Multimodal convolutional neural network–based algorithm for real-time detection and differentiation of malignant and inflammatory biliary strictures in cholangioscopy: a proof-of-concept study (with video)

Joceline Ziegler, MSc,[1,*] Philipp Dobsch, MD,[2,*] Marten Rozema, MSc,[1] Ina Zuber-Jerger, MD,[2] Kilian Weigand, MD,[2,3] Stefan Reuther, MSc,[1] Martina Müller, MD,[2] Arne Kandulski, MD[2]

Regensburg, München, Koblenz, Germany

GRAPHICAL ABSTRACT

Multimodal CNN-Based Algorithm for Real-Time Detection and Differentiation of Malignant and Inflammatory Biliary Strictures in Cholangioscopy

Context

The characterization of biliary strictures stands out as a challenging endeavor in the field of cholangioscopy.

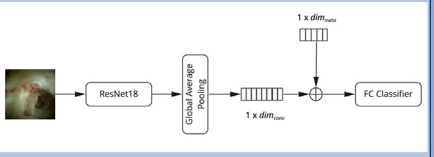
Convolutional neural networks have the potential to enhance the differentiation of biliary lesions through image-based analysis.

Incorporating clincal metadata holds the potential to enhance the accuracy and efficacy of CNN performance.

Study

A CNN algorithm was developed based on >200,000 video frames and 806 cholangioscopic still images to distinguish malignant, inflamed, and normal-appearing bile ducts.

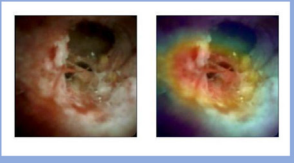
Clinical metadata was incorporated using a concatenation-based approach.



Results

CNN performance was analyzed on a prospective test cohort with full video sequences on a case-based scenario.

Concatenation based CNN algorithm
Sensitivity 87,4%
Specificity 83,0%



Model embeddings visualized; class-activation maps pinpoint relevant image regions.

Background and Aims: Deep learning algorithms gained attention for detection (computer-aided detection [CADe]) of biliary tract cancer in digital single-operator cholangioscopy (dSOC). We developed a multimodal convolutional neural network (CNN) for detection (CADe), characterization and discriminating (computer-aided diagnosis [CADx]) between malignant, inflammatory, and normal biliary tissue in raw dSOC videos. In addition, clinical metadata were included in the CNN algorithm to overcome limitations of image-only models.

Methods: Based on dSOC videos and images of 111 patients (total of 15,158 still frames), a real-time CNN-based algorithm for CADe and CADx was developed and validated. We established an image-only model and metadata injection approach. In addition, frame-wise and case-based predictions on complete dSOC video sequences were validated. Model embeddings were visualized, and class activation maps highlighted relevant image regions.

Results: The concatenation-based CADx approach achieved a per-frame area under the receiver-operating characteristic curve of .871, sensitivity of .809 (95% CI, .784-.832), specificity of .773 (95% CI, .761-.785), positive predictive value of .450 (95% CI, .423-.467), and negative predictive value of .946 (95% CI, .940-.954) with respect to malignancy

on 5715 test frames from complete videos of 20 patients. For case-based diagnosis using average prediction scores, 6 of 8 malignant cases and all 12 benign cases were identified correctly.

Conclusions: Our algorithm distinguishes malignant and inflammatory bile duct lesions in dSOC videos, indicating the potential of CNN-based diagnostic support systems for both CADe and CADx. The integration of non-image data can improve CNN-based support systems, targeting current challenges in the assessment of biliary strictures. (Gastrointest Endosc 2025;101:830-42.)

(footnotes appear on last page of article)

Early diagnosis of biliary tract cancer (BTC) is of high clinical importance.[1,2] The differentiation from inflammation is a significant task in clinical decision-making.[3]

Abnormal constrictions of the bile duct, known as indeterminate biliary strictures or stenosis, are typically observed fluoroscopically during endoscopic retrograde cholangiography but do not allow a type-specific classification.

Digital single-operator cholangioscopy (dSOC) has enabled endoscopists to conduct a "bile duct endoscopy," elevating cholangioscopic visualization and guided biopsies (which have been proven superior to brush cytology, particularly in confirming malignancy).[4] Apart from malignancy, inflammation and scarification leading to narrowing are the primary culprits behind indeterminate strictures. The clinical scenario is further compounded by the fact that certain inflammatory conditions, such as primary sclerosing cholangitis (PSC), serve as risk factors for BTC.

In both blinded assessments (without additional clinical data) and unblinded evaluations, experienced examiners continue to encounter limited precision in arriving at a visual diagnosis based on dSOC images.[5] This lack of standardization results in wide interobserver variations in identifying relevant features and malignancies.[5,6]

Deep learning (DL) has swiftly permeated the realm of endoscopy, providing robust support for the detection (computer-aided detection [CADe]) and characterization (computer-aided diagnosis [CADx]) of pathologic observations based on endoscopic images and video recordings.[7-10] Unlike previous endeavors that required manual crafting of visual attributes for lesion detection and characterization,[11] convolutional neural networks (CNNs) learn their own parameters through exposure to labeled data. This empowers them to autonomously extract relevant features linked to the corresponding diagnosis.

Current research has been focused on using CNNs to detect neoplastic changes in bile ducts.[12-14] Differentiating these changes from inflammation remains a challenge, however. In clinical practice, physicians rely also on contextual information such as demographic characteristics, medical history, imaging, and laboratory results. Integrating relevant nonimage data seems promising, analogous to how physicians incorporate diverse information for better diagnostic results.

Addressing the clinical rationale to differentiate between normal-appearing bile ducts, inflammation, and malignancy in indeterminate biliary strictures, we developed and validated a CNN algorithm combining images and nonimage data, providing frame-by-frame characterizations and continuous real-time predictions in dSOC videos for enhanced diagnostic accuracy.

METHODS

Data collective and patients' characteristics

Based on analogous studies using CNN algorithms with image data from cholangioscopy and addressing similar scientific inquiries, we assumed a sample size of around 100 patients. Retrospective data from 122 patients collected at the University Hospital Regensburg (Regensburg, Germany) between January 2018 and June 2023 were included. Patients underwent dSOC investigations to assess uninvestigated biliary strictures. The data include still images and full video sequences obtained with SpyGlass DS II (Boston Scientific, Marlborough, Mass, USA).

The final diagnoses were confirmed either by histology of biopsy specimens (SpyBite, SpyBite Max; Boston Scientific) or surgical resection specimens. Regarding the comparable low diagnostic accuracy of bioptic specimens or brush cytology, all patients with inflammatory changes in the report were included in a subsequent clinical follow-up to set up the final diagnosis. The mean follow-up was 22.3 months (range, 1-88 months).

From 122 patients, 11 were excluded due to poor video quality, inconclusive visual assessment, or missing final histologic diagnosis. After exclusion, 177 dSOC investigations from 111 patients (37 female, 74 male; mean age 56.5 ± 19.0 years [range, 16-85 years]) were analyzed. Video recordings were available for 64 patients, and still images were available for 47 patients. The raw dataset consists of 270,343 frames extracted from videos and 806 individual still images.

Data collected between January 2018 and August 2022 were used for algorithm development and were split into a training set (n = 69 [approximately 76% of the development set]) and a validation set (n = 22 [approximately 24% of the development set]). Data collected after August 2022 until

TABLE 1. Clinical characteristics of dedicated training, validation, and test sets

	Training set (n = 69)	Validation set (n = 22)	Test set (n = 20)
Diagnosis			
PSC	25 (36.2)	6 (27.3)	2 (10.0)
SSC	10 (14.5)	1 (4.5)	0 (0)
ITBL	1 (1.4)	0 (0.0)	2 (10.0)
CBS/other inflammation	19 (27.5)	6 (27.3)	7 (35.0)
Malignant	14 (20.3)	8 (36.4)	8 (40.0)
Normal	0 (0)	1 (4.5)	1 (5.0)
Sex			
Female	21 (30.4)	10 (45.5)	6 (30.0)
Male	48 (69.6)	12 (54.5)	14 (70.0)
Age			
Female	53.0 ± 20.5 (16-82)	64.1 ± 11.7 (40-81)	51.2 ± 19.0 (24-68)
Male	55.5 ± 15.9 (16-85)	52.1 ± 15.4 (28-73)	63.3 ± 16.4 (25-81)
Age group			
<30 y	6 (8.7)	1 (4.5)	2 (10.0)
30-49 y	18 (26.1)	4 (18.2)	2 (10.0)
50-59 y	12 (17.4)	5 (22.7)	4 (20.0)
≥60 y	33 (47.8)	12 (54.5)	12 (60.0)

Values are n (%) or mean ± standard deviation. The absolute number of patients and the relative percentage is reported per diagnosis, sex, and age category.
PSC, Primary sclerosing cholangitis; *SSC*, secondary sclerosing cholangitis; *ITBL*, ischemic-type biliary lesion; *CBS*, common bile duct stones.

