algorithms

MDPI

Article

Reduction of Video Capsule Endoscopy Reading Times Using Deep Learning with Small Data

Hunter Morera [1], Roshan Warman [2], Azubuogu Anudu [3], Chukwudumebi Uche [3], Ivana Radosavljevic [4], Nikhil Reddy [4], Ahan Kayastha [4], Niharika Baviriseaty [4], Rahul Mhaskar [4], Andrew A. Borkowski [5,6,7], Patrick Brady [3,4], Satish Singh [8], Gerard Mullin [9], Jose Lezama [6], Lawrence O. Hall [1], Dmitry Goldgof [1,*] and Gitanjali Vidyarthi [6]

[1] Department of Computer Science and Engineering, University of South Florida, Tampa, FL 33620, USA
[2] Department of Computer Science, Yale University, New Haven, CT 06520, USA
[3] Department of Internal Medicine, University of South Florida, Tampa, FL 33620, USA
[4] Morsani College of Medicine, University of South Florida, Tampa, FL 33620, USA
[5] Department of Pathology and Cell Biology, Morsani College of Medicine, University of South Florida, Tampa, FL 33620, USA
[6] James A. Haley Veterans Hospital, Tampa, FL 33612, USA
[7] National Artificial Intelligence Institute (NAII), Washington, DC 20422, USA
[8] Veterans Affairs Boston Healthcare System, Boston, MA 02130, USA
[9] School of Medicine, Johns Hopkins University, Baltimore, MD 21205, USA
* Correspondence: goldgof@usf.edu

Abstract: Video capsule endoscopy (VCE) is an innovation that has revolutionized care within the field of gastroenterology, but the time needed to read the studies generated has often been cited as an area for improvement. With the aid of artificial intelligence, various fields have been able to improve the efficiency of their core processes by reducing the burden of irrelevant stimuli on their human elements. In this study, we have created and trained a convolutional neural network (CNN) capable of significantly reducing capsule endoscopy reading times by eliminating normal parts of the video while retaining abnormal ones. Our model, a variation of ResNet50, was able to reduce VCE video length by 47% on average and capture abnormal segments on VCE with 100% accuracy on three VCE videos as confirmed by the reading physician. The ability to successfully pre-process VCE footage as we have demonstrated will greatly increase the practicality of VCE technology without the expense of hundreds of hours of physician annotated videos.

Keywords: endoscopy; deep learning; capsule endoscopy

check for updates

Citation: Morera, H.; Warman, R.; Anudu, A.; Uche, C.; Radosavljevic, I.; Reddy, N.; Kayastha, A.; Baviriseaty, N.; Mhaskar, R.; Borkowski, A.A.; et al. Reduction of Video Capsule Endoscopy Reading Times Using Deep Learning with Small Data. *Algorithms* **2022**, *15*, 339. https://doi.org/10.3390/a15100339

Academic Editor: Antonio Sarasa-Cabezuelo

Received: 29 August 2022
Accepted: 17 September 2022
Published: 21 September 2022

1. Introduction

Upon its implementation in 2001, video capsule endoscopy (VCE) revolutionized the field of gastroenterology (GI) by allowing physicians to easily visualize areas of the bowel that were previously impossible to examine [1,2]. Despite being a powerful solution for examining 600 cm of small bowel, the field of capsule endoscopy has largely stood still since then [3]. Additionally, although VCE has a wide range of clinical applications, interpretation of VCE studies can be a daunting task, as it requires an individual with expertise to review more than 50,000–70,000 image frames, with 8 h of recording, looking for abnormal pathology that is often only present in one or two frames [4,5]. As a result of the limitations of human concentration [6], VCE readings have a significant miss rate of 5.9% for vascular lesions, 0.5% for ulcers, and 18.9% for neoplasms [7]. Over the years, VCE manufacturers have continuously refined their software to include advances in interpretation aids such as blood indicators, Quickview modes, and adaptive framerates, yet capsule reading remains a tedious and time-consuming process [8–10]. With the evolution of artificial intelligence (AI), computer-aided diagnosis (CAD) has shown promise in many areas of medicine, including pathology, dermatology, radiology and gastroenterology, where it has been used to decrease observational oversights (i.e., human error). With the above-mentioned limitations of VCE, CAD seems to be well poised to resolve several gaps in maximizing the function of VCE [11]. Though previous studies have been published regarding the use of CAD with VCE [12–19], these studies are either conducted on tens to hundreds of non-publicly available videos, or only use publicly available frame data. The aim of this study is to use CAD to identify abnormal vascular and inflammatory capsule endoscopy frames from full video footage, using a limited number of publicly available videos.

