




Article

An Optimal Artificial Intelligence System for Real-Time Endoscopic Prediction of Invasion Depth in Early Gastric Cancer

Jie-Hyun Kim ^{1,*}, Sang-Il Oh ², So-Young Han ¹, Ji-Soo Keum ², Kyung-Nam Kim ², Jae-Young Chun ¹, Young-Hoon Youn ¹ and Hyojin Park ¹

¹ Department of Internal Medicine, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul 06273, Republic of Korea

² Waycen Inc., Seoul 03722, Republic of Korea

* Correspondence: otilia94@yuhs.ac

Simple Summary: We previously constructed a VGG-16-based artificial intelligence (AI) model (image classifier [IC]) to predict the invasion depth in early gastric cancer (EGC) using static images. However, images cannot capture the spatio-temporal information available during real-time endoscopy. Thus, we constructed a video classifier [VC] using videos by attaching sequential layers to the last convolutional layer of the IC. We computed the standard deviation (SD) of output probabilities for a video clip and the sensitivities in the manner of frame units to observe consistency. The sensitivity, specificity, and accuracy of the IC for video clips were 33.6%, 85.5%, and 56.6%, respectively. The VC performed better analysis of the videos (sensitivity 82.3%, specificity 85.8%, and accuracy 83.7%, respectively). Furthermore, the mean SD was lower for the VC than the IC. The AI model developed utilizing videos can predict invasion depth in EGC more precisely and consistently than image-trained models, and is more appropriate for real-world situations.



Citation: Kim, J.-H.; Oh, S.-I.; Han, S.-Y.; Keum, J.-S.; Kim, K.-N.; Chun, J.-Y.; Youn, Y.-H.; Park, H. An Optimal Artificial Intelligence System for Real-Time Endoscopic Prediction of Invasion Depth in Early Gastric Cancer. *Cancers* **2022**, *14*, 6000. <https://doi.org/10.3390/cancers14236000>

Academic Editor: Sachio Fushida

Received: 8 November 2022

Accepted: 2 December 2022

Published: 5 December 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Abstract: We previously constructed a VGG-16 based artificial intelligence (AI) model (image classifier [IC]) to predict the invasion depth in early gastric cancer (EGC) using endoscopic static images. However, images cannot capture the spatio-temporal information available during real-time endoscopy—the AI trained on static images could not estimate invasion depth accurately and reliably. Thus, we constructed a video classifier [VC] using videos for real-time depth prediction in EGC. We built a VC by attaching sequential layers to the last convolutional layer of IC v2, using video clips. We computed the standard deviation (SD) of output probabilities for a video clip and the sensitivities in the manner of frame units to observe consistency. The sensitivity, specificity, and accuracy of IC v2 for static images were 82.5%, 82.9%, and 82.7%, respectively. However, for video clips, the sensitivity, specificity, and accuracy of IC v2 were 33.6%, 85.5%, and 56.6%, respectively. The VC performed better analysis of the videos, with a sensitivity of 82.3%, a specificity of 85.8%, and an accuracy of 83.7%. Furthermore, the mean SD was lower for the VC than IC v2 (0.096 vs. 0.289). The AI model developed utilizing videos can predict invasion depth in EGC more precisely and consistently than image-trained models, and is more appropriate for real-world situations.

Keywords: gastric cancer; artificial intelligence; convolutional neural networks; video; endoscopy



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Several studies have reported the usefulness of artificial intelligence (AI) in diagnostic imaging, including endoscopic imaging, to detect or diagnose neoplastic lesions. However, the role of artificial intelligence in endoscopy is distinct from that of other imaging modalities, such as computed tomography or magnetic resonance imaging [1]. In other medical devices, AI can detect and diagnose lesions from saved images after the procedure has been performed. In contrast, in endoscopy, simultaneous and continuous support of

AI is essential while the patient is undergoing the procedure to enable reliable evaluation when continuous frames are being sequentially captured.

All previous studies have used endoscopic static images to train AI models for detection and diagnosis of early gastric cancer (EGC) [2]. Previously, we also developed a convolutional neural network (CNN) model to detect EGC and predict the depth of invasion using static images [3]. In our previous model, the sensitivity, specificity, and positive predictive value (PPV) for EGC detection were 91.0%, 97.6%, and 97.5%, respectively. The sensitivity, specificity, and PPV of depth prediction were 79.2%, 77.8%, and 79.3%, respectively. The performance of depth prediction will be further improved.

The differentiation between mucosal and submucosal invasion was used to predict the tumor invasion depth in EGC. It is essential to make a diagnosis from observing subtle differences between two invasion depths, compared with the detection of EGC from normal mucosa. In real-world situations, endoscopists can predict invasion depth more precisely and reliably when observing subtle changes in the EGC lesion during an endoscopic procedure rather than still images of the lesion. Furthermore, since fine-grained visual features which come from spatio-temporal information were not considered, CNN trained on static images cannot reliably estimate tumor invasion depth in EGC. The structure of a neural network is inspired by the human brain; hence, training a CNN model in the same way as the human brain is more effective. Therefore, we developed an AI model to estimate the depth of tumor invasion in EGC using endoscopic videos. This allowed us to investigate and prove training an AI using endoscopic videos was more effective for predicting the tumor invasion depth in EGC as compared to endoscopic static images.

We trained our previous CNN model using more static images to improve the depth prediction performance. In addition, we developed the AI model using endoscopic videos and compared the performances of the two AI models using endoscopic videos.

