary sclerosing cholangitis	14 (8.2)	–	14 (16.5)	

IQR, interquartile range; DSOC, digital single-operator cholangioscopy; ERCP, endoscopic retrograde cholangiopancreatography; IPMN, intraductal papillary mucinous neoplasm.
[1] Wilcoxon rank sum test with continuity correction.
[2] Pearson's chi-squared test with Yates' continuity correction.
[3] Not mutually exclusive categories.
[4] Fisher's exact test for count data

▶ **Table 3** Comparison of diagnostic accuracy between experts using the Carlos Robles-Medranda and Mendoza classifications and convolutional neural network model version 2 in clinical validation.

	Sensitivity n/N (%) [95 %CI]	Specificity n/N (%) [95 %CI]	PPV n/N (%) [95 %CI]	NPV n/N (%) [95 %CI]	Observed agreement n/N (%) [95 %CI]	AUC (*P* value)
Expert 1 (n = 94)						
CRM	43/47 (91.5) [79.6–97.6]	36/47 (76.6) [61.4–89.7]	43/54 (79.6) [66.5–89.4]	36/40 (90.0) [76.3–97.2]	79/94 (84.0) [75.1–90. 8]	0.836 (0.82)
Mendoza	47/47 (100) [92.5–100.0]	4/47 (8.5) [2.4–20.4]	47/90 (52.2) [41.4–62.9]	4/4 (100) [32.8–100.0]	51/94 (54.3) [43.7–64.6]	0.761 (0.06)
CNN2	46/47 (97.9) [88.7–99.9]	28/47 (59.6) [44.3–73.6]	46/65 (70.8) [58.2–81.4]	28/29 (96.6) [82.2–99.9]	74/94 (78.7) [69.1–86.5]	0.848
Expert 2 (n = 135)						
CRM	60/67 (89.6) [79.7–95.7]	38/68 (55.9) [43.3–67.9]	60/90 (66.7) [55.9–76.3]	38/45 (84.4) [70.5–93.5]	98/135 (72.6) [64.3–79.9]	0.755 (0.50)
Mendoza	67/67 (100) [94.6–100.0]	29/68 (42.7) [30.7–55.2]	67/106 (63.2) [53.3–72.4]	29/29 (100) [88.1–100.0]	96/135 (71.1) [62.7–78.6]	0.816 (0.54)
CNN2	59/67 (88.1) [77.8–94.7]	46/68 (67.7) [55.2–78.5]	59/81 (72.8) [61.8–82.1]	46/54 (85.2) [72.9–93.4]	105/135 (77.8) [69.8–84.5]	0.790
Expert 3 (n = 136)						
CRM	57/68 (83.8) [72.9–91.6]	44/68 (64.7) [52.2–75.9]	57/81 (70.4) [59.2–80.0]	44/55 (80.0) [67.0–89.6]	101/136 (74.3) [66.1–81.4]	0.803 (0.78)
Mendoza	68/68 (100) [94.7–100.0]	24/68 (35.3) [24.1–47.8]	68/112 (60.7) [51.0–69.8]	24/24 (100) [85.8–100.0]	92/136 (67.7) [59.1–75.4]	0.751 (0.43)
CNN2	60/68 (88.2) [78.1–94.8]	46/68 (67.7) [55.2–78.5]	60/82 (73.2) [62.2–82.3]	46/54 (85.2) [72.9–93.4]	106/136 (77.9) [70.0–84.6]	0.791
Expert 4 (n = 136)						
CRM	63/68 (92.7) [83.7–97.6]	33/68 (48.5) [36.2–60.9]	63/98 (64.3) [53.9–73.7]	33/38 (86.8) [71.9–95.6]	96/136 (70.6) [62.2–78.1]	0.755 (0.005)
Mendoza	68/68 (100) [94.7–100.0]	2/68 (2.9) [0.4–10.2]	68/134 (50.8) [41.9–59.5]	2/2 (100) [15.8–100.0]	70/136 (51.5) [42.8–60.1]	0.753 (0.04)
CNN2	67/68 (98.5) [92.1–99.9]	42/68 (61.8) [49.2–73.3]	67/93 (72.0) [61.9–80.9]	42/43 (97.7) [87.7–99.9]	109/136 (80.2) [72.5–86.5]	0.848