2. Materials and Methods

In this work, we used publicly available data from the *κάψουλα* interactive database (KID) [20,21]. The database consists of two types of data: still frame images and videos. All data in the database was collected with a MiroCam device developed by IntroMedic Co in Seoul, Korea. The MiroCam has a recording frame rate of 3 frames per second [22]. However, the videos provided in the database are all MP4 files with a frame rate of five frames per second. Previous work [23–30] using this database focused on classification on the frame only data known as KID Datasets 1 and 2. These datasets were annotated by expert clinicians and come with both classification and segmentation labels. For the videos, only the KID video 1 contains classification labels and it is the only one to have been used in previous work [31] referencing the KID dataset. The other two videos are without expert annotation; with the help of our local Veteran's Affairs (VA) Hospital, we are the first to use these previously unannotated KID videos. Our senior author, a VA physician with 20 years of GI experience, labeled videos 2 and 3 and adjusted the labels on the first video according to their findings. This resulted in three full-length capsule endoscopy videos with annotations that we used for cross-validation of our deep learning approach.

The goal of this work is to reduce the physician time needed to read the VCE videos; thus, our framework focused on removing confidently predicted normal frames which are not needed for diagnoses. Before training a model, we had to pre-process the videos by separating each into frames using the five frames per second speed of the MP4 files. All frames contain a time code in the top corner as seen in Figure 1, which are what the physician uses to denote the location of any abnormalities. The ground truth for this data is a list of segments with a start and ending time of when the abnormality is visible in the video. To provide the label for each image, we used optical character recognition to read the time code and determine if the frame was indicated as abnormal by the physician. The final step before training the deep learner was to remove the time code and IntroMedic text from the images as seen in Figure 2, to prevent the model from seeing this information and using it to make predictions. We decided to use a well-known and robust convolutional neural network architecture, ResNet50 [32], which has been trained to achieve high performance on various classification benchmark datasets. ResNet50, with initial weights from training on ImageNet [33], was fine-tuned with the KID videos dataset. The classification layer of ResNet was replaced by a single fully connected layer with two neurons as seen in Figure 3. This output layer used a sigmoid activation function to classify the frames as either normal or abnormal. All models were trained using stochastic gradient descent as the optimizer using a learning rate of 0.001 with class weights to account for the abnormal frames being a minority.

Typically, a network would have all its starting weights generated randomly prior to training, and learn optimal weights based on the training data. However, we had a very small dataset in this study, so we decided to use the final weights learned on the ImageNet dataset as the starting point [33]. As we only had three annotated videos to work with, it was decided that experiments would be conducted using a three-fold cross validation scheme. Thus, we would train a model with two videos, testing on the third, and repeat for all combinations. With our architecture and pre-processing pipeline chosen, our first experiment was to train a model using the videos and make predictions with that model. We found that because our dataset was small, this single model approach did not perform well. With small datasets, it is useful to use multiple networks in an ensemble approach to increase performance. Traditional ensembles require training multiple models that are used for voting on final predictions, which can be very time consuming. However, more recently, it has been shown that saving intermediate model's weights while training can save time and still produce the desired effect of an ensemble [34].

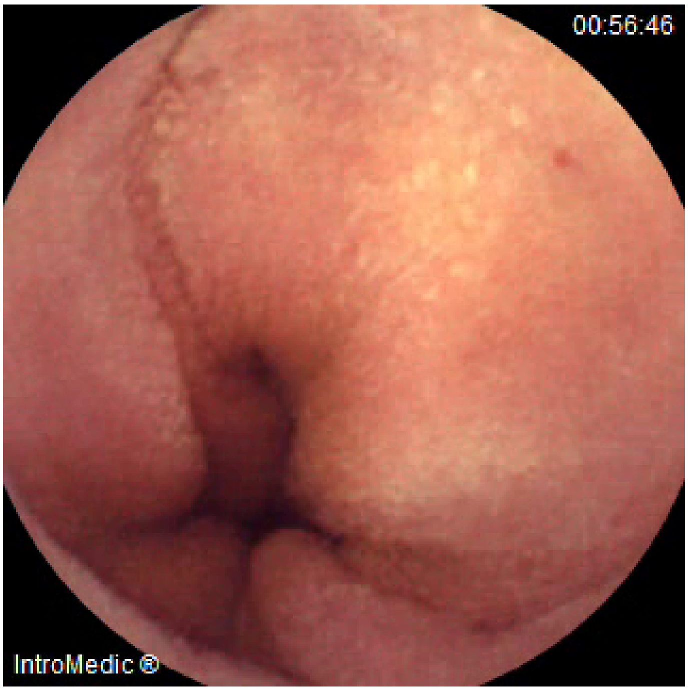

Figure 1. KID video 1 frame.