June 2023 served as a dedicated test set (n = 20). Patient characteristics and distribution in the particular groups (including number of frames) are displayed in Table 1 and Supplementary Figure 1 (available online at www.giejournal.org).

In addition, we considered sets of clinical metadata that were available at the time of examination to test if they affect performance of the CNN algorithm. For incorporation of this metadata, 2 different approaches were analyzed: the multiplication-based method and the concatenation-based incorporation. These sets of metadata included patient demographic characteristics, laboratory measurements, and characteristics from fluoroscopy and patients' history.

Patient demographic characteristics consisted of patient age (grouped as <30 years, 30-49 years, 50-59 years, and ≥60 years) and sex (male/female). Because they are routinely available at the time of investigation, 5 laboratory values were included in the analysis: bilirubin, alkaline phosphatase, gamma-glutamyltransferase, alanine transaminase, and aspartate aminotransferase.

Characteristics from fluoroscopy comprised the length of stenosis (short or long), localization and distribution of stenosis (unifocal or multifocal), previous stent placement or dilation therapy, duration of the medical history of bile duct changes (<3 months vs ≥3 months), and previous diagnosis of PSC or BTC that had previously been resected (Table 2).

The study was conducted and the results are reported according to the proposed protocol and guidelines for model studies based on artificial intelligence (TRIPOD [Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis]-AI).[15] The study was approved by the local ethics committee (No. 23-3229-104) and was performed according to the updated guidelines of Good Clinical Practice and the updated Declaration of Helsinki.

Data processing

Raw video and image files were pseudonymized before further processing. All videos were resampled from 25 to 10 frames per second. Videos were annotated on a frame-by-frame basis by an experienced endoscopist (A.K.) and labeled as either malignant or inflamed (acute inflammation or scarring tissue). The region of interest was cropped from each frame by identifying the largest area enclosed by a solid contour line corresponding to the recording window. Labeling was performed for malignancy or inflammation. Diffuse or blurry regions within each frame were excluded and not enclosed by the counter lines. Frames not labeled for "malignancy" or "inflammation" were assumed to be "normal." Each annotated video and image was reviewed for accuracy and consistency (A.K., P.D., J.Z., I.Z.-J., K.W.).

Frames were excluded if they did not show any part of the biliary duct system (eg, recordings inside the "mother" endoscope, pancreatic duct) or if they were corrupted due to technical issues.

For the training and validation sets, each patient's data were randomly subsampled to a maximum of 200 frames per class. This left 6709 frames (of 69 patients) for training

TABLE 2. Clinical metadata that were incorporated in the CNN algorithm by multiplication- and concatenation-based injection and analyzed for its impact on CNN performance

Clinical metadata	Quantitative value	One-hot encoding
Laboratory values		
Bilirubin	mg/dL	
Alkaline phosphatase	U/L	
Gamma-glutamyl transferase	U/L	
Alanine transaminase	U/L	
Aspartate aminotransferase	U/L	
Age		
<30 y		0
30-39 y		1
40-49 y		2
50-59 y		3
>60 y		4
Sex		
Female		0
Male		1
Results of fluoroscopy		
Length of stenosis		
Short		0
Long (>5 mm)		1
None		2
Location of stenosis		
Intrahepatic		0
Hilar		1
Common bile duct		2
Multiple loci		3
Anastomosis		4
None		5
Distribution of stenosis		
Singular		0
Multifocal		1
None		2
Biliary stent therapy		
Previous stent therapy		
No		0
Yes		1
Duration of stent therapy		
<3 mo		0
>3 mo		1
No stent therapy		2
Biliary dilation therapy		
No		0
Yes		1
PSC in patients' history		
No		0
Yes		1
Resected BTC in patients' history		
No		0
Yes		1
Medical history of bile duct pathology		
<3 mo		0
>3 mo		1

Laboratory results are given in quantitative values whereas the additional metadata from patients' medical history and the results of fluoroscopy were incorporated by using one-hot vector encoding.

CNN, Convolutional neural network; *PSC*, primary sclerosing cholangitis; *BTC*, biliary tract cancer.

and 2734 frames (of 22 patients) for validation. The test set comprised 20 previously unencountered patients examined between August 2022 and June 2023. One complete, representative video sequence of approximately 30 seconds was selected per patient, amounting to 5715 frames. There was no overlap of patients between the training, validation, and test sets.

Of patients with benign findings, random assignment with a distribution between the training set and the validation set of approximately 80% (training set) and 20% (validation set) was enforced. To ensure meaningful validation on the fewer amount of data available of patients with malignancy, 70% of patients within that group were assigned to the training set and 30% to the validation set.

The images were standardized by histogram equalization, per-channel normalization using the mean and standard deviation of the ImageNet dataset, and resizing to 224 × 224 pixels. During training, a combination of the following augmentation methods with random parameters was applied to each frame: horizontal and vertical flipping, rotation, changes in brightness and contrast, RGB (red, green, blue color model) shifting, pixel dropout, sharpening, and grayscale conversion.[7,16-18]

Patient age and sex were translated into one-hot vectors encoding their respective categories. Laboratory values were z-normalized, and missing values were replaced with their mean. Other metadata were encoded nominally as one-hot vectors.

Model development and integration of metadata

ResNet18 pretrained on the ImageNet dataset was used as model backbone.[19] A custom final classification unit was

added consisting of a fully connected layer, a batch normalization layer with a dropout rate of .3, and another fully connected layer mapping to the desired number of classes. We considered a binary classifier distinguishing between malignant/nonmalignant (CADe) as well as a multi-class classifier distinguishing between malignant/inflammation/normal (CADx).

Patient demographic characteristics and laboratory values were selected for metadata injection, as those are routinely available and accessible during every examination. Other metadata that were tested included the fluoroscopic characterization and distribution of stenosis (see Data collective and patients' characteristics) but did not affect the performance of our CNN model. Finally, the aforementioned set of metadata was selected, as it has been shown to further improve the CNN performance.

Two CNN different models were tested to incorporate the metadata. For the concatenation-based metadata injection model, the metadata vector was appended to the image feature vector obtained from the last convolutional layer, and this combined vector was passed to the final classification unit (Fig. 1A).[20-23] For the multiplication-based model, the metadata were preprocessed by a separate fully connected network unit and then multiplied with the feature vector, with the former effectively acting as a weight on the image feature map (Fig. 1B).[24]

All experiments were implemented with PyTorch version 1.13.1 (The PyTorch Foundation, copyright: The Linux Foundation, San Francisco, Calif, USA). Model hyperparameters were optimized using grid search. We trained each model for 50 epochs with a batch size of 64 using the Adam optimizer with a constant learning rate of 10^{-6}, $\beta_1 = .9$, and $\beta_2 = .999$. Class weights were applied to the cross entropy loss proportional to the number of samples per class. The average inference time for a single frame with metadata was approximately 140 milliseconds (7 frames per second) using an Apple M1 Pro processing unit (Apple Inc, Cupertino, Calif, USA), with data preprocessing included.

Model performance metrics

Model performance on the validation set was monitored during model training. For multi-class models, we computed the macro-average of one-versus-one per class area under the receiver-operating characteristic curves (AUCs) as a global metric that is insensitive to class imbalance and one-versus-rest AUCs per class for class-specific evaluation.

Finally, the CNN models were evaluated on the independent test set of 20 patients.

For each frame, the model's final prediction was the diagnosis with the highest assigned probability. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score (harmonic mean between sensitivity and PPV) in reference to malignancy on the separate test set are reported. Confidence intervals were calculated by using the Wilson score interval with $\alpha = .05$.