2. Materials and Methods

2.1. Study Design and Data Preparation

The study design is summarized in Figure 1. We constructed the CNN model, which was termed as image classifier (IC) *v2*, by adding newer endoscopic static images (1582 images of mucosal cancer and 1697 images of submucosal cancer) of the training image dataset to IC *v1*, which was developed in a previous study [3]. We used independent static images of EGC (1060 images of mucosal cancer and 1060 images of submucosal cancer) as the testing set to evaluate the improvement in IC *v2* for depth prediction.

In addition, we built a CNN model called the video classifier (VC) to analyze the data from endoscopic white-light videos. The used video data were captured for 20 s at 30 frames per second (FPS) because a lesion has to be intactly placed on whole frames composing a video. A single EGC lesion was centered on each video. To use the video data before training the model, we organized video clips by setting up the exclusion criteria. Because extreme scene changes in a video can disrupt sequential continuity, we split the video into two video clips when a discontinuity was recognized. To avoid presenting endoscopy recordings to the VC that involved occlusion of the lesion, videos containing medical procedures and/or bleeding scenes were omitted. Finally, a consecutive frame section in which an EGC lesion was detected by the EGC detector developed in previous study [3] for more than 5 s was clipped from 600 frames (20 s × 30 FPS). In total, 354 video clips of EGC (189 video clips of mucosal cancer and 165 video clips of submucosal cancer) were used to train the model. Independent video clips of EGC (40 video clips of mucosal cancer and 23 video clips of submucosal cancer) were used as the validation set. After training, the model was evaluated using the test set (44 video clips of mucosal cancer and 23 video clips of submucosal cancer).

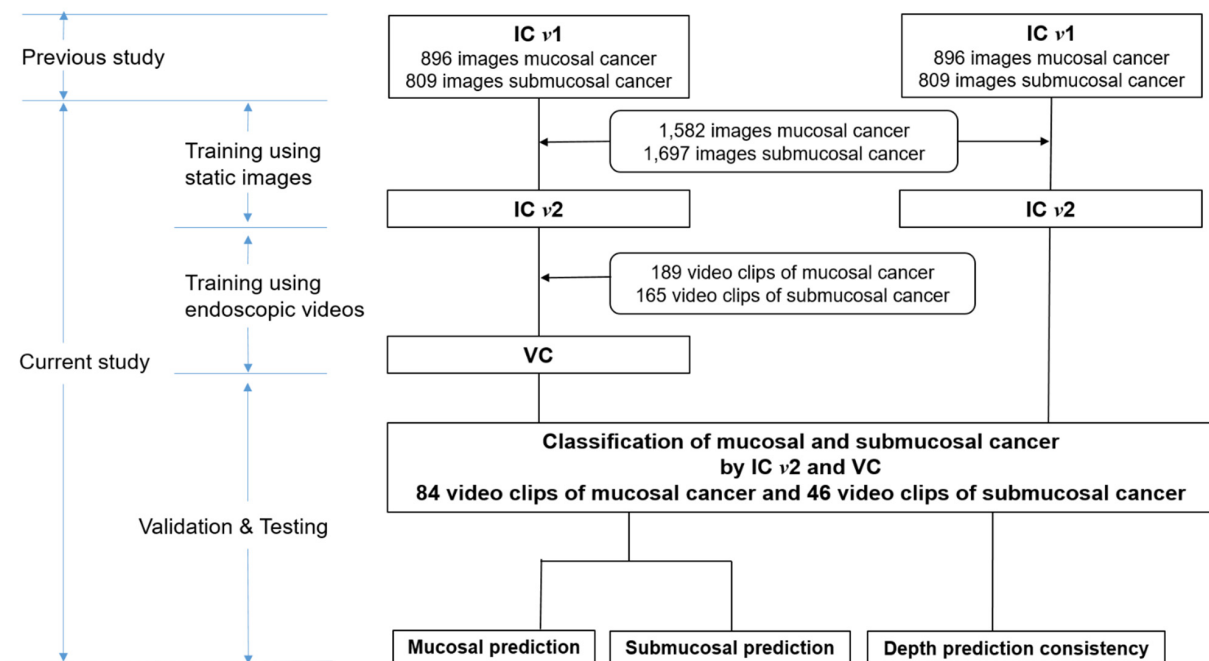


Figure 1. Flowchart of the study design. IC, image classifier; VC, video classifier.

To reflect the diversity of the video data and thus minimize the overfitting, we employed some image data augmentation methods to the training video dataset. Random flipping, random rotation, random gaussian blur, random motion blur, and sequence reverses were used to generate augmented video clips. In general, there are many types of augmentation techniques including elastic transformation and random cropping. However, the continuity which is the basic nature of video data, makes it hard to apply image augmentation methods to videos because the augmentation process may disrupt sequential information.

All endoscopic static images and videos in the current study were retrospectively consecutively collected at the Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, Korea. Baseline clinicopathological characteristics of EGC for IC *v2* and the VC are shown in Table 1. Static images and endoscopic videos were obtained using standard endoscopes (GIF-Q260J, GIF-H260, and GIF-H290; Olympus Medical Systems Co. Ltd., Tokyo, Japan). This study was approved by the institutional review board of Gangnam Severance Hospital (no. 3-2021-0341).

Table 1. Baseline clinicopathological characteristics of early gastric cancer for image classifier (IC) *v2* and the video classifier (VC).