PPV, positive predictive value; NPV, negative predictive value; AUC, area under the receiver operating characteristic curve; CRM, Carlos Robles-Medranda classification; CNN2, convolutional neural network model version 2.

gies to aid image interpretation has been proposed; however, the proposed models could only be applied to images [15, 16]. In the present study, we developed a new DSOC-based CNN for recognizing neoplasia in indeterminate biliary lesions in prerecorded videos and real-time DSOC procedures, and compared the model with DSOC experts and nonexperts using the CRM and Mendoza classifications. We found that the model achieved high diagnostic accuracy for detecting neoplastic lesions when analyzing frames and real-time endoscopic procedures, with a significantly better performance than the nonexpert group. This is the first international multicenter study to develop and validate a CNN model applicable to prerecorded videos and live procedures for detection of neoplastic lesions in treatment-naïve patients.

The DSOC classifications (CRM and Mendoza) used to identify macroscopic features of biliary duct neoplasia are dependent on the expertise of the physician, and thus, the observed agreement varies, even among experts [8, 11]. We compared our CNN2 model against experts and nonexperts using both classifications. In the expert group, the observed agreement between the model and experts using the CRM classification ranged between 70.6 % and 84.0 %, while the corresponding values for the Mendoza classification ranged between 51.5 % and 71.1 %. In the nonexpert group, the observed agreement of nonexperts using the two classifications was lower than that in the expert group: 67.2 % vs. 75.4 % and 60.8 % vs. 61.1 % for the CRM and Mendoza classification, respectively. Additionally, when using the Mendoza classification in this study, experts and nonexperts achieved a low specificity and PPV, with values ranging from 2.9 % to 42.7 % and 50.8 % to 63.2 %, respectively. The difference between the two classifications may be attributed to the difference in characteristics used to assess neoplastic

▶ **Table 4** Pooled analysis of diagnostic accuracy of the artificial intelligence model, expert, and nonexpert endoscopists.

	Sensitivity, % (95 %CI)	Specificity, % (95 %CI)	PPV, % (95 %CI)	NPV, % (95 %CI)	Agreement, % (95 %CI)
AICNN2	90.6 (82.3–95.8)	68.2 (77.9–57.2)	74.0 (82.1–64.5)	87.9 (77.5–94.6)	80.0 (72.5–85.2)
Experts (CRM)	89.4 (83.1–95.6)	61.4 (42.2–80.7)	70.2 (59.5–80.9)	85.3 (78.6–92.0)	75.4 (65.9–84.9)
Nonexperts (CRM)	82.9 (57.6–100)	51.1 (25.1–77.1)	63.7 (57.7–69.8)	78.6 (59.9–97.3)	67.2 (60.9–73.4)
Experts (Mendoza)	100	22.4 (–8.8–53.5)	56.7 (46.9–66.5)	100	61.1 (45.7–76.6)
Nonexperts (Mendoza)	93.3 (81.3–100)	27.9 (15.7–40.1)	56.7 (52.2–61.3)	82.8 (60.1–100)	60.8 (53.2–68.5)
Kappa Fleiss of experts using CRM:			$\kappa = 0.053$ ($P = 0.13$)		
Kappa Fleiss of experts using Mendoza:			$\kappa = 0.009$ ($P = 0.80$)		
Kappa Fleiss of nonexperts using CRM:			$\kappa = 0.237$ ($P < 0.001$)		
Kappa Fleiss of nonexperts using Mendoza:			$\kappa = 0.441$ ($P < 0.001$)		

PPV, positive predictive value; NPV, negative predictive value; CRM, Carlos Robles-Medranda classification.

lesions. For the CRM classification, a physician detects different macroscopic patterns [5]; in contrast, the Mendoza classification requires the identification of only one out of five parameters to determine whether a lesion is neoplastic or non-neoplastic [8], which may increase the number of false-positive cases.