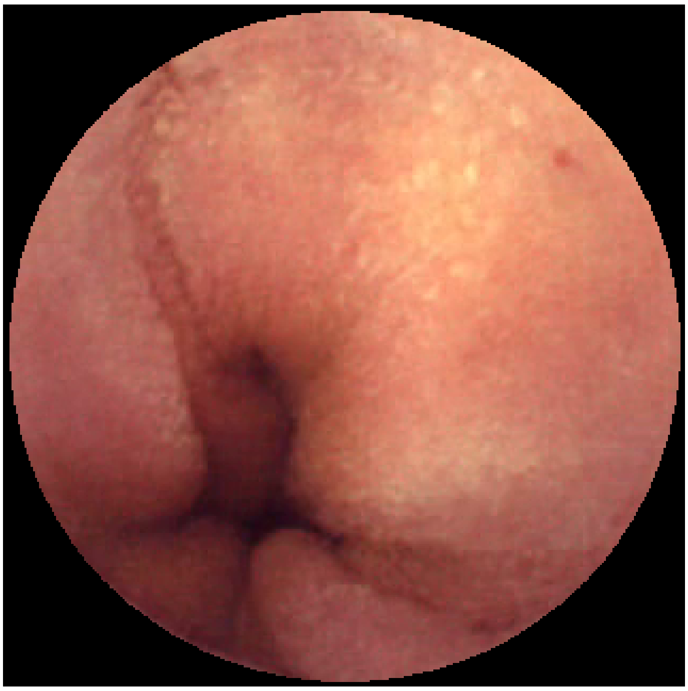

Figure 2. KID video 1 frame after pre-processing.

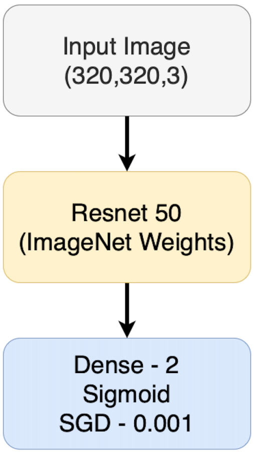

Figure 3. Network Architecture.

While training the single model, we found that the training accuracy began to plateau around 99% after ten epochs. If training continues past this point, the model begins to over-fit to the training data and not generalize well on the test data. Thus, we decided to train for only ten epochs, saving the model weights after each epoch to use for voting. We then conducted a second experiment using the weights of nine of these intermediate models to vote on the class of each frame in the video. We found an increase in performance over the single model; however, overall, there were still many missing abnormal segments. KID videos 1 and 3 had significantly less abnormal frames than KID video 2, an imbalance which we identified as a potential problem.

Our solution was to augment the abnormal frames in these two videos using rotation. KID video 2 has 5050 abnormal frames; therefore, for KID video 1, we rotated each of the abnormal frames by 30 degrees resulting in 12 rotations, and in KID video 3, we rotated each of the abnormal frames by 72 degrees resulting in 5 rotations. These rotations increased the number of abnormal frames from 429 and 988 to 5148 and 4940 for videos 1 and 3 respectively. In our final experiment, the rotated abnormal images were only used in training the model, not in testing. An example of this rotation can be seen in Figure 4. This augmentation, combined with the use of intermediate model weights to vote on predictions, provided us with our best results.