Characteristics	IC <i>v2</i> (N = 714)	VC (N = 81)
Tumor size (mm, mean ± SD)	23.7 ± 14.4	31.0 ± 19.3
Location (n, %)		
Upper one-third	58 (8.2)	9 (11.1)
Middle one-third	171 (23.9)	12 (14.8)
Lower one-third	485 (67.9)	60 (74.1)
Gross type (n, %)		
Elevated	111 (15.5)	19 (23.4)
Flat	331 (46.4)	37 (45.7)
Depressed	272 (38.1)	25 (30.9)
* Depth of invasion (n, %)		
Mucosa (T1a)	426 (59.7)	50 (61.7)
Submucosa (T1b)	288 (40.3)	31 (38.3)
Japanese classification (n, %)		
Differentiated	419 (58.7)	45 (55.6)
Undifferentiated	295 (41.3)	36 (44.4)

* Confirmed by pathologic findings after endoscopic or surgical resection.

2.2. Image Classifiers *v1* and *v2*

A 16-layer network of the Visual Geometry Group (VGG) was employed as the backbone architecture for both IC *v1* and IC *v2*. IC *v1* was designed to mimic the architecture of VGG-16 [3], and some layers were replaced on IC *v2*. The differences between the two classifiers are shown in Figure S1. To reduce the overfitting possibility caused by unbalances between the number of parameters and the number of training data, we replaced the flattening operation used for traditional CNN feed flows to a global average pooling (GAP) layer on IC *v2*. The GAP layer measures the average values of the output from the last convolutional layer in a two-dimensional manner. Although the number of the training parameters was dynamically decreased by using the GAP layer, there was a known result that the performance losses were insignificant [4]. After flattening the output of the final convolutional layer, the two intermediate fully connected dense layers were skipped from the architecture. The IC parameters were optimized using the lesion-based training procedure proposed in our previous study³ with an adaptive moment estimation (ADAM) optimizer.

2.3. The Video Classifier

To prove the effectiveness of utilizing endoscopic videos for training the deep learning models, we constructed and evaluated two representative sequential models. Figure 2 presents an overview of the models. The preliminary convolutional layers were formed using all the convolutional layers of the VGG-16 networks. Sequential frames from the video clip were independently fed into the convolutional layers. All outputs from the last convolutional layer were flattened through the GAP layer, and concatenated for use as input features for a sequential layer. In this study, we applied a long short-term memory (LSTM) layer and gated recurrent unit (GRU) layer for each sequential model. The LSTM layer introducing a short-term state and a long-term state to the cell was proposed to overcome the long-term dependency problem of the traditional recurrent neural network (RNN), where the information is gradually faded out by an increase in timesteps [5]. The GRU layer is a streamlined version of the LSTM layer [6]. For the GRU layer, the short-term state and long-term state of the LSTM layer were integrated into a single parameter. Both layers have been widely adopted to the AI models covering sequential problems. Finally, two probability scores for the invasion depth of the input sequence, which ranged from 0 to 1, were predicted through the subsequent fully connected dense layer(s) and a softmax layer.

To train and test the deep learning models, we used TensorFlow (version 2.4) for Python (version 3.7.10) as a backend. The models were trained on dual NVIDIA GeForce RTX 3090 with 24 GB of CUDA memory for each, whereas an NVIDIA GeForce RTX 2080Ti with 11 GB of CUDA memory was used to evaluate the performance. The weights of the convolutional layers of each VC were initialized and fixed using the weights of the convolutional layers from IC *v2*. The sequential layer and fully connected dense layer(s) were trained using an ADAM optimizer with an initial learning rate of 0.0001 on the gathered video clip dataset. To measure the training losses, we applied a cross-entropy loss. In this study, the VCs used seven consecutive frames as input frame segments. The performance differences by changing the number of input frames are shown in Figure S2.

2.4. Outcome Measures

A single inference procedure of ICs outputs the probabilities of an input frame, whereas the output of the VCs is the probability of the last frame among consecutive input frames. Therefore, the output of the VCs represents the predicted invasion depth of the seventh frame in this study.

To evaluate the performance of invasion depth prediction, we derived three types of final decisions from the outputs of each classifier as follows: (1) frame-by-frame prediction, (2) video clip-based prediction by averaging outputs, and (3) video clip-based prediction by voting outputs. For frame-by-frame prediction, we assumed that the invasion depth in all frames was the same as in the source video clip. To evaluate the frame-by-frame

prediction, we counted the frames that were classified to have the correct invasion depth. The prediction from the video clip was deduced using the outputs of all the frames that constituted a video clip. There are two methods for determining the video clip-based prediction by the classifiers—averaging (2) and voting (3). In the averaging method, all probabilities were averaged to determine the maximum argument, whereas the voting method was used to determine the most frequent invasion depth that was estimated from among the all-frame results, which was denoted as the estimated invasion depth of EGC.