To clinically validate CNN2, we included 170 additional treatment-naïve patients from four different endoscopy units. When the diagnostic accuracy was analyzed in terms of frames, our model achieved a 98.6 % sensitivity, 98.0 % specificity, 89.2 % PPV, and 99.2 % NPV. These results are similar to those of two recent studies by Pereira et al. and Saraiva et al., who presented CNN models that detected tumor vessels during DSOC using 85 subjects [15, 16]. The authors extracted frames from these patients in both studies (11 855 and 6475, respectively) and divided the frames into training (80 % of the frames) and testing (the remaining 20 %) datasets. In their studies, they obtained the diagnostic accuracy of their models in terms of frames, with 94.7 % sensitivity, 92.1 % specificity, 94.8 % PPV, and 84.2 % NPV [16], with similar results in the second study [15]. However, these models could not be applied to prerecorded videos nor live procedures.

Given that AI assistance should be applied in real time (rather than after the procedure) if it is to aid in diagnosis and image interpretation, we consider that the diagnostic accuracy and clinical validation of any CNN model should be based on cases rather than frames. Thus, we obtained the diagnostic accuracy of CNN2 in detecting neoplastic lesions and compared the results with the final diagnoses based on histological findings and 12-month follow-up data. CNN2 achieved a 90.5 % sensitivity, 68.2 % specificity, 74.0 % PPV, and 87.8 % NPV, with an observed agreement of 80.0 %. Hence, using frames for diagnostic accuracy cannot estimate the true diagnostic value of CNN models in clinical practice, and using cases instead of frames would provide a more accurate clinical validation. Furthermore, as we continue to upload samples to the cloud, the model will automatically update itself, which will further increase its diagnostic accuracy.

Nonetheless, the key advantage of our CNN model over other DSOC-based CNN models, is that it can be applied during real-time DSOC procedures, leading to more conclusive diagnostic results. This advantage could eliminate the need for repeated invasive procedures or possible delays in curative surgery. Furthermore, the accurate diagnosis of neoplastic lesions in patients with biliary disorders and prompt therapeutic responses may improve overall survival and/or decrease differences between visual examinations and histological results by improving targeted biopsy sampling [18]. Additionally, this model may be able to shorten the DSOC learning curve, as it may help increase a trainee's confidence in their diagnostic visual impression, thus reducing the missed lesion rate.

The potential clinical benefits of using the AI system during DSOC include: a) provision of a second opinion on lesions suggestive of neoplasia, helping expert and nonexpert endoscopists obtain a targeted sample; b) improvement in cost-effectiveness following adequate AI-guided tissue sampling; c) reduction in interobserver agreement mismatch among experts and nonexperts; d) reduction in the variance between visual impression and histology; and e) potential use as a training tool. Therefore, further studies evaluating the clinical application of the proposed DSOC-based CNN model in terms of assisting with diagnosis, biopsy sampling, and training new endoscopists to recognize neoplastic signs in biliary lesions should be conducted. Moreover, this cholangioscopy CNN model may lead to increased DSOC availability, which has been proven to facilitate clinical care for such patients.

In conclusion, the proposed CNN2 model accurately recognized and classified biliary lesions as neoplastic in prerecorded videos and real-time DSOC procedures in treatment-naïve patients. Furthermore, our proposed CNN model effectively outperformed experts and nonexperts.

Competing Interests

C. Robles-Medranda is a key opinion leader and consultant for Pentax Medical, Steris, Micro-tech, G-Tech Medical Supply, CREO Medical, and mdconsgroup, and is a board member and consultant for Endo-Sound. M. Kahaleh is a consultant for Boston Scientific, Interscope Med, and Abbvie; a grant recipient from Boston Scientific, Conmed, Gore, Pinnacle, Merit Medical, Olympus Medical, and Ninepoint Medical; and the chief executive officer and founder of Innovative Digestive Health Education & Research Inc. A. Tyberg is a consultant for Ninepoint Medical, EndoGastric Solutions, and Obalon Therapeutics. I. Raijman is a speaker for Boston Scientific, ConMed, Medtronic, and GI Supplies; an advisory board member for Microtech; and a co-owner of EndoRx. R. Kunda is a consultant for Olympus, Boston Scientific, Omega Medical Imaging, M.I. Tech, Tigen Pharma, and Ambu. J. Baquerizo-Burgos, J. Alcivar-Vasquez, M. Puga-Tejada, M. Egas-Izquierdo, M. Arevalo-Mora, J.C. Mendez, A. Sarkar, H. Shahid, R. del Valle-Zavala, J. Rodriguez, R.C. Merfea, J. Barreto-Perez, G. Saldaña-Pazmiño, D. Calle-Loffredo, H. Alvarado, and H.P. Lukashok declare that they have no conflict of interest.