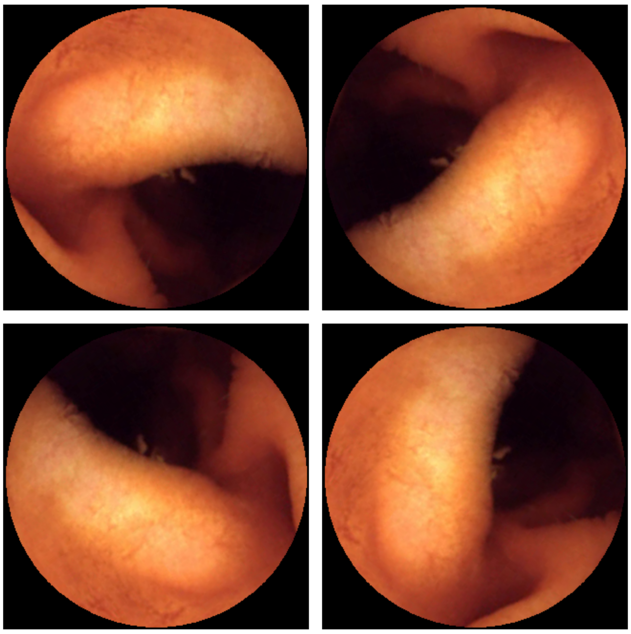

Figure 4. Rotation augmentation examples from KID video 3.

3. Results

The ground truth for the videos is a list of time segments where an abnormality was observed and annotated by an expert. An abnormality may be visible for as short as a few seconds to as long as several minutes. We considered an abnormal segment detected if at least one frame from the segment is predicted as abnormal. In our first experiment, we trained the deep learning model detailed in Figure 3, for ten epochs, using the final epoch to make predictions on the test videos. The results for the three videos are shown in Table 1. This single network was able to identify 21 of the 119 abnormalities, while reducing the number of frames on average by 97%.

Table 1. Single Network Trained on Non-Augmented Videos.

Video Name	Total Frames	Number of Abnormal Segments	Reduced Number of Frames	Abnormal Segments Detected
KID Video 1	28,480 (95 min)	22	616 (2 min)	3
KID Video 2	117,565 (391 min)	86	956 (3 min)	10
KID Video 3	74,762 (249 min)	11	4522 (15 min)	8

Though this method produces a sizeable reduction in video length, the number of abnormalities detected is less than 18%. In real world applications, there will always be a physician to look over the predictions; thus, normal frames being incorrectly classified as abnormal is acceptable. However, incorrectly classifying abnormal frames as normal could have dire consequences for patient outcomes. Therefore, our goal is to limit the number of missed abnormal segments. Since the single model did not perform well, we decided to use an ensemble approach to increase performance. As discussed above, we decided to save intermediate models while training to have them vote on predictions. In our second experiment, we again trained the model described in Figure 3 for ten epochs; however, we saved each of the ten intermediate models. We then had the first nine intermediate models make a prediction for each frame in the videos. Since the most important aspect is not to miss an abnormality, we decided on a voting scheme where all of the nine networks had to agree on a normal frame for it to be given that label. A paired *t*-test with a confidence interval of 95% showed a statistically significant improvement in the detection of abnormal segments with this ensemble in the three-fold cross validation over a single model as seen in Table 2 below.

Table 2. Paired *t*-test Results Comparing Accuracy of Single Network vs. Ensemble with No Augmentation.

Model	n	Mean	Variance	t-calc	t-crit	df	*p*
Single Network	119	0.176	0.147	11.72	1.98	118	1.66×10^{-21}
Ensemble No Aug	119	0.714	0.206				

As shown in Table 3, the ensemble method was able to identify 85 of the 119 abnormal segments, while reducing the number of frames by 75%. The voting was repeated using the last nine models, and it was found to perform marginally worse with only 80 of the 119 abnormal segments detected, and a reduction in number of frames by 76%.

Table 3. Ensemble of Networks Trained on Non-Augmented Videos.

Video Name	Total Frames	Number of Abnormal Segments	Reduced Number of Frames	Abnormal Segments Detected
KID Video 1	28,480 (95 min)	22	10,747 (36 min)	21
KID Video 2	117,565 (391 min)	86	10,960 (37 min)	53
KID Video 3	74,762 (249 min)	11	33,963 (113 min)	11

This method performed very well on videos 1 and 3 but missed a large number of abnormal segments on video 2. We hypothesized this was due to the number of abnormal examples in the training set. Video 1 has 22 abnormal segments with a total of 429 abnormal frames, video 2 has 86 abnormal segments with a total of 5050 abnormal frames, and video 3 has 11 abnormal segments with a total of 988 abnormal frames. This means when testing on video 2 and training with videos 1 and 3, we have a significantly lower number of abnormal frames for training than the other two folds. Thus, to compensate for the large variance in abnormal frames per video, we decided to use augmentation to increase the number of frames in videos 1 and 3 as described previously. We then re-trained the networks using the videos with the augmented abnormal frames, again saving the ten intermediate models. Using the first nine models, the same voting procedure was used such that all networks needed to agree a frame was normal for it to be given that label. The results of this experiment with augmentation on the abnormal frames can be seen in Table 4 below.