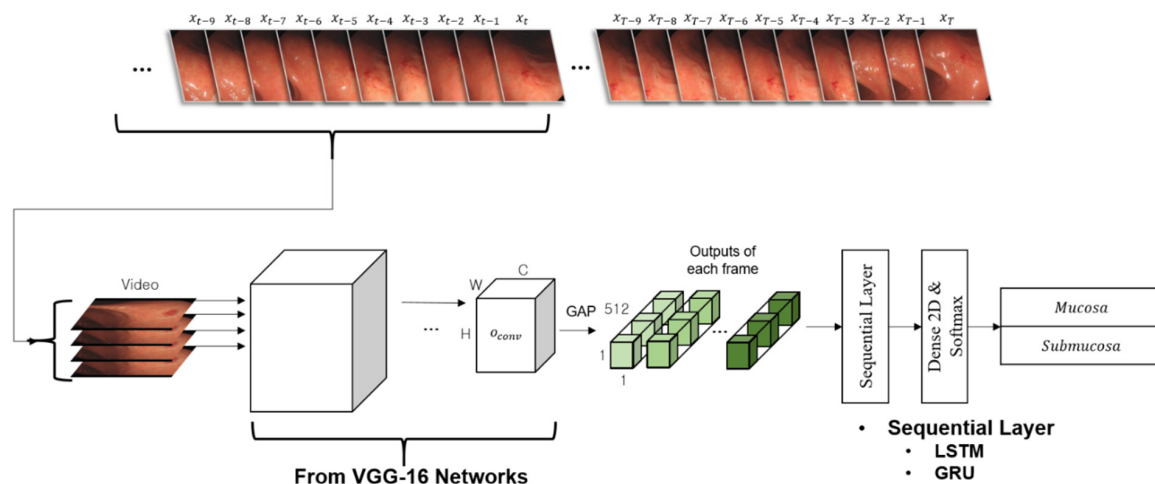


Figure 2. The network of the video classifier. The preliminary layers as convolutional layers have been formed by the all convolutional layers of visual geometry group (VGG)-16 networks. Sequential frames from a video clip were independently fed into the convolutional layers. All outputs from the last convolutional layer were stacked to be used as an input feature for a sequential layer. In this study, we have applied a long short-term memory (LSTM) layer and a gated recurrent unit (GRU) layer for each sequential-CNN type. Finally, two probability scores ranging from 0 to 1 for invasion depth of the input sequence were predicted through the subsequent fully connected dense layer(s) and a softmax layer.

To prove that consistency increased with the video data, we also examined the probability distribution for the frame-by-frame predictions.

The class activation maps (CAM) were designed to identify the discriminative image regions which are highly contributing to the final classification result of CNNs [7]. To extract CAM, the output of the last convolutional layer was multiplied by the weight matrix of the last fully connected layer. Each component of CAM has a value ranging from 0 to 1, and the higher component value indicates the more significant impact on the final decision of the model. Using this concept, we extracted boundaries from CAM by thresholding the component values to 0.5 to visualize the activated regions.

2.5. Statistical Methods

The diagnostic performance of the CNN system per image and frame was measured in terms of accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the curve (AUC) of the receiver operating characteristic (ROC) graph. The standard deviation (SD) of the output probabilities was computed using a video clip, and the sensitivities in the manner of frame units to observe consistency.

3. Results

3.1. Diagnostic Performance of IC v1 and v2 for Static Images

Using the testing set, consisting of independent static images of EGC (1060 images of mucosal cancer and 1060 images of submucosal cancer), we evaluated the diagnostic performance of IC v1 and v2 in predicting invasion depth (Table 2). The accuracy, sensitivity, specificity, PPV, and NPV of IC v1 were 79.8%, 79.2%, 77.8%, 79.3%, and 77.7%, respectively.

The values of IC *v2* were 82.7%, 82.5%, 82.9%, 82.9%, and 82.6%, respectively, which showed a better performance than IC *v1*.

Table 2. Diagnostic performance for depth prediction of image classifier (IC) *v1* and *v2* for static images.

Predicting Depth	IC <i>v1</i>	IC <i>v2</i>
Accuracy (%)	79.8	82.7
Sensitivity (%)	79.2	82.5
Specificity (%)	77.8	82.9
PPV (%)	79.3	82.9
NPV (%)	77.7	82.6

PPV, positive predictive value; NPV, negative predictive value.

3.2. Diagnostic Performance of IC *v2* and the VC for Endoscopic Videos

The diagnostic performance of IC *v2* and the VC with GRU layers were compared in terms of frame-by-frame and video clip-based prediction when endoscopic videos (44 clips of mucosal cancer and 23 clips of submucosal cancer) were used for testing. (Table 3). The diagnostic performance of the VC with LSTM layers is described in Table S2. According to the frame-by-frame prediction (Table 3A), accuracy, sensitivity, specificity, PPV, NPV, and AUC were 56.6%, 33.6%, 85.5%, 74.4%, 50.6%, and 0.615, respectively, in IC *v2*, and 83.7%, 82.3%, 85.8%, 88.0%, 79.4%, and 0.865, respectively, in the VC with GRU layers. Video clip-based prediction was evaluated using two methods: averaging output values and the voting outputs method (Table 3B). For video clip-based prediction by averaging of output values, the accuracy, sensitivity, specificity, PPV, and NPV were 50.8%, 25.0%, 100.0%, 100.0%, and 41.1%, respectively, in IC *v2*, and 85.1%, 81.8%, 91.3%, 94.7%, and 72.4%, respectively, in the VC with GRU layers. By the voting output method, the accuracy, sensitivity, specificity, PPV, and NPV were 50.8%, 25.0%, 100.0%, 100.0%, and 41.1%, respectively, and for the VC with GRU layers, the respective values were 82.1%, 81.8%, 82.6%, 90.0%, and 70.4%.