Uniform Manifold Approximation and Projection (UMAP) was used to map the high-dimensional embedding vector obtained from the model's final convolutional layer for better explainability in a 2-dimensional representation.[25] In addition, the Gradient-weighted Class Activation Mapping (Grad-CAM++) method was applied, which uses the gradients of a CNN to produce a class-activation map highlighting the important regions of an input image that contribute most to the model's prediction.[26]

Descriptive statistics include means, standard error of the mean, and range where indicated.

Temporal smoothing

To address the video's temporal nature, context windows of different sizes were studied by averaging per-frame probabilities over the last k frames within a video, where k ranges stepwise from 0 to 100 (0, 5, 10, 20, 30, 50, 100) (Supplementary Fig. 2, available online at www.giejournal.org).

Sensitivity and specificity for malignancy simultaneously increased as the context window increased up to 30 frames. Coincidentally, the model lost predictive ability for inflammatory lesions and normal mucosa. Joint improvements were observed for all classes at a context window of 5 frames. The performance for malignancy detection at $k = 5$ was .833 (.809-.854) for sensitivity, .782 (.770-.794) for specificity, .462 (.440-.484) for PPV, and .954 (.947-.960) for NPV.

The prediction scores generated by the multi-class concatenation-based model for a complete test video, including malignant changes in a patient with PSC not encountered during the training phase, were illustrated. Figure 2A displays the initial predicted scores, and Figure 2B shows the scores after smoothing with $k = 5$. In both cases, the model accurately and confidently identified both the malignant lesion and the inflammatory aspect within the video.

Case-based prediction of malignancy

Frame-wise predictions allow a continuous, real-time assessment and localization of lesions in individual dSOC video frames and images (Video 1, available online at www.giejournal.org). However, we highlight the importance of a comprehensive evaluation of the CNN algorithm as a diagnostic support system for a complete patient examination, which requires a risk measure on a case-based level.[13] To derive case-based predictions from the frame-based algorithm, we exploited that the model's confidence was greater for correctly identified samples ($.752 \pm .167$) compared with incorrectly predicted samples ($.613 \pm .148$).

We defined a malignancy score for each individual case by calculating the average of the model's output probabilities for all frames classified as malignant. A frame was considered malignant if its probability of malignancy was the highest among all 3 classes. In cases in which no frame was classified as malignant, the malignancy score was set to zero. The malignancy score was used to create a binary

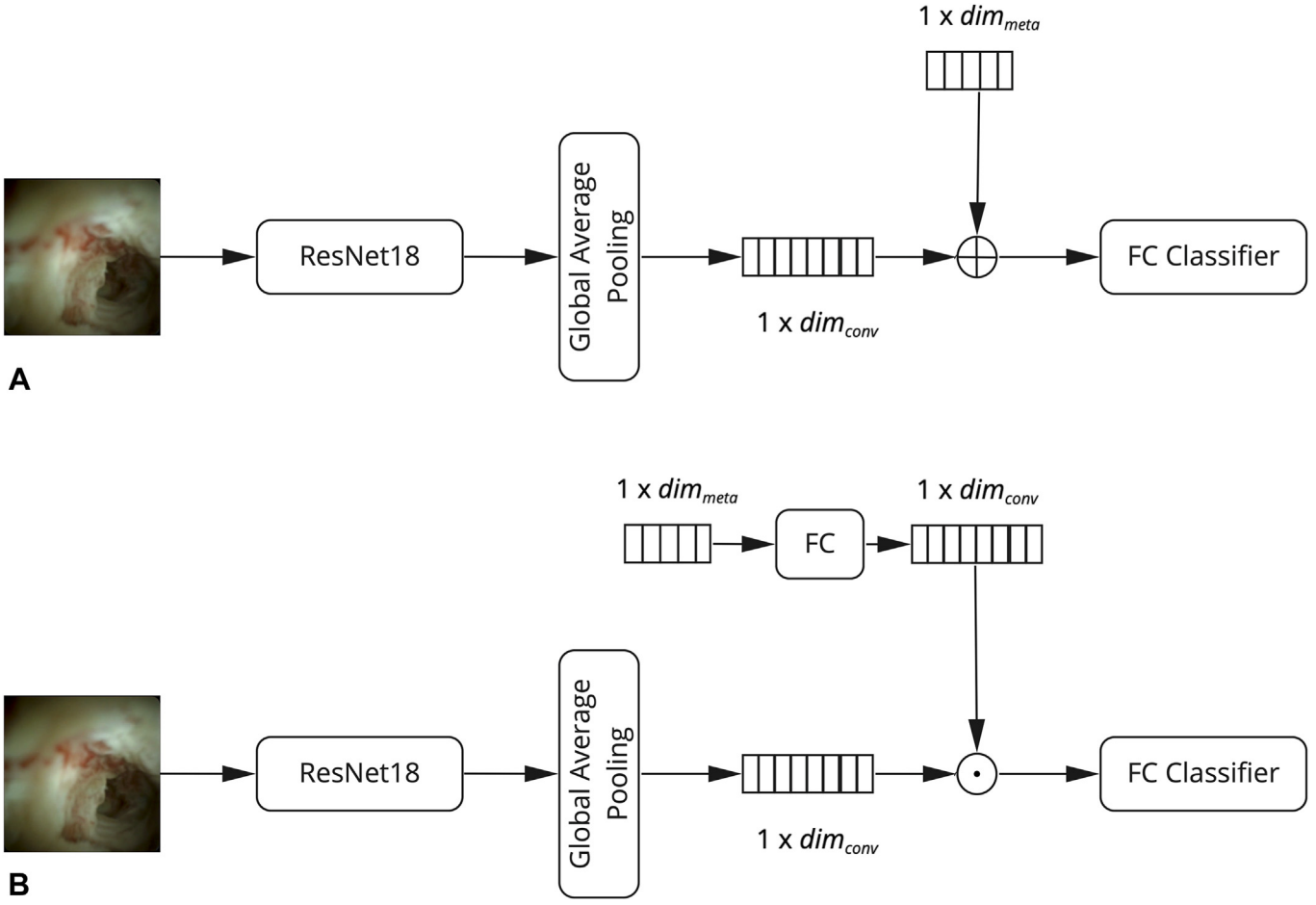


Figure 1. Inclusion of clinical metadata by concatenation- and multiplication-based metadata injection. **A,** The concatenation-based metadata injection fuses the feature map and the metadata vector before passing them to the final classification layers. **B,** In the multiplication-based approach, the processed metadata acts as a weight for the feature map. Visualization based on Li et al.[24] *FC,* Fully connected layer.

classification system that determines malignancy based on whether the average malignancy score surpasses a specific threshold. We determined the optimal threshold to be .61 by maximizing Youden's J statistic on the malignancy scores of cases in the validation dataset.

RESULTS

Frame-based evaluation of the binary model for the detection of malignancy (CADe) and a multi-class model for the differentiation between inflammation and malignancy (CADx)

The models improved continuously during training for both CADe (malignant/nonmalignant) and CADx (malignant/inflammation/normal) as depicted in Figure 3. The binary model for detection (CADe) achieved a validation AUC of .815 (without metadata), .840 (concatenation-based metadata injection), and .857 (multiplication-based metadata injection).

In the test set, we achieved an AUC for CADe of .845, .873, and .866, respectively (Table 3). The multi-class model for characterization and diagnosis (CADx) indicated converging class-averaged validation AUCs of .850, .865, .860, and AUCs in the test set of .864, .871, and .873 (Table 4). The best CADe model (with multiplication-based injection) showed a sensitivity of .890 (.869-.907), a specificity of .681 (.668-.694), a PPV of .385 (.366-.404), and an NPV of .965 (.958-.970). The best CADx model (concatenation-based injection) achieved a sensitivity of .809 (.784-.832), specificity of .773 (.761-.785), PPV of .450 (.423-.467), and NPV of .946 (.940-.954). It accurately identified 80.9% of frames with malignant lesions, 71.9% of inflammatory lesions, and 57.6% of normal tissue.

Multimodal CNN algorithm by incorporating clinical metadata for CADe and CADx in cholangioscopic videos

Metadata injection with demographic information and laboratory measurements improved the sensitivity compared with both image-only models (CADe, .878 [.856-.896]/.890 [.869-.907] vs .833 [.809-.854]; CADx, .809 [.784-.832]/.793 [.768-.817] vs .734 [.707-.760]).