Table 4. Ensemble of Networks Trained on Augmented Videos.

Video Name	Total Frames	Number of Abnormal Segments	Reduced Number of Frames	Abnormal Segments Detected
KID Video 1	28,480 (95 min)	22	14,828 (49 min)	22
KID Video 2	117,565 (391 min)	86	84,672 (282 min)	86
KID Video 3	74,762 (249 min)	11	27,099 (90 min)	11

The average reduction in number of frames is 47%, while the number of abnormal segments detected is 119 of 119. A paired *t*-test with a confidence interval of 95% showed that this ensemble method with augmented data had a statistically significant improvement in the detection of abnormal segments in the three-fold cross validation over the ensemble without augmentation as shown in Table 5. The same voting procedure was repeated using the last nine models. It was slightly worse in detecting 118 of 119 abnormalities while reducing the videos on average by the same 47% as voting the first nine models. Despite working with a small dataset of only two videos for training in each fold, we were able to develop an algorithm that successfully detected 100% of abnormal segments while reducing the reading time for a physician by 47% on average.

Table 5. Paired *t*-test Results Comparing Accuracy of Ensemble without Augmentation vs. Ensemble with Augmentation.

Model	n	Mean	Variance	t-calc	t-crit	df	*p*
Ensemble No Aug	119	0.714	0.206	6.87	1.98	118	3.21×10^{-10}
Ensemble With Aug	119	1	0				

4. Discussion

The goal of our study was to reduce the physician time needed to read VCE studies by excluding frames classified as normal without sacrificing diagnostic accuracy. Our study used a robust convolutional neural network (CNN) that was trained and tested to classify a frame as normal or abnormal, with the goal of minimizing the number of false negatives. As such, using our CNN model, the physician can confidently exclude normal frames and carefully analyze abnormal frames to ensure accurate diagnosis. Using the method proposed in our study, we were able to reduce the video length on average by 47%. By doing this, our model was able to remove frames confidently classified as normal while accurately classifying frames within abnormal segments. Using this model, we captured frames from all 119 of the abnormal segments, for an accuracy of 100%. We would note that KID video 2 is of lesser quality than the two other videos in the database. Our approach has the promise of significantly reducing the reading time needed by a physician to review VCE images. This study contributes to the growing field of CNN in the detection of subtle lesions in the small bowel by demonstrating the ability to perform well with small training data. This is important due to the cost associated with collecting and annotating large amounts of medical data.

Similar studies have efficiently detected pathology in VCE in shorter time periods. As described in [35], the authors attempted to use AI to improve the performance of VCE using machine learning called Computer-Assisted Diagnosis for CAPsule Endoscopy (CAD-CAP). CAD-CAP had a sensitivity of 100% in the detection of angiectasias; however, this was done with clean and still images. Our study had the benefit of having the physician review all the abnormal frames, thus ensuring an accurate diagnosis. As described in [36], a custom CNN was trained and evaluated using a total of 5825 WCE videos and was able to detect 95.9% of abnormal findings. In [37], capsule endoscopy data was used to train a CNN to find protruding lesions in the small bowel. Using 30,000 images from 292 subjects, they obtained 0.91 AUC with a sensitivity of 90.7%. In [38], the authors obtained a sensitivity of 88.2% using a CNN trained on 5360 images when looking for erosions and ulcerations in the small bowel. In [39], the authors were able to train a CNN to classify abnormal pathology from normal pathology with sensitivities and specificity as high as 99.88% and 99.09% compared with physician reads that had sensitivities and specificity of 74.57% and 76.89%. Their method reduces reading times on average from 96.6 min to 5.9 min. However, their study involved the use of around 5000 videos collected and annotated by 21 physicians in China with unknown specialties, experience, and skill.

Our study has some limitations which offer opportunity for improvement. Our CNN model was trained only on two videos and tested on one, which limits the scale of our study and the results achieved. We hope to improve this by increasing the data available in training our CNN model, with the goal of establishing an algorithm that can delineate between normal and abnormal frames instantly. Our goal is to expand our data and to continue to train our CNN model to detect abnormal frameworks, ultimately reducing physician reading time.