Clinical trial

Trial Registration: ClinicalTrials.gov | Registration number (trial ID): NCT05147389 | Type of study: Prospective, Multi-Center Study

References

[1] Navaneethan U, Njei B, Lourdusamy V et al. Comparative effectiveness of biliary brush cytology and intraductal biopsy for detection of malignant biliary strictures: a systematic review and meta-analysis. Gastrointest Endosc 2015; 81: 168–176

[2] Yoon WJ, Brugge WR. Endoscopic evaluation of bile duct strictures. Gastrointest Endosc Clin N Am 2013; 23: 277–293

[3] Robles-Medranda C, Oleas R, Sánchez-Carriel M et al. Vascularity can distinguish neoplastic from non-neoplastic bile duct lesions during digital single-operator cholangioscopy. Gastrointest Endosc 2021; 93: 935–941

[4] Chung HG, Chang JI, Lee KH et al. Comparison of EUS and ERCP-guided tissue sampling in suspected biliary stricture. PLoS One 2021; 16: 1–12

[5] Robles-Medranda C, Valero M, Soria-Alcivar M et al. Reliability and accuracy of a novel classification system using peroral cholangioscopy for the diagnosis of bile duct lesions. Endoscopy 2018; 50: 1059–1070

[6] Gerges C, Beyna T, Tang RSY et al. Digital single-operator peroral cholangioscopy-guided biopsy sampling versus ERCP-guided brushing for indeterminate biliary strictures: a prospective, randomized, multicenter trial (with video). Gastrointest Endosc 2020; 91: 1105–1013

[7] Fukuda Y, Tsuyuguchi T, Sakai Y et al. Diagnostic utility of peroral cholangioscopy for various bile-duct lesions. Gastrointest Endosc 2005; 62: 374–382

[8] Kahaleh M, Gaidhane M, Shahid HM et al. Digital single-operator cholangioscopy interobserver study using a new classification: the Mendoza Classification (with video). Gastrointest Endosc 2022; 95: 319–326

[9] de Oliveira PVAG, de Moura DTH, Ribeiro IB et al. Efficacy of digital single-operator cholangioscopy in the visual interpretation of indeterminate biliary strictures: a systematic review and meta-analysis. Surg Endosc 2020; 34: 3321–3329

[10] Sethi A, Tyberg A, Slivka A et al. Digital single-operator cholangioscopy (DSOC) improves interobserver agreement (IOA) and accuracy for evaluation of indeterminate biliary strictures: the Monaco Classification. J Clin Gastroenterol 2022; 56: e94–e97

[11] Kahaleh M, Raijman I, Gaidhane M et al. Digital cholangioscopic interpretation: when North meets the South. Dig Dis Sci 2021; 67: 1345–1351

[12] Sarvamangala DR, Kulkarni Raghavendra V. Convolutional neural networks in medical image understanding: a survey. Evol Intell 2022; 15: 1–22

[13] Chan HP, Samala RK, Hadjiiski LM et al. Deep learning in medical image analysis. Adv Exp Med Biol 2020; 1213: 3–21

[14] le Berre C, Sandborn WJ, Aridhi S et al. Application of artificial intelligence to gastroenterology and hepatology. Gastroenterology 2020; 158: 76–94

[15] Pereira P, Mascarenhas M, Ribeiro T et al. Automatic detection of tumor vessels in indeterminate biliary strictures in digital single-operator cholangioscopy. Endosc Int Open 2022; 10: E262–268

[16] Saraiva MM, Ribeiro T, Ferreira JPS et al. Artificial intelligence for automatic diagnosis of biliary stricture malignancy status in single-operator cholangioscopy: a pilot study. Gastrointest Endosc 2022; 95: 339–348

[17] Kohl M. CRAN package MKmisc. Accessed 18 April 2022. https://cran.r-project.org/web/packages/MKmisc/index.html

[18] Shah RJ, Raijman I, Brauer B et al. Performance of a fully disposable, digital, single-operator cholangiopancreatoscope. Endoscopy 2017; 49: 651–658