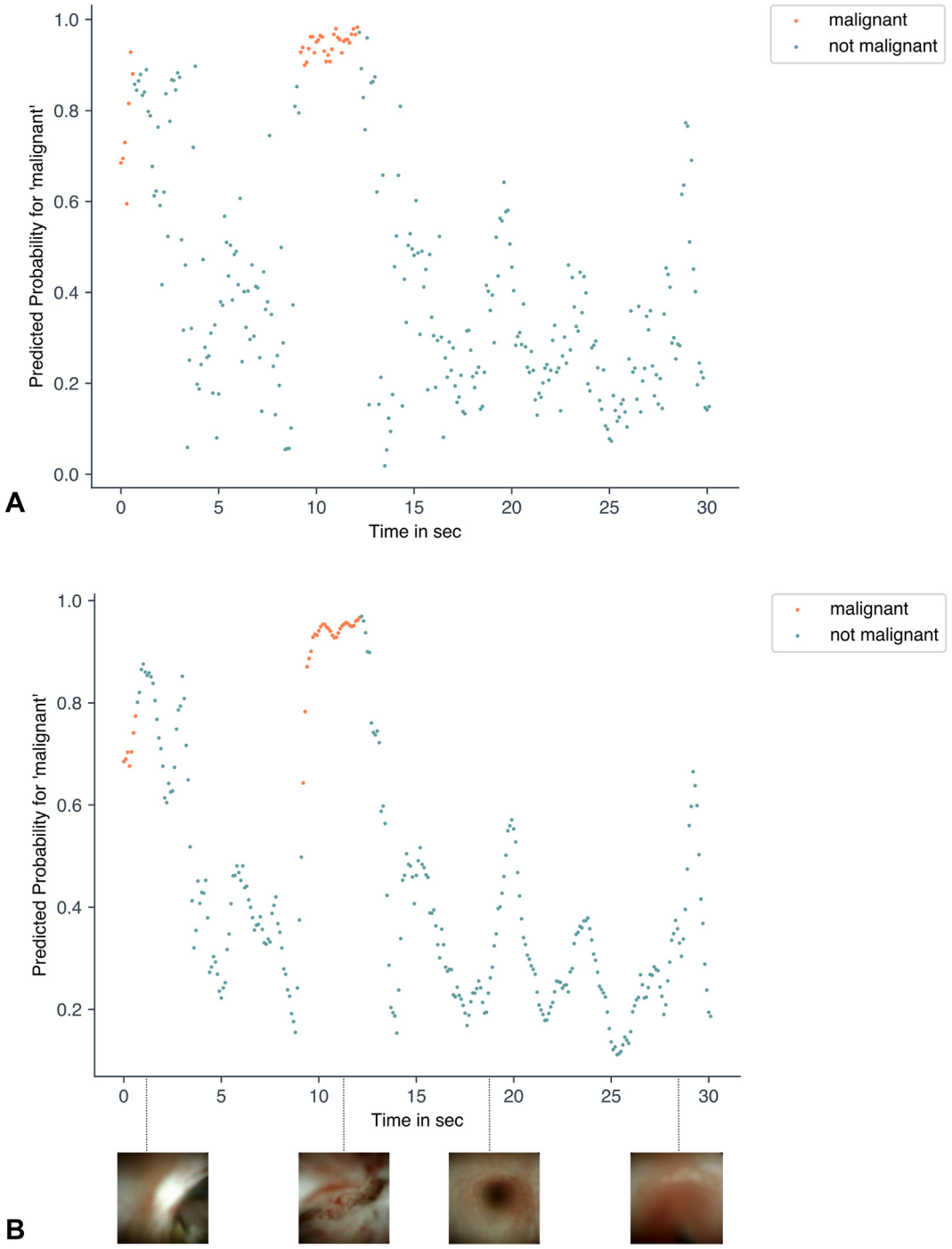


Figure 2. Probability scores for malignancy of convolutional neural network–based diagnosis on a full video sequence. The original **(A)** and smoothed ($k = 5$) **(B)** probability scores for malignancy on an example video were compared with actual malignancy highlighted in *red*. The model correctly identifies the malignant lesion and inflammatory aspect with high confidence. Smoothing contributes to the consistency of the prediction. *FC,* Fully connected layer.

Including additional metadata from fluoroscopy or patients' history significantly decreased the concatenation-based CADx model's performance to a sensitivity of .685 (.656-.712; analysis not shown). We therefore proceed with the combination of patient demographic characteristics and laboratory values and further investigated the concatenation-based CADx model.

The corresponding video (Video 1) exemplifies the real-time utilization of our CNN algorithm (CADe and CADx) on the entire video sequence on a frame-based level. The

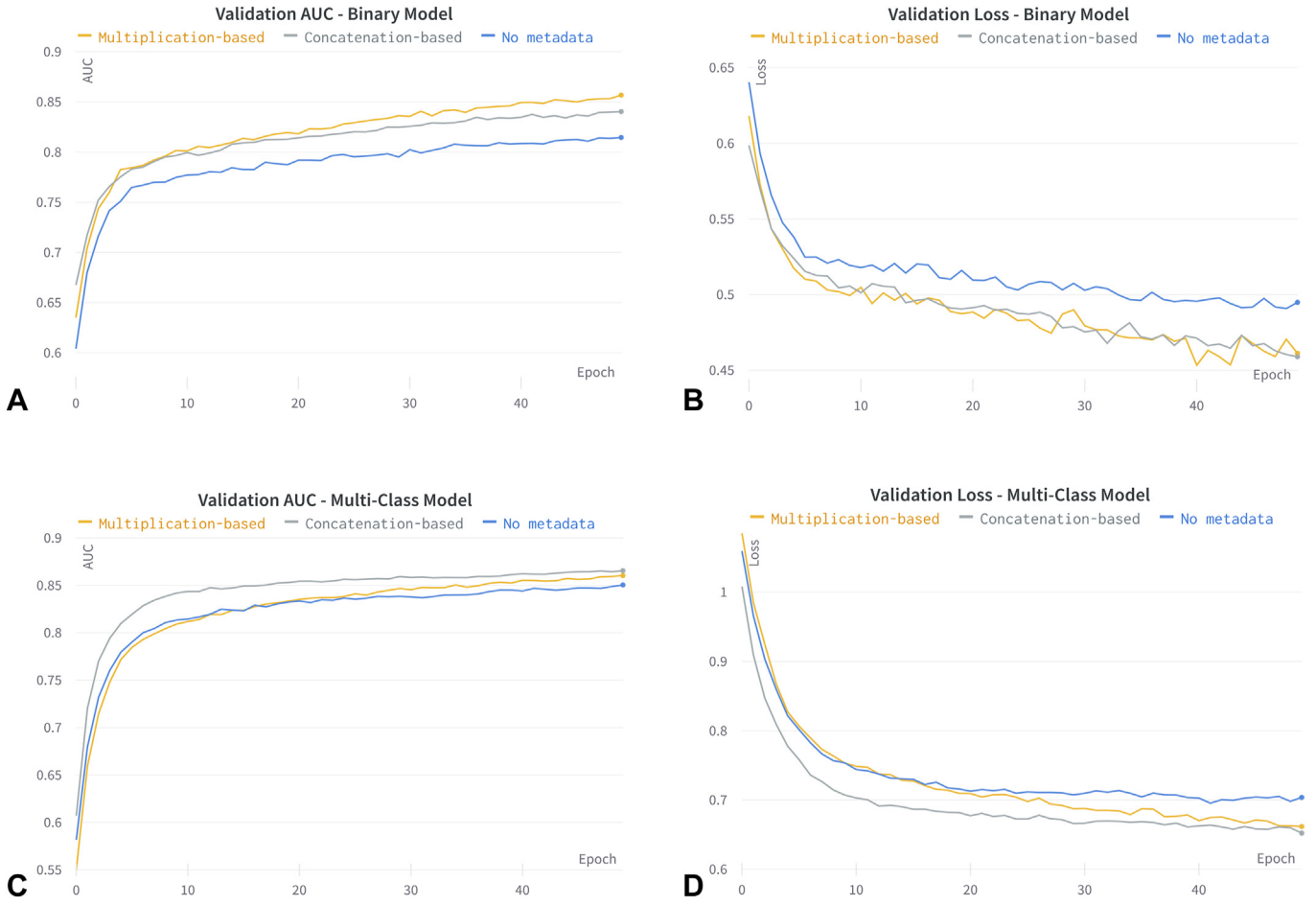


Figure 3. Training progress of the binary model for detection of malignancy (computer-aided detection) and multi-class model to differentiate malignant changes from malignant, inflammatory, and normal bile duct mucosa (computer-aided diagnosis). The area under the receiver-operating characteristic curve (AUC) increased consistently while the training loss decreased on the validation set for the binary models for detection **(A, B)** and multi-class models for characterization **(C, D)** with and without metadata injection. The model converges with a greater AUC when trained with either metadata injection approach. For the multi-class model, the one-versus-one macro-averaged AUC over all 3 classes (malignant, inflammation, and normal) is displayed.

projected diagnosis, accompanied by confidence scores, is showcased in the lower left quadrant of the video throughout each frame of the complete sequence. Simultaneously, the degree of confidence pertaining to any diagnosis (malignant, inflammation, normal) is presented. As an alert mechanism, the color of the predicted diagnosis transitions to red in the event of high malignancy risk.