Artificial intelligence (AI) is an emerging field in gastroenterology, particularly in VCE. Our study focused on training a CNN-based model to identify subtle inflammatory and vascular lesions, with the goal of developing an algorithm that can quickly differentiate between normal and abnormal frames. Our model demonstrated the ability to find all abnormal regions with significant reduction in physician reading time. Our study results demonstrate the benefit of using a CNN in processing VCE images. We believe our study lays an excellent foundation for further validation in large multi-center trials.

Author Contributions: Conceptualization, R.W., R.M., A.A.B., P.B., S.S., G.M., J.L. and G.V.; Methodology, H.M., L.O.H. and D.G.; Software, H.M., L.O.H. and D.G.; Supervision, L.O.H., D.G. and G.V.; Validation, H.M., L.O.H. and D.G.; Writing—original draft, H.M., A.A., C.U. and G.V.; Writing—review and editing, H.M., I.R., N.R., A.K., N.B., R.M., A.A.B., P.B., S.S., G.M., J.L., L.O.H., D.G. and G.V. All authors have read and agreed to the submitted version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The KID dataset is publicly available upon request https://mdss.uth.gr/datasets/endoscopy/kid/, accessed on 2 April 2022. The code repository will be released upon acceptance for publication.

Acknowledgments: We would like to thank the physicians at our local Veterans Affairs Hospital for their input in the development of this study. We would also like to thank Haley Futch, for her extensive work in reviewing and editing this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

