Medical Image Analysis 94 (2024) 103157

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Robustness evaluation of deep neural networks for endoscopic image analysis: Insights and strategies

Tim J.M. Jaspers [a,*], Tim G.W. Boers [a], Carolus H.J. Kusters [a], Martijn R. Jong [b], Jelmer B. Jukema [b], Albert J. de Groof [b], Jacques J. Bergman [b], Peter H.N. de With [a], Fons van der Sommen [a]

[a] Department of Electrical Engineering, Video Coding & Architectures, Eindhoven University of Technology, Eindhoven, The Netherlands
[b] Department of Gastroenterology and Hepatology, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, The Netherlands

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Deep learning
Endoscopy
Robustness
Image degradation
Image quality

ABSTRACT

Computer-aided detection and diagnosis systems (CADe/CADx) in endoscopy are commonly trained using high-quality imagery, which is not representative for the heterogeneous input typically encountered in clinical practice. In endoscopy, the image quality heavily relies on both the skills and experience of the endoscopist and the specifications of the system used for screening. Factors such as poor illumination, motion blur, and specific post-processing settings can significantly alter the quality and general appearance of these images. This so-called *domain gap* between the data used for developing the system and the data it encounters after deployment, and the impact it has on the performance of deep neural networks (DNNs) supportive endoscopic CAD systems remains largely unexplored. As many of such systems, for e.g. polyp detection, are already being rolled out in clinical practice, this poses severe patient risks in particularly community hospitals, where both the imaging equipment and experience are subject to considerable variation. Therefore, this study aims to evaluate the impact of this domain gap on the clinical performance of CADe/CADx for various endoscopic applications. For this, we leverage two publicly available data sets (KVASIR-SEG and GIANA) and two in-house data sets. We investigate the performance of commonly-used DNN architectures under synthetic, clinically calibrated image degradations and on a prospectively collected dataset including 342 endoscopic images of lower subjective quality. Additionally, we assess the influence of DNN architecture and complexity, data augmentation, and pretraining techniques for improved robustness. The results reveal a considerable decline in performance of 11.6% (±1.5) as compared to the reference, within the clinically calibrated boundaries of image degradations. Nevertheless, employing more advanced DNN architectures and self-supervised in-domain pre-training effectively mitigate this drop to 7.7% (±2.03). Additionally, these enhancements yield the highest performance on the manually collected test set including images with lower subjective quality. By comprehensively assessing the robustness of popular DNN architectures and training strategies across multiple datasets, this study provides valuable insights into their performance and limitations for endoscopic applications. The findings highlight the importance of including robustness evaluation when developing DNNs for endoscopy applications and propose strategies to mitigate performance loss.

1. Introduction

Gastrointestinal (GI) cancers are a significant health concern worldwide, with a high incidence and mortality rate (Arnold et al., 2020; Chen et al., 2016b). Early detection of GI cancers is crucial for improving patient outcomes and reducing the burden of the disease and significantly increasing the prognosis. Endoscopy is the golden standard for GI cancer screening and diagnosis, providing direct visualization of the GI tract and the ability to obtain tissue samples for histological examination. However, the yield of endoscopic examination depends highly on the skills of the physician, where especially early neoplasia are difficult to detect (Beveridge et al., 2023; Schölvinck et al., 2017). This has resulted into an increase in attention towards computer-aided detection and diagnosis (CADe/CADx) systems for GI cancer screening (de Groof et al., 2020; Ebigbo et al., 2019; Guimarães et al., 2020; Ozawa et al., 2020; Byrne et al., 2019; Chen et al., 2018; Fockens et al., 2023a).

* Corresponding author.
E-mail address: t.j.m.jaspers@tue.nl (T.J.M. Jaspers).

https://doi.org/10.1016/j.media.2024.103157
Received 25 July 2023; Received in revised form 19 March 2024; Accepted 21 March 2024
Available online 29 March 2024