Case-based evaluation

In addition to the application of the CNN model to individual frames, the algorithm was evaluated on the case level using the determined threshold of .61 for malignancy on averaged frame scores. From 20 test cases, 6 of 8 malignant cases were correctly classified (averaged prediction confidence scores, .656 to .899). All 12 benign cases were correctly identified as benign (scores, .439-.590).

Model interpretability and qualitative validation

We projected high-dimensional model embeddings to 2-dimensional space for visualization using UMAP. Figure 4A shows the low-dimensional frame embeddings per their predicted diagnosis, and Figure 4B shows UMAP model embeddings with model confidence scores for the predicted diagnosis. It visualizes the frame embeddings and their corresponding predicted probabilities. The confidence is higher in areas in which the 3 diagnostic classes do not overlap. Furthermore, the mean confidence for correctly predicted frames (.752 ± .167) is higher than that for incorrectly predicted frames (.613 ± .148).

Regions of the image that are important to the CNN model are highlighted with class-activation heat maps using Grad-CAM++ on correctly and incorrectly classified frames from the test set (Fig. 5). The model consistently attends to relevant areas containing the visual characteristics associated with the respective diagnosis.

DISCUSSION

We developed a multimodal CNN model designed not only to detect BTC within dSOC videos but also to effectively

TABLE 3. Test performance of the binary model for detection of malignancy (CADe) without clinical metadata, with concatenation-based and with multiplication-based metadata injection

Clinical metadata	AUC	Sensitivity	Specificity	PPV	NPV
No metadata injection	.845	.833 (.809-.854)	**.710** (.697-.722)	**.392** (.372-.412)	.950 (.942-.957)
Concatenation based	**.873**	.878 (.856-.896)	.687 (.673-.700)	.386 (.367-.406)	.961 (.954-.968)
Multiplication based	.866	**.890** (.869-.907)	.681 (.668-.694)	.385 (.366-.404)	**.965** (.958-.970)

Both metadata models outperformed the image-only model with respect to sensitivity. Best performance metrics are highlighted (bold).
CADe, Computer-aided detection; *AUC*, area under the receiver-operating characteristic curve; PPV, positive predictive value; *NPV*, negative predictive value.

TABLE 4. Test performance of the multi-class model for differentiation (CADx) between inflammation, malignancy, and normal mucosa without clinical metadata, with concatenation-based, and multiplication-based metadata injection

Clinical metadata	AUC average, malignant, inflammation, normal	Sensitivity	Specificity	PPV	NPV
No metadata injection	.864, .853, .874, .872	.734 (.707-.760)	**.817** (.805-.827)	**.473** (.450-.500)	.932 (.924-.940)
Concatenation-based	.871, .865, .884, .861	**.809** (.784-.832)	.773 (.761-.785)	.450 (.423-.467)	**.946** (.940-.954)
Multiplication-based	**.873,** .870, .886, .857	.793 (.768-.817)	.793 (.781-.804)	.462 (.440-.486)	.945 (.937-.951)

Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) are reported with respect to malignancy. Again, the metadata models outperformed the image-only model with respect to sensitivity. Best performance metrics are highlighted (bold).
CADx, Computer-aided diagnosis; *AUC*, area under the receiver-operating characteristic curve.

distinguish BTC from inflammatory biliary mucosa (CADx). This model can be applied frame-wise and in real-time to entire video sequences comprising both high- and low-quality imaging content. We developed the CNN model on a challenging patient cohort with a high prevalence of patients with PSC (training set, 36.2%; validation set, 27.3%) and other inflammatory conditions (training set, 43.4%; validation set, 32.3%). This study further highlights the technical potential to integrate clinical metadata to improve a video-based CNN algorithm. This is in line with observations that clinicians' diagnoses are influenced by supplementary clinical information.[5]

We analyzed features that were available at the time of every cholangioscopic investigation, including 5 laboratory values, age, sex, and information about the fluoroscopic morphology and patients' history. The combination of age, sex, and laboratory values performed best in the CNN algorithm.

Traditional clinical information and metadata, which have long guided physicians, did not enhance the performance of our CNNs as we expected. This outcome is partly attributable to how CNN architecture connects data points and processes them into a unified vector for classification. These data points may not align perfectly with the nuanced medical characteristics relied upon by human intelligence in clinical practice. We translated the information extracted from fluoroscopy into valuable medical insights, including parameters such as stenosis length, distribution patterns, and the presence of unifocal or multifocal lesions before incorporation into the CNN algorithm. In the near future, a promising task is integrating diverse sources of image data such as cholangioscopic and fluoroscopy-originated images. Another plausible explanation for the detrimental effect of incorporating fluoroscopy results or patients' history may be due to the increased complexity that has been included. This increased complexity may exceed the model's ability to learn effectively, primarily due to the limited variability in metadata feature combinations within the available dataset. As a proof-of-concept study for the incorporation of clinical metadata into primary image-based information, our results nonetheless direct to the future combination of different sources including not only metadata but also different image datasets (cholangioscopy, magnetic resonance imaging, fluoroscopy).

On complete videos from 20 patients in the independent test set, the best multi-class CADx model with concatenation-based metadata injection achieved a class-average AUC of .871, per-frame sensitivity of .809 (.784-.832), specificity of .773 (.761-.785), PPV of .450 (0.423-.467), and NPV of .946 (.940-.954).

The low PPV of our algorithm can be attributed to the relatively larger proportion of frames depicting normal tissue encountered in the raw video sequences. These frames exhibit considerable variability, including low-quality frames that may hinder accurate classification. Although it is crucial to account for low-quality data in the evaluation, the impact of these data on clinical practice is limited, as such data are seldom consulted for diagnostic decisions. As evidenced by the case-based evaluation and the incorporation of temporal smoothing, context is crucial for deploying the algorithm effectively. Analyzing differences between CADe and CADx, the CADe approach exhibited higher sensitivity for malignancy, and its specificity was generally lower compared with CADx. This discrepancy can be anticipated due to the narrower focus of the CADe model, which only requires distinguishing between