VCE	Video Capsule Endoscopy
CNN	Convolutional Neural Network

References

1. Byrne, F.M.; Donnellan, F. Artificial intelligence and capsule endoscopy: Is the truly "smart" capsule nearly here? *Gastrointest. Endosc.* **2019**, *89*, 195–197. [CrossRef] [PubMed]
2. Pogorelov, K.; Suman, S.; Azmadi Hussin, F.; Saeed Malik, A.; Ostroukhova, O.; Riegler, M.; Halvorsen, P.; Hooi Ho, S.; Goh, K.L. Bleeding detection in wireless capsule endoscopy videos - Color versus texture features. *J. Appl. Clin. Med. Phys.* **2019**, *20*, 141–154. [CrossRef] [PubMed]
3. Iddan, G.; Meron, G.; Glukhovsky, A.; Swain, P. Wireless capsule endoscopy. *Nature* **2000**, *405*, 405–417. [CrossRef] [PubMed]
4. Rondonotti, E.; Pennazio, M.; Toth, E.; Koulaouzidis, A. How to read small bowel capsule endoscopy: A practical guide for everyday use. *Endosc. Int. Open* **2020**, *8*, E1220–E1224. [CrossRef] [PubMed]
5. Kim, S.H.; Lim, Y.J. Artificial Intelligence in Capsule Endoscopy: A Practical Guide to Its Past and Future Challenges. *Diagnostics* **2021**, *11*, 1722. [CrossRef]
6. Beg, S.; Card, T.; Sidhu, R.; Wronska, E.; Ragunath, K.; Ching, H.L.; Koulaouzidis, A.; Yung, D.; Panter, S.; Mcalindon, M.; et al. The impact of reader fatigue on the accuracy of capsule endoscopy interpretation. *Dig. Liver Dis.* **2021**, *53*, 1028–1033. [CrossRef]
7. Lewis, B.S.; Eisen, G.M.; Friedman, S. A pooled analysis to evaluate results of capsule endoscopy trials. *Endoscopy* **2005**, *37*, 960–965. [CrossRef]
8. Xavier, S.; Monteiro, S.; Magalhães, J.; Rosa, B.; Moreira, M.J.; Cotter, J. Capsule endoscopy with PillCamSB2 versus PillCamSB3: Has the improvement in technology resulted in a step forward? *Rev. Española Enfermedades Dig.* **2018**, *110*, 155–159. [CrossRef]
9. Buscaglia, J.M.; Giday, S.A.; Kantsevoy, S.V.; Clarke, J.O.; Magno, P.; Yong, E.; Mullin, G.E. Performance characteristics of the suspected blood indicator feature in capsule endoscopy according to indication for study. *Clin. Gastroenterol. Hepatol.* **2008**, *6*, 298–301. [CrossRef]
10. Shiotani, A.; Honda, K.; Kawakami, M.; Kimura, Y.; Yamanaka, Y.; Fujita, M.; Matsumoto, H.; Tarumi, K.I.; Manabe, N.; Haruma, K. Analysis of small-bowel capsule endoscopy reading by using Quickview mode: Training assistants for reading may produce a high diagnostic yield and save time for physicians. *J. Clin. Gastroenterol.* **2012**, *46*, 92–95. [CrossRef]
11. Dray, X.; Iakovidis, D.; Houdeville, C.; Jover, R.; Diamantis, D.; Histace, A.; Koulaouzidis, A. Artificial intelligence in small bowel capsule endoscopy-current status, challenges and future promise. *J. Gastro Hepatol.* **2021**, *36*, 12–19. [CrossRef] [PubMed]
12. Noya, F.; Álvarez-González, M.A.; Benitez, R. Automated angiodysplasia detection from wireless capsule endoscopy. In Proceedings of the 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju, Korea, 11–15 July 2017; pp. 3158–3161.
13. Iakovidis, D.K.; Koulaouzidis, A. Software for enhanced video capsule endoscopy: Challenges for essential progress. *Nat. Rev. Gastroenterol. Hepatol.* **2015**, *12*, 172–186. [CrossRef] [PubMed]
14. Muruganantham, P.; Balakrishnan, S.M. Attention aware deep learning model for wireless capsule endoscopy lesion classification and localization. *J. Med. Biol. Eng.* **2022**, *42*, 157–168. [CrossRef]
15. Hosoe, N.; Horie, T.; Tojo, A.; Sakurai, H.; Hayashi, Y.; Limpias Kamiya, K.J.L.; Sujino, T.; Takabayashi, K.; Ogata, H.; Kanai, T. Development of a Deep-Learning Algorithm for Small Bowel-Lesion Detection and a Study of the Improvement in the False-Positive Rate. *J. Clin. Med.* **2022**, *11*, 3682. [CrossRef] [PubMed]
16. Ding, Z.; Shi, H.; Zhang, H.; Zhang, H.; Tian, S.; Zhang, K.; Cai, S.; Ming, F.; Xie, X.; Liu, J.; et al. Artificial intelligence-based diagnosis of abnormalities in small-bowel capsule endoscopy. *Endoscopy* **2022**. [CrossRef]
17. Son, G.; Eo, T.; An, J.; Oh, D.J.; Shin, Y.; Rha, H.; Kim, Y.J.; Lim, Y.J.; Hwang, D. Small Bowel Detection for Wireless Capsule Endoscopy Using Convolutional Neural Networks with Temporal Filtering. *Diagnostics* **2022**, *12*, 1858. [CrossRef]
18. Raut, V.; Gunjan, R.; Shete, V.V.; Eknath, U.D. Small Bowel Gastrointestinal tract disease segmentation and classification in wireless capsule endoscopy using intelligent deep learning model. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* **2022**, 1–17. [CrossRef]
19. Aoki, T.; Yamada, A.; Kato, Y.; Saito, H.; Tsuboi, A.; Nakada, A.; Niikura, R.; Fujishiro, M.; Oka, S.; Ishihara, S.; et al. Automatic detection of various abnormalities in capsule endoscopy videos by a deep learning-based system: A multicenter study. *Gastrointest. Endosc.* **2021**, *93*, 165–173. [CrossRef]
20. Iakovidis, D.K.; Georgakopoulos, S.V.; Vasilakakis, M.; Koulaouzidis, A.; Plagianakos, V.P. Detecting and Locating Gastrointestinal Anomalies Using Deep Learning and Iterative Cluster Unification. *IEEE Trans. Med. Imaging* **2018**, *37*, 2196–2210. [CrossRef]
21. Koulaouzidis, A.; Iakovidis, D.K.; Yung, D.E.; Rondonotti, E.; Kopylov, U.; Plevris, J.N.; Toth, E.; Eliakim, A.; Johansson, G.W.; Marlicz, W.; et al. KID Project: An internet-based digital video atlas of capsule endoscopy for research purposes. *Endosc. Int. Open* **2017**, *5*, 477–483. [CrossRef]
22. Fernandez-Urien, I.; Carretero, C.; Borobio, E.; Borda, A.; Estevez, E.; Galter, S.; Gonzalez-Suarez, B.; Gonzalez, B.; Lujan, M.; Martinez, J.L.; et al. KID Project: An internet-based digital video atlas of capsule endoscopy for research purposes. *World J. Gastroenterol.* **2014**, *20*, 14472. [CrossRef]
23. Amiri, Z.; Hassanpour, H.; Beghdadi, A. Feature extraction for abnormality detection in capsule endoscopy images. *Biomed. Signal Process. Control* **2022**, *71*, 103219. [CrossRef]
24. Caroppo, A.; Leone, A.; Siciliano, P. Deep transfer learning approaches for bleeding detection in endoscopy images. *Comput. Med Imaging Graph.* **2021**, *88*, 1–8. [CrossRef]