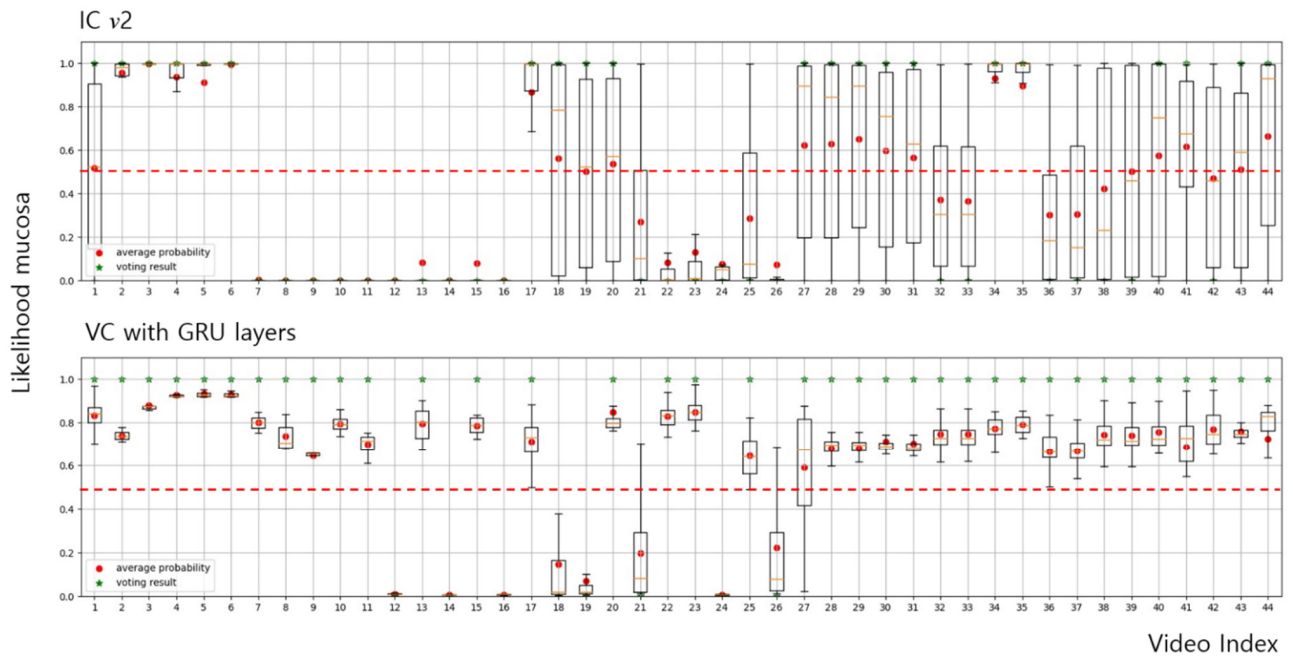
Table 3. Diagnostic performance of depth prediction of image classifier (IC) *v2* and the video classifier (VC) with GRU layers for endoscopic videos by (A) frame-by-frame prediction (B) video clip-based prediction.

(A)						
Predicting Depth	IC <i>v2</i>	VC			VC	
Accuracy (%)	56.6				83.7	
Sensitivity (%)	33.6				82.3	
Specificity (%)	85.5				85.8	
PPV (%)	74.4				88.0	
NPV (%)	50.6				79.4	
AUC	0.615				0.865	
(B)						
Predicting depth	IC <i>v2</i>	Voting		IC <i>v2</i>	Average	
		VC			VC	
Accuracy (%)	50.8	82.1		50.8		85.1
Sensitivity (%)	25.0	81.8		25.0		81.8
Specificity (%)	100.0	82.6		100.0		91.3
PPV (%)	100.0	90.0		100.0		94.7
NPV (%)	41.1	70.4		41.1		72.4

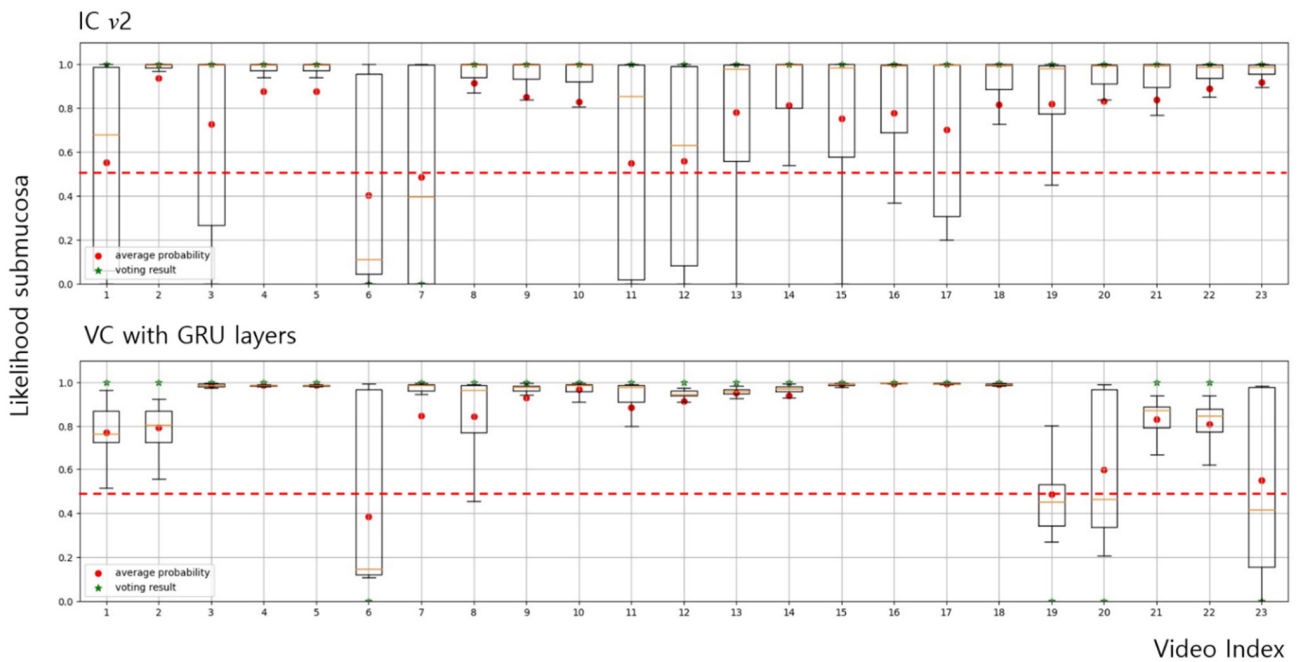
(A) PPV, positive predictive value; NPV, negative predictive value; AUC, area under the curve. (B) PPV, positive predictive value; NPV, negative predictive value.