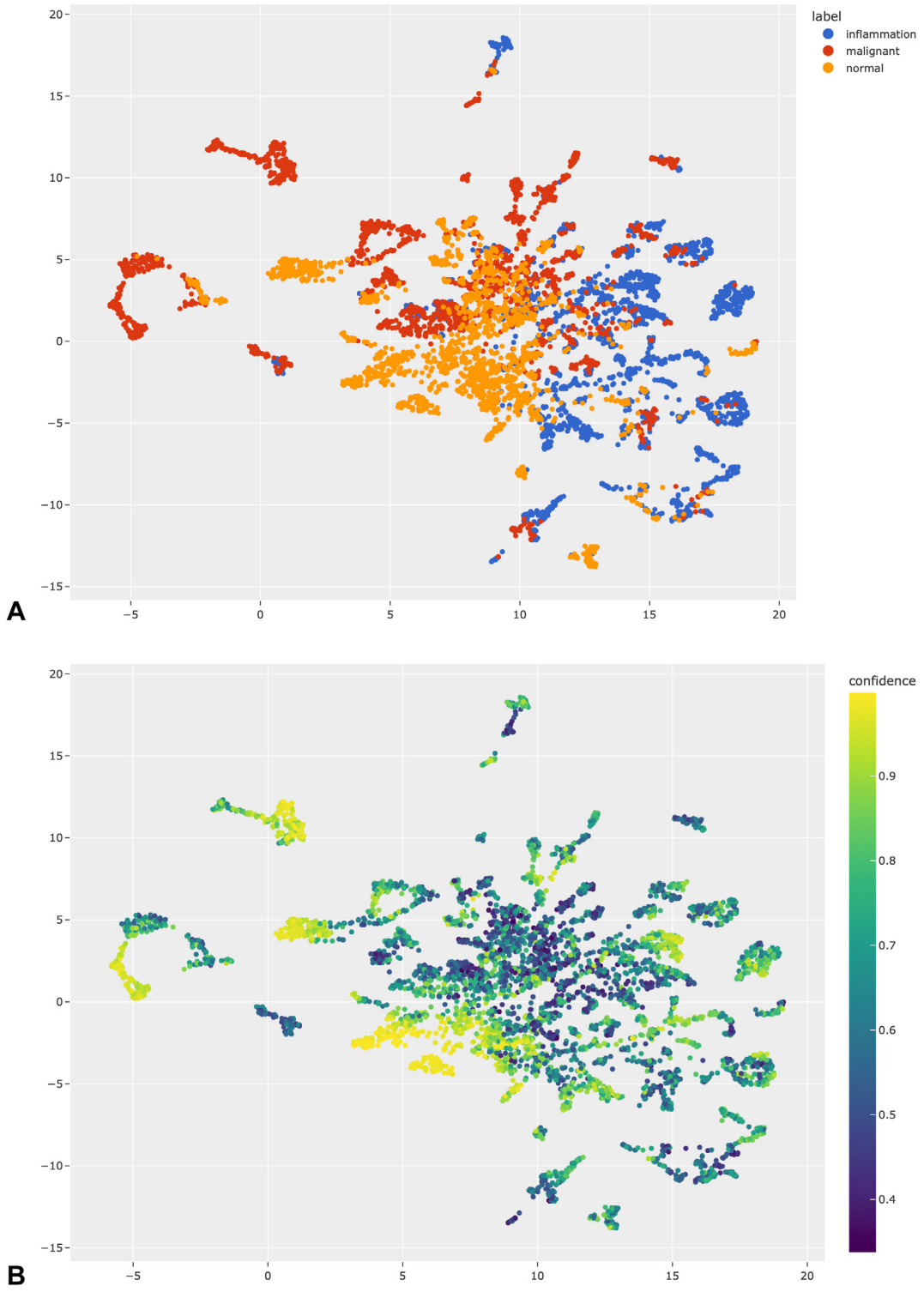


Figure 4. Uniform Manifold Approximation and Projection results of convolutional neural network–embedded frames according to their predicted diagnosis **(A)** and model confidence scores **(B)**. Uniform Manifold Approximation and Projection results on convolutional embeddings with metadata of the concatenation-based multi-class model colored by their predicted diagnosis **(panel A)** outline decision boundaries learned by the model. Confidence scores for the predicted class **(panel B)** are higher in areas where data points of a specific class accumulate.

2 classes, thereby prioritizing recognition of malignancy at the expense of specificity.

In contrast to other DL models that have been proposed exclusively for the purpose to identify malignancy,[12,13] we are the first, to the best of our knowledge, to address the differentiation between malignancy and inflammation as CADx in a real-time scenario of a full video stream.

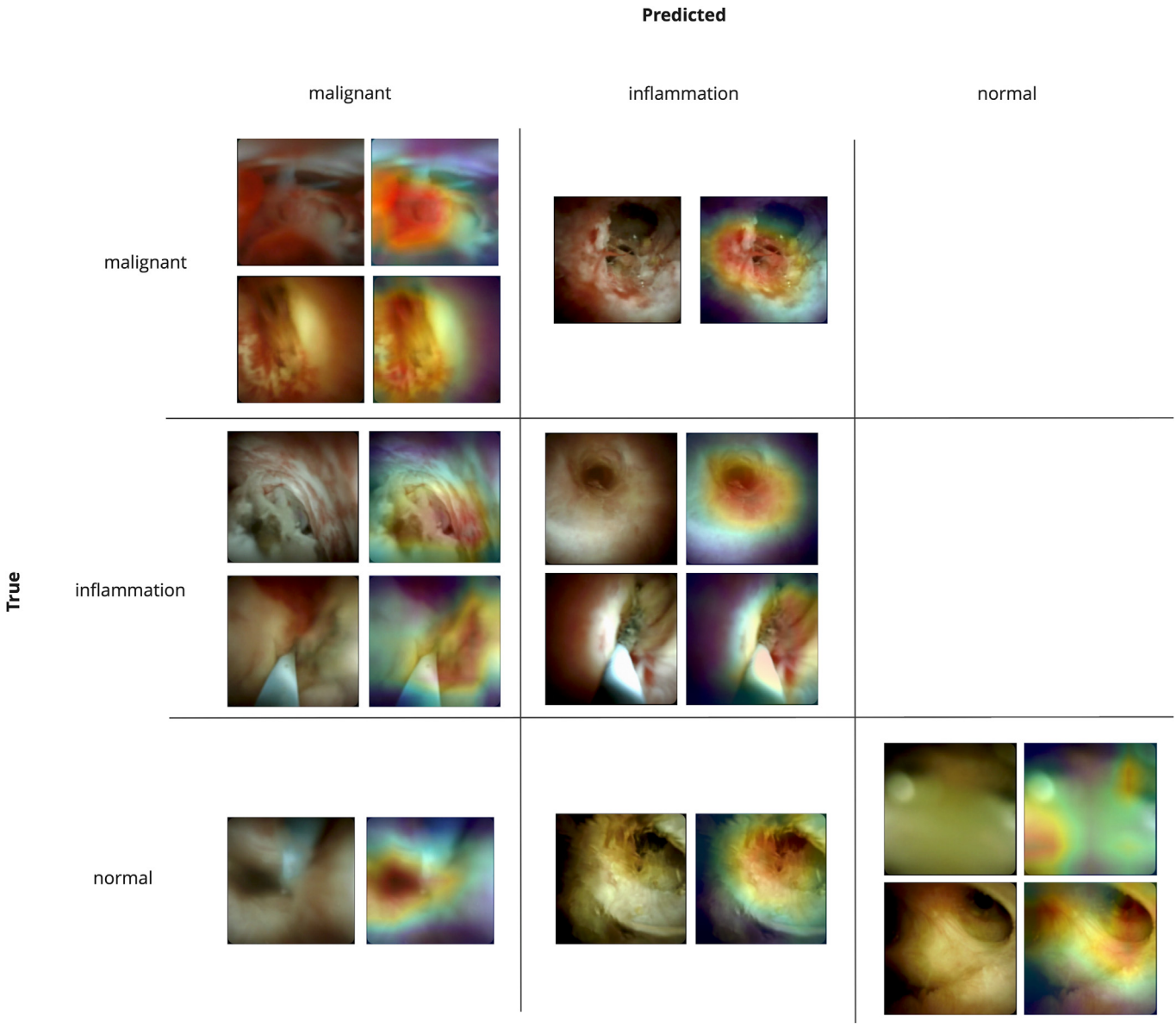


Figure 5. Heat activation maps of predicted frames and diagnosis. The class-activation maps generated with Gradient-weighted Class Activation Mapping (Grad-CAM++) highlight regions of importance to the model. Examples of correctly and incorrectly predicted frames across the 3 diagnostic classes are shown. Both malignant and inflammatory lesions are rarely misclassified as normal.

Previously, the identification of malignancy in dSOC images using DL has been based on individual indicators, including dilated and tortuous vessels,[27,28] papillary projections,[28,29] and ulceration,[28] or across multiple or an undefined set of the mentioned features.[12-14] These DL models have been shown to detect single features associated with malignancy with an AUC of up to 1.0.[27,29] Malignancy was identified with sensitivity/specificity of .905/.682[13] and 0.933/0.882,[12] and outperformed brush cytology, forceps biopsy,[12] and human visual evaluations.[13] In contrast to a DL algorithm trained on the recognition of distinct features of malignancy, our CNN algorithm operates as a self-learning system, eliminating the need for specific feature engineering. We visually examined the significance of image regions contributing to the neural network's classification outcomes using an explainability algorithm (Grad-CAM++). This qualitative analysis confirms that the neural network indeed focuses on image features relevant to human observers.