25. Vasilakakis, M.; Sovatzidi, G.; Iakovidis, D.K. Explainable classification of weakly annotated wireless capsule endoscopy images based on a fuzzy bag-of-colour features model and brain storm optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2021; pp. 488–498.
26. Vieira, P.M.; Freitas, N.R.; Lima, V.B.; Costa, D.; Rolana, C.; Lima, C.S. Multi-pathology detection and lesion localization in WCE videos by using the instance segmentation approach. *Artif. Intell. Med.* **2021**, *119*, 102141. [CrossRef]
27. Jain, S.; Seal, A.; Ojha, A.; Yazidi, A.; Bures, J.; Tacheci, I.; Krejcar, O. A deep CNN model for anomaly detection and localization in wireless capsule endoscopy images. *Comput. Biol. Med.* **2021**, *137*, 104789. [CrossRef]
28. Jain, S.; Seal, A.; Ojha, A.; Krejcar, O.; Bureš, J.; Tachecí, I.; Yazidi, A. Detection of abnormality in wireless capsule endoscopy images using fractal features. *Comput. Biol. Med.* **2020**, *127*, 104094. [CrossRef]
29. Diamantis, D.E.; Iakovidis, D.K.; Koulaouzidis, A. Look-behind fully convolutional neural network for computer-aided endoscopy. *Biomed. Signal Process. Control* **2019**, *49*, 192–201. [CrossRef]
30. Guo, X.; Yuan, Y. Triple ANet: Adaptive abnormal-aware attention network for WCE image classification. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019; pp. 293–301.
31. Raut, V.; Gunjan, R. Transfer learning based video summarization in wireless capsule endoscopy. *Int. J. Inf. Technol.* **2022**, 1–8. [CrossRef]
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
33. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
34. Huang, G.; Li, Y.; Pleiss, G.; Liu, Z.; Hopcroft, J.E.; Weinberger, K.Q. Snapshot ensembles: Train 1, get m for free. *arXiv* **2017**, arXiv:1704.00109.
35. Leenhardt, R.; Li, C.; Le Mouel, J.P.; Rahmi, G.; Saurin, J.C.; Cholet, F.; Boureille, A.; Amiot, X.; Delvaux, M.; Duburque, C.; et al. CAD-CAP: A 25,000-image database serving the development of artificial intelligence for capsule endoscopy. *Endosc. Int. Open* **2020**, *8*, 415–420. [CrossRef]
36. Xie, X.; Xiao, Y.F.; Zhao, X.Y.; Li, J.J.; Yang, Q.Q.; Peng, X.; Nie, X.B.; Zhou, J.Y.; Zhao, Y.B.; Yang, H.; et al. Development and Validation of an Artificial Intelligence Model for Small Bowel Capsule Endoscopy Video Review. *JAMA Netw. Open* **2022**, *5*, 2221992. [CrossRef]
37. Saito, H.; Aoki, T.; Aoyama, K.; Kato, Y.; Tsuboi, A.; Yamada, A.; Fujishiro, M.; Oka, S.; Ishihara, S.; Matsuda, T.; et al. Automatic detection and classification of protruding lesions in wireless capsule endoscopy images based on a deep convolutional neural network. *Gastrointest. Endosc.* **2020**, *92*, 144–151. [CrossRef]
38. Aoki, T.; Yamada, A.; Aoyama, K.; Saito, H.; Tsuboi, A.; Nakada, A.; Niikura, R.; Fujishiro, M.; Oka, S.; Ishihara, S.; et al. Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network. *Gastrointest. Endosc.* **2019**, *89*, 357–363. [CrossRef]
39. Ding, Z.; Shi, H.; Zhang, H.; Meng, L.; Fan, M.; Han, C.; Zhang, K.; Ming, F.; Xie, X.; Liu, H.; et al. Gastroenterologist-Level Identification of Small-Bowel Diseases and Normal Variants by Capsule Endoscopy Using a Deep-Learning Model. *Nat. Rev. Gastroenterol.* **2019**, *157*, 1044–1054. [CrossRef]