3.3. Diagnostic Consistency between IC *v2* and the VC for Endoscopic Videos

Figure 3 shows the probability distribution in the frame-by-frame prediction results between IC *v2* and the VC with GRU layers. The mean SD of the output probabilities was lower for the VC-based analysis than for IC *v2* (0.096 vs. 0.289). That is, in many frames of IC *v2*, the invasion depth prediction showed values ranging between both extremes spanning the mucosa and submucosa. In contrast, the likelihood of depth prediction was more reliable with the VC. These findings indicate that the VC has higher diagnostic consistency than the IC. Videos S1 and S2 show the differences between IC *v2* and the VC in real-world situations.



(A)



(B)

Figure 3. The probability distribution in frame-by-frame prediction results between IC *v2* and the VC with GRU layers on (A) mucosa videos and (B) submucosa videos. IC, image classifier; VC, video classifier. The box plots are the probability distribution of each video for each invasion depth. The red dashed lines are a cut-off value to decide the correct prediction. In this study, we set the cut-off value to 0.5 because it has been widely used as a threshold value in a binary classification task. The red circles and green stars refer to the average probabilities and the voting results, respectively. In the many frames of IC *v2*, the likelihood of depth prediction showed both extremes between mucosa and submucosa. In contrast, the likelihood of depth prediction was more reliable in the VC (mean SD of output probabilities in the VC vs. IC *v2*, 0.096 vs. 0.289).

4. Discussion

Many studies have reported the development of AI models to detect gastric cancer, including our previous study, which showed favorable results [2,3,8–12]. All studies have used endoscopic static images to train AI models, whereas few have used endoscopic videos to test the AI model. However, the frame selection was suspicious and insufficient to arrive at a real-time simultaneous diagnosis during the procedure. The most important feature of real-time endoscopy is the availability of spatio-temporal information, especially temporal information, which differs from other static imaging modalities. The 3D reconstruction of static images can provide spatial information; however, temporal information can only be provided from the sequential frames extracted from real-time video. In real-world situations, endoscopists perform endoscopic procedures and make the diagnoses simultaneously. Hence, it is necessary to train AI models using endoscopic videos to develop an optimal AI model for real-world endoscopic environments. In this study, we developed a CNN model and trained it using endoscopic videos, which resulted in a better and more consistent performance than the image-trained model.

Several CNN models were developed to predict the invasion depth in EGC [3,13–17]. All CNN models showed good performance in predicting the tumor invasion depth in EGC from static endoscopic images, including our previous model [3]. However, no study has evaluated the diagnostic consistency of endoscopic videos to predict invasion depth in EGC. Diagnostic consistency should be assessed to appraise diagnostic performance precisely. The evaluation of each still image or frame-selected video cannot accurately reflect real-world scenarios. In the present study, we improved the ability of our previous model (IC *v1*) to predict the invasion depth by training it with more static images. However, the performance deteriorated when testing was performed using endoscopic videos. Furthermore, the diagnostic performance was very unreliable, and values representing both extremes between the mucosa and submucosa (Figure 3, Videos S1 and S2) were depicted. If the answer frames were selected, the CNN model would have high diagnostic performance. However, real-world situations are different. Our VC model, trained using videos, could predict the tumor invasion depth more reliably and accurately than an image-trained model (IC *v2*), as observed in the unedited video (Videos S1 and S2). The role of AI during the endoscopic procedure is decision support [18,19]. If the AI can give us information about the exact invasion depth during endoscopy, the endoscopist can make a final decision to predict the invasion depth easily and rapidly. Thus, in the real-world endoscopic environment, AI must give consistent and exact information to endoscopists. Accordingly, we developed the optimal AI model to predict the invasion depth in EGC more accurately and consistently. Our AI model can give endoscopists exact and consistent information during an endoscopic procedure.

To observe the effectiveness of utilizing video data containing spatial and temporal information, we modeled two types of video classifiers using sequential layers. In this study, we attached an LSTM and GRU to the tail of the convolutional layers. The architectures of the convolutional layers and their trained weights were adopted from the pre-trained IC *v2* for the front layers of the two video classifiers. Several sequential frames were independently fed into the convolutional layers to extract spatial features. Subsequently, the sequential layer analyzed the spatial-temporal relationship using sequentially stacked convolutional outputs as the input. Similar to our concept, the algorithms to detect anatomical landmark in the upper endoscopy have been proposed for temporal correlations [20–24]. CNN models can detect anatomical sites in the upper gastrointestinal tract but could not explain the temporal correlations among the frames. Thus, the algorithms have been proposed to consider temporal correlations.

Generally, the number of training datasets that are employed determines the efficiency of deep learning. That is, the performance of deep learning grows with increasing data [25–28]. Namikawa et al. constructed an advanced CNN model by adding more training datasets that consisted of benign lesions, to improve the low PPV of the original CNN model in detecting gastric cancer [29]. They could enhance the PPV from 30.6 to

92.5% [8,29]. The size of the dataset is significant for developing a CNN algorithm [30]. However, the types of datasets are also crucial for developing a CNN algorithm that reflects real-world situations. Our data suggest that an AI algorithm trained using endoscopic videos may be more appropriate for predicting the invasion depth of EGC in a real-time endoscopic environment. However, our data need to be externally validated using many video images.