Comparison of different CNN algorithms requires a detailed discussion of data selection, data processing, algorithm development, and evaluation. Earlier studies mainly focused on processing individual high-quality images and still frames, potentially leading to limited applicability when dealing with unprocessed video streams encountered in real-time dSOC procedures, and disregarding the inherent temporal characteristic of video data. The per-

frame sensitivity of DL models trained on still images has been shown to decline when evaluated on full videos.[17,30,31] This suggests an optimistic estimation of model performance that cannot necessarily be translated to full recordings in clinical practice.

Addressing clinical decision-making, we tested the application of our CNN in a case-based scenario. Six of 8 malignant cases and all 12 benign cases were correctly identified.

Evaluating our multimodal CADx model on all frames from complete videos is a major strength. For CNN training and validation, low-quality or low-informative data are present in the data due to random sampling of frames. We thus expect a reasonable generalization of results toward raw real-world videos. In addition, patients with PSC were included. This is in contrast to studies in which these patients have been explicitly omitted.[12,13] Averaging adjacent per-frame predictions has been shown to boost the identification of malignancy on the video level.[12] Our results endorse this conjecture, showing an improvement in the concatenation-based CADx model when averaging per-frame predictions over a context window of 5 frames.

Our work is limited by the retrospective character of data. Although these results suggest the superiority of a metadata-enhanced model over an image-only approach, further verification is required. Access to diverse and multicentric data is therefore essential to avoid model overfitting, especially regarding metadata integration, and to enable development of a robust CADx algorithm working under real-time conditions during dSOC investigations.

CONCLUSION

To our knowledge, this is the first CNN-based algorithm to integrate clinical metadata into its architecture, enabling not only the detection but also the differentiation between malignancy and inflammation in cholangioscopy.

This algorithm shows promise for integration into cholangioscopy as a real-time clinical decision support system.

DISCLOSURE

The following authors disclosed financial relationships: A. Kandulski: Speaker and advisory committee for Roche Pharma AG, Eisai GmbH, AbbVie Germany AG, Janssen-Cilag GmbH, MSD Sharp & Dohme GmbH, Boston Scientific, Fujifilm Germany, Micro-Tech Germany, and Bayer Pharma AG Germany. K. Weigand: scientific presentations and scientific advisory activities: AstraZeneca Germany, Bristol-Myers Squibb, Abbvie Germany AG, Janssen-Cilag GmbH, MSD Sharp & Dohme GmbH, Dr. Falk Pharma AG, Roche Pharma AG, Gilead Sciences GmbH, Orifarm GmbH, Olympus Germany. M. Müller: scientific presentations and scientific advisory activities: AstraZeneca Germany, Dr. Falk Pharma AG, Gilead Sciences GmbH, Ipsen Pharma GmbH. This research project was supported by the Federal Government of Bavaria, Germany (Bayerische Forschungsstiftung, AZ-1534-21). The funders had no impact on either study design, data collection, analysis, and interpretation of data, drafting of the manuscript, or the decision for publication.

ACKNOWLEDGMENTS

The authors thank Jochen Weigt (Department of Gastroenterology, Hepatology and Infectious Diseases, Otto-von-Guericke University Magdeburg, Magdeburg, Germany) for support in data management. They also thank Mark Ellrichmann (Department of Interdisciplinary Endoscopy, University Medical Center Schleswig-Holstein, Campus Kiel, Kiel, Germany) and all collaborators of the German Cholangioscopy Group for helpful scientific discussions.

REFERENCES

1. Ogawa T, Kanno Y, Koshita S, et al. Cholangioscopy- versus fluoroscopy-guided transpapillary mapping biopsy for preoperative evaluation of extrahepatic cholangiocarcinoma: a prospective randomized crossover study. Surg Endosc 2021;35:6481-8.
2. Valle JW, Kelley RK, Nervi B, et al. Biliary tract cancer. Lancet 2021;397:428-44.
3. Banales JM, Marin JJG, Lamarca A, et al. Cholangiocarcinoma 2020: the next horizon in mechanisms and management. Nat Rev Gastroenterol Hepatol 2020;17:557-88.
4. 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;91:1105-13.
5. Stassen PMC, Goodchild G, de Jonge PJF, et al. Diagnostic accuracy and interobserver agreement of digital single-operator cholangioscopy for indeterminate biliary strictures. Gastrointest Endosc 2021;94:1059-68.
6. Subhash A, Abadir A, Iskander JM, et al. applications, limitations, and expansion of cholangioscopy in clinical practice. Gastroenterol Hepatol 2021;17:110-20.
7. Byrne MF, Chapados N, Soudan F, et al. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. Gut 2019;68:94-100.
8. Zhang R, Zheng Y, Mak TWC, et al. Automatic detection and classification of colorectal polyps by transferring low-level CNN Features from nonmedical domain. IEEE J Biomed Health Inform 2017;21:41-7.
9. Hirasawa T, Aoyama K, Tanimoto T, et al. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. Gastric Cancer 2018;21:653-60.
10. Wang P, Berzin TM, Brown JRG, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut 2019;68:1813-9.
11. Kominami Y, Yoshida S, Tanaka S, et al. Computer-aided diagnosis of colorectal polyp histology by using a real-time image recognition system and narrow-band imaging magnifying colonoscopy. Gastrointest Endosc 2016;83:643-9.
12. Marya NB, Powers PD, Petersen BT, et al. Identification of patients with malignant biliary strictures using a cholangioscopy-based deep learning artificial intelligence (with video). Gastrointest Endosc 2023;97:268-78.
13. Robles-Medranda C, Baquerizo-Burgos J, Alcívar-Vásquez J, et al. Artificial intelligence for diagnosing neoplasia on digital cholangioscopy:

development and multicentric validation of a convolutional neural network model. Endoscopy 2023;55:719-27.
14. 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;95:339-48.
15. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021;11:e048008.
16. Pacal I, Karaboga D. A robust real-time deep learning based automatic polyp detection system. Comput Biol Med 2021;134:104519.
17. Qadir HA, Balasingham I, Solhusvik J, et al. Improving automatic polyp detection using CNN by exploiting temporal dependency in colonoscopy video. IEEE J Biomed Health Inform 2020;24:180-93.
18. Jheng Y-C, Wang Y-P, Lin H-E, et al. A novel machine learning-based algorithm to identify and classify lesions and anatomical landmarks in colonoscopy images. Surg Endosc 2022;36:640-50.
19. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Available at: https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html. Accessed March 31, 2023.
20. Calderisi M, Galatolo G, Ceppa I, et al. Improve image classification tasks using simple convolutional architectures with processed metadata injection. In: 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE): IEEE; 2019. p. 223-30.
21. Ningrum DNA, Yuan S-P, Kung W-M, et al. Deep learning classifier with patient's metadata of dermoscopic images in malignant melanoma detection. J Multidiscip Healthc 2021;14:877-85.
22. Thomas SA. Combining image features and patient metadata to enhance transfer learning. Annu Int Conf IEEE Eng Med Biol Soc 2021;2021:2660-3.
23. Tang P, Yan X, Nan Y, et al. FusionM4Net: a multi-stage multi-modal learning algorithm for multi-label skin lesion classification. Med Image Anal 2022;76:102307.
24. Li W, Zhuang J, Wang R, et al. Fusing metadata and dermoscopy images for skin disease diagnosis. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI): IEEE; 2020. p. 1996-2000.
25. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv 2020;1802.03426v3.
26. Chattopadhyay A, Sarkar A, Howlader P, et al. Grad-CAM++: improved visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV): IEEE; 2018. p. 839-47.
27. 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:E262-8.
28. Ghandour B, Hsieh H-W, Akshintala V, et al. s1 machine learning for classification of indeterminate biliary strictures during cholangioscopy. Am J Gastroenterol 2021;116:S1.
29. Ribeiro T, Saraiva MM, Afonso J, et al. Automatic identification of papillary projections in Indeterminate biliary strictures using digital single-operator cholangioscopy. Clin Transl Gastroenterol 2021;12:e00418.
30. Wang P, Xiao X, Glissen Brown JR, et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nat Biomed Eng 2018;2:741-8.
31. Guo L, Xiao X, Wu C, et al. Real-time automated diagnosis of precancerous lesions and early esophageal squamous cell carcinoma using a deep learning model (with videos). Gastrointest Endosc 2020;91:41-51.