5. Conclusions

We developed an AI model utilizing endoscopic videos which can predict the invasion depth of EGC more precisely and consistently than an image-trained model. It is more appropriate for real-world endoscopic situations.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/cancers14236000/s1>, Figure S1: The architectures of the image classifiers (IC) v1 and v2. IC v1 mimicked the architecture of the visual geometry group (VGG)-16,3 while some layers had been replaced on IC v2; Figure S2: The performance differences by varying the number of input frames. Both on the perspective of performances and computations costs, taking consecutive seven frames as the input of the network was relatively suitable for experiments; Table S1: Diagnostic performance of depth prediction of video classifier (VC) with LSTM layers for endoscopic videos Video S1: The lesion was a mucosal-invasive gastric cancer located in the lower body. The green lines were extracted by thresholding the activation values to 0.5. To show real operation environments, we only drew green lines when a frame was classified as a lesion by the EGC detector introduced in our previous study.³ In the upper row of the video, the image classifier v2 predicted the depth unreliably, showing both extremes between the mucosa and submucosa. In contrast, in the lower row of the video, the video classifier predicted mucosal invasion more reliably; Video S2: The lesion was a submucosal invasive gastric cancer located at the antrum. The green lines were extracted by thresholding the activation values to 0.5. To show real operation environments, we only drew green lines when a frame was classified as a lesion by the EGC detector introduced in our previous study.³ In the video in the upper row, the image classifier v2 predicted the depth unreliably, showing both extremes between the mucosa and submucosa. In contrast, in the lower row of the video, the video classifier predicted submucosal invasion more reliably.

Author Contributions: Conception and design: J.-H.K. and S.-I.O. Funding obtainment: J.-H.K., Provision of study data: J.-H.K., Collection and assembly of data: J.-H.K. and S.-I.O. Data analysis and interpretation: J.-H.K., S.-I.O. and J.-S.K. Manuscript writing: J.-H.K. and S.-I.O. Final approval of the manuscript: J.-H.K., S.-I.O., S.-Y.H., J.-S.K., K.-N.K., J.-Y.C., Y.-H.Y. and H.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science, and Technology (2021R1A2C2011296).

Institutional Review Board Statement: This study was conducted according to the guidelines of the Declaration of Helsinki. All imaging data were handled in accordance with institutional policies, and approval by an institutional review board of Gangnam Severance Hospital (no. 3-2021-0341).

Informed Consent Statement: Patient consent was waived because of the retrospective nature of this study and the analysis used anonymous clinical and imaging data.

Data Availability Statement: The corresponding author will provide the data presented in this study upon reasonable request.

Conflicts of Interest: Oh, S.-I., Keum, J.-S. and Kim, K.-N. are employed by Waycen, Inc. All other authors disclosed no conflict of interest.