Abbreviations: AUC, area under the receiver-operating characteristic curve; BTC, biliary tract cancer; CADe, computer-aided detection; CADx, computer-aided diagnosis; CNN, convolutional neural network; dSOC, digital single-operator cholangioscopy; DL, deep learning; NPV, negative predictive value; PPV, positive predictive value; PSC, primary sclerosing cholangitis; UMAP, Uniform Manifold Approximation and Projection.

DIVERSITY, EQUITY, AND INCLUSION: We worked to ensure gender balance in the recruitment of human subjects. We worked to ensure ethnic or other types of diversity in the recruitment of human subjects. We worked to ensure that the language of the study questionnaires reflected inclusion. The author list of this paper includes contributors from the location where the research was conducted who participated in the data collection, design, analysis, and/or interpretation of the work.

De-identified data may be viewed and code can be made available upon request directed at the corresponding author.

This video can be viewed directly from the *GIE* website or by using the QR code and your mobile device. Download a free QR code scanner by searching "QR Scanner" in your mobile device's app store.

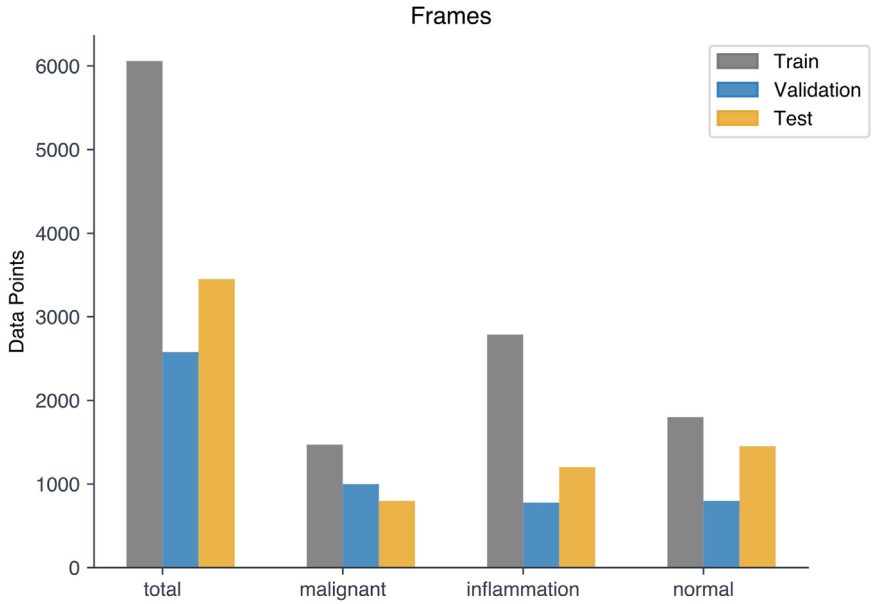
*Ms Ziegler and Dr Dobsch contributed equally to this article.

0016-5107
https://doi.org/10.1016/j.gie.2024.09.001

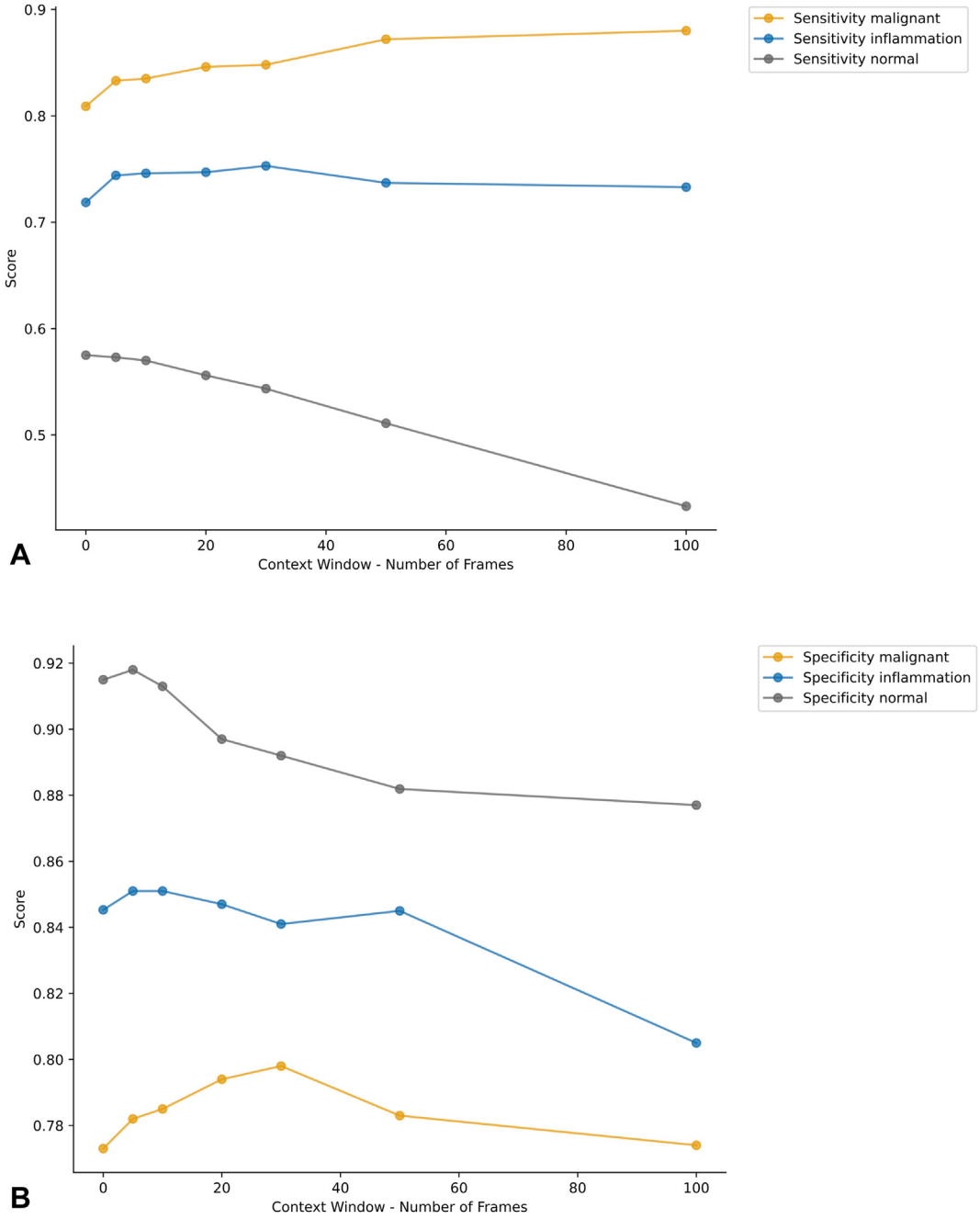
Received February 5, 2024. Accepted September 1, 2024.

Current affiliations: Unetiq GmbH, München, Germany (1), Department of Internal Medicine I, Gastroenterology, Hepatology, Endocrinology, Rheumatology and Infectious Diseases, University Hospital Regensburg, Regensburg, Germany (2), Department of Internal Medicine, Gastroenterology, Gastrointestinal Oncology and Diabetology, Gemeinschaftsklinikum Mittelrhein, Koblenz, Germany (3).

Reprint requests to: Arne Kandulski, MD, Department of Internal Medicine I, Gastroenterology, Hepatology, Endocrinology, Rheumatology and Infectious Diseases, University Hospital Regensburg, Franz-Josef-Strauß-Allee 11, 93053 Regensburg. E-mail: Arne.Kandulski@ukr.de.



Supplementary Figure 1. Data distribution between training, validation, and test sets. Video frames were subsampled randomly per patient to balance the number of frames per class (malignancy; inflammation; normal) in the training and validation sets, enforcing a distribution between the training set and the validation set of approximately 80%/20% of patients with only benign findings and 70%/30% of patients with malignancy. The test set includes complete videos of 20 patients.



Supplementary Figure 2. Smoothing of per-frame predictions on full video sequences. Temporal smoothing of different context windows affects the per-frame sensitivity **(A)** and specificity **(B)** with respect to each diagnostic class. Averaging predicted class probabilities over the last 5 frames improves sensitivity and specificity with respect to all diagnoses. For k $\leq$30, detection of malignancy improves, but differentiation of inflammatory lesions and normal mucosa is impeded.