References

1. Yoshida, S.; Tanaka, S. Artificial intelligence for the detection of gastric precancerous conditions using image-enhanced endoscopy: What kind of abilities are required for application in real-world clinical practice? *Gastrointest. Endosc.* **2021**, *94*, 549–550. [[CrossRef](#)] [[PubMed](#)]
2. Okagawa, Y.; Abe, S.; Yamada, M.; Oda, I.; Saito, Y. Artificial Intelligence in Endoscopy. *Dig. Dis. Sci.* **2021**, *67*, 1553–1572. [[CrossRef](#)] [[PubMed](#)]
3. Yoon, H.J.; Kim, S.; Kim, J.H.; Keum, J.S.; Oh, S.I.; Jo, J.; Chun, J.; Youn, Y.H.; Park, H.; Kwon, I.G.; et al. A Lesion-Based Convolutional Neural Network Improves Endoscopic Detection and Depth Prediction of Early Gastric Cancer. *J. Clin. Med.* **2019**, *8*, 1310. [[CrossRef](#)] [[PubMed](#)]
4. Lin, M.; Chen, Q.; Yan, S. Network in network. *arXiv* **2013**, arXiv:1312.4400.
5. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural. Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)] [[PubMed](#)]
6. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
7. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
8. Hirasawa, T.; Aoyama, K.; Tanimoto, T.; Ishihara, S.; Shichijo, S.; Ozawa, T.; Ohnishi, T.; Fujishiro, M.; Matsuo, K.; Fujisaki, J.; et al. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. *Gastric Cancer* **2018**, *21*, 653–660. [[CrossRef](#)]
9. Sakai, Y.; Takemoto, S.; Hori, K.; Nishimura, M.; Ikematsu, H.; Yano, T.; Yokota, H. Automatic detection of early gastric cancer in endoscopic images using a transferring convolutional neural network. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2018**, *2018*, 4138–4141.
10. Luo, H.; Xu, G.; Li, C.; He, L.; Luo, L.; Wang, Z.; Jing, B.; Deng, Y.; Jin, Y.; Li, Y.; et al. Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: A multicentre, case-control, diagnostic study. *Lancet Oncol.* **2019**, *20*, 1645–1654. [[CrossRef](#)]
11. Tang, D.; Wang, L.; Ling, T.; Lv, Y.; Ni, M.; Zhan, Q.; Fu, Y.; Zhuang, D.; Guo, H.; Dou, X.; et al. Development and validation of a real-time artificial intelligence-assisted system for detecting early gastric cancer: A multicentre retrospective diagnostic study. *EBioMedicine* **2020**, *62*, 103146. [[CrossRef](#)]
12. Yu, H.; Singh, R.; Shin, S.H.; Ho, K.Y. Artificial intelligence in upper GI endoscopy—current status, challenges and future promise. *J. Gastroenterol. Hepatol.* **2021**, *36*, 20–24. [[CrossRef](#)]
13. Kubota, K.; Kuroda, J.; Yoshida, M.; Ohta, K.; Kitajima, M. Medical image analysis: Computer-aided diagnosis of gastric cancer invasion on endoscopic images. *Surg. Endosc.* **2012**, *26*, 1485–1489. [[CrossRef](#)]
14. Zhu, Y.; Wang, Q.C.; Xu, M.D.; Zhang, Z.; Cheng, J.; Zhong, Y.S.; Zhang, Y.Q.; Chen, W.F.; Yao, L.Q.; Zhou, P.H.; et al. Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy. *Gastrointest. Endosc.* **2019**, *89*, 806–815.e1. [[CrossRef](#)]
15. Cho, B.J.; Bang, C.S.; Lee, J.J.; Seo, C.W.; Kim, J.H. Prediction of Submucosal Invasion for Gastric Neoplasms in Endoscopic Images Using Deep-Learning. *J. Clin. Med.* **2020**, *9*, 1858. [[CrossRef](#)]
16. Nagao, S.; Tsuji, Y.; Sakaguchi, Y.; Takahashi, Y.; Minatsuki, C.; Niimi, K.; Yamashita, H.; Yamamichi, N.; Seto, Y.; Tada, T.; et al. Highly accurate artificial intelligence systems to predict the invasion depth of gastric cancer: Efficacy of conventional white-light imaging, nonmagnifying narrow-band imaging, and indigo-carmin dye contrast imaging. *Gastrointest. Endosc.* **2020**, *92*, 866–873.e1. [[CrossRef](#)]
17. Nam, J.Y.; Chung, H.J.; Choi, K.S.; Lee, H.; Kim, T.J.; Soh, H.; Kang, E.A.; Cho, S.J.; Ye, J.C.; Im, J.P.; et al. Deep learning model for diagnosing gastric mucosal lesions using endoscopic images: Development, validation, and method comparison. *Gastrointest. Endosc.* **2022**, *95*, 258–268.e10. [[CrossRef](#)]
18. Pannala, R.; Krishnan, K.; Melson, J.; Parsi, M.A.; Schulman, A.R.; Sullivan, S.; Trikudanathan, G.; Trindade, A.J.; Watson, R.R.; Maple, J.T.; et al. Artificial intelligence in gastrointestinal endoscopy. *VideoGIE* **2020**, *5*, 598–613. [[CrossRef](#)]
19. Parasher, G.; Wong, M.; Rawat, M. Evolving role of artificial intelligence in gastrointestinal endoscopy. *World J. Gastroenterol.* **2020**, *26*, 7287–7298. [[CrossRef](#)]
20. Renna, F.; Martins, M.; Neto, A.; Cunha, A.; Libânio, D.; Dinis-Ribeiro, M.; Coimbra, M. Artificial Intelligence for Upper Gastrointestinal Endoscopy: A Roadmap from Technology Development to Clinical Practice. *Diagnostics* **2022**, *12*, 1278. [[CrossRef](#)]
21. Choi, S.J.; Khan, M.A.; Choi, H.S.; Choo, J.; Lee, J.M.; Kwon, S.; Keum, B.; Chun, H.J. Development of artificial intelligence system for quality control of photo documentation in esophagogastroduodenoscopy. *Surg. Endosc.* **2022**, *36*, 57–65. [[CrossRef](#)]
22. Wu, L.; Zhang, J.; Zhou, W.; An, P.; Shen, L.; Liu, J.; Jiang, X.; Huang, X.; Mu, G.; Wan, X.; et al. Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. *Gut* **2019**, *68*, 2161–2169. [[CrossRef](#)] [[PubMed](#)]
23. Chen, D.; Wu, L.; Li, Y.; Zhang, J.; Liu, J.; Huang, L.; Jiang, X.; Huang, X.; Mu, G.; Hu, S.; et al. Comparing blind spots of unsedated ultrafine, sedated, and unsedated conventional gastroscopy with and without artificial intelligence: A prospective, single-blind, 3-parallel-group, randomized, single-center trial. *Gastrointest. Endosc.* **2020**, *91*, 332–339.e3. [[CrossRef](#)] [[PubMed](#)]

24. Ding, A.; Li, Y.; Chen, Q.; Cao, Y.; Liu, B.; Chen, S.; Liu, X. Gastric Location Classification During Esophagogastroduodenoscopy Using Deep Neural Networks. In Proceedings of the 2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE), Kragujevac, Serbia, 25 October 2021; pp. 1–8.
25. Schmidt-Erfurth, U.; Sadeghipour, A.; Gerendas, B.S.; Waldstein, S.M.; Bogunovic, H. Artificial intelligence in retina. *Prog. Retin. Eye Res.* **2018**, *67*, 1–29. [[CrossRef](#)] [[PubMed](#)]
26. Alom, M.Z.; Taha, T.M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M.S.; Hasan, M.; Van Essen, B.C.; Awwal, A.A.S.; Asari, V.K. A State-of-the-Art Survey on Deep Learning Theory and Architectures. *Electronics* **2019**, *8*, 292. [[CrossRef](#)]
27. Sarker, I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 420. [[CrossRef](#)]
28. Shorten, C.; Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [[CrossRef](#)]
29. Namikawa, K.; Hirasawa, T.; Nakano, K.; Ikenoyama, Y.; Ishioka, M.; Shiroma, S.; Tokai, Y.; Yoshimizu, S.; Horiuchi, Y.; Ishiyama, A.; et al. Artificial intelligence-based diagnostic system classifying gastric cancers and ulcers: Comparison between the original and newly developed systems. *Endoscopy* **2020**, *52*, 1077–1083. [[CrossRef](#)]
30. Samarasena, J.B. Guns, germs, and steel . . . and artificial intelligence. *Gastrointest. Endosc.* **2021**, *93*, 99–101. [[CrossRef](#)]