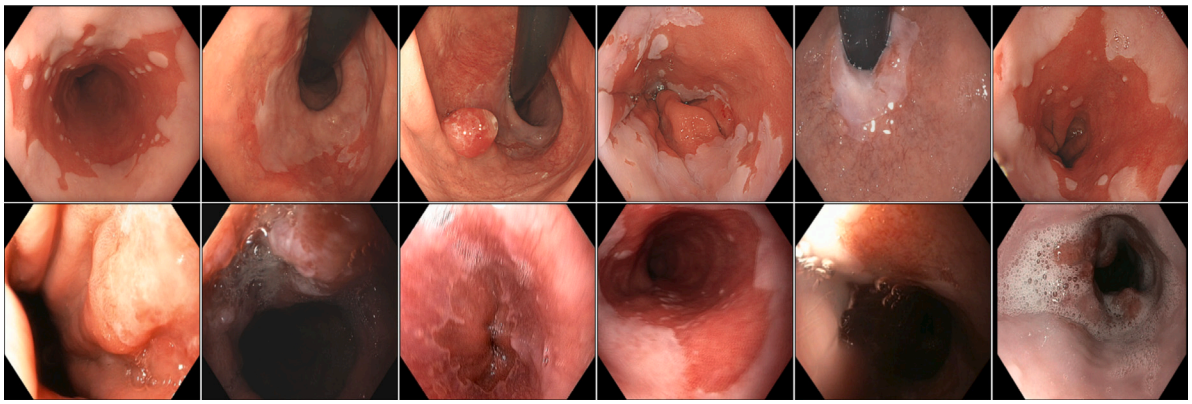

Fig. 1. Images of Barrett's esophagus with a high subjective quality, obtained from expert centers (top row), compared to lower quality images likely to be encountered in clinical practice (bottom row).

In recent years, the focus of research on CADe/CADx systems for endoscopy has shifted towards deep neural network-based (DNN) approaches. These models have demonstrated remarkable success in various computer vision challenges (Cordts et al., 2016; Deng et al., 2009). As a result, DNNs are increasingly being applied in safety-critical domains, such as self-driving cars and medical image analysis. Therefore, it is crucial to ensure that these models are robust and reliable, even under more diverse and sub-optimal imaging conditions faced in real-world applications. Previous studies have already shown that minor image degradation can significantly impact the performance of DNNs (Karahan et al., 2016a; Dodge and Karam, 2016, 2017; Hendrycks and Dietterich, 2019; Pei et al., 2021). As such, it is crucial to develop robust DNN models that can withstand these challenges and operate with reliable performance even under reduced Image Quality (IQ).

In medical imaging, the effect and severity of degraded image quality on DNN performance have been pointed out in earlier studies (Eche et al., 2021; Boone et al., 2023; Maron et al., 2021). The challenges posed by varying image qualities, particularly within DNN-based CADe/CADx models, extend beyond specific imaging modalities. DNN models designed for MRI or CT image analysis must exhibit robustness against different kinds of signal noise and artifacts (Boone et al., 2023; Eche et al., 2021), compared to e.g. DNN models for ultrasound (Jiang et al., 2023). However, all of these studies emphasize the significance of evaluating the robustness against lower IQ prior to the deployment of these models in clinical practice.

This sensitivity to degradations particularly holds for the context of DNN-based CADe/CADx models for GI cancer screening, as the IQ heavily relies on the skill and experience of the endoscopist and the exact imaging system used for screening. Endoscopic CADe/CADx algorithms have gained considerable momentum in recent years, and an increasing number of AI systems have been introduced to the market (Yuba and Iwasaki, 2022; Food and Administration, 2021). Such systems are typically trained and validated on carefully selected pre-processed, high-quality images, while in clinical practice, the IQ is considerably more heterogeneous (see Fig. 1). The IQ can be compromised by various factors such as inadequate lighting, motion blur, and video compression. Additionally, manufacturers typically apply virtual chromoendoscopy technologies to improve the perceived IQ (see Fig. 2). This can have unknown and perhaps undesirable consequences on the performance of CADe/CADx systems. If these models are unable to perform well under such heterogeneous clinical conditions, they become ineffective in clinical practice. Therefore, it is crucial to identify these blind spots and quantify their impact to facilitate the development of robust DNNs, delivering reliable performance under diverse imaging conditions.

In light of these challenges, this study aims to investigate the robustness of the most frequently used DNNs for endoscopic image analysis under realistic and clinically calibrated IQ perturbations. Specifically, we evaluate the performance of DNNs for three endoscopic tasks and various forms of image perturbations to assess their influences on diagnostic outcomes. The chosen image degradation types and their severity are determined in collaboration with two clinical experts to safeguard clinical relevance and realism. The primary evaluation is conducted on three test sets for detecting neoplasia in patients with Barrett's esophagus. To further validate and interpret the results, a subset of the experiments is performed on two publicly available endoscopic datasets, each focusing on different endoscopic tasks. Additionally, this study provides insights into approaches for improving the robustness of DNNs. In summary, the paper presents four key contributions to the field as listed below.

- A comprehensive evaluation of the robustness of popular DNN architectures and training strategies across multiple datasets, using clinically relevant and calibrated levels of IQ degradation.
- A comparison between model performance on synthetically created low image quality datasets and a manually collected test set with lower subjective image quality.
- An analysis showing that peak performance and robustness are not necessarily correlated. Higher peak performance does not guarantee increased robustness, and conversely, higher robustness does not necessarily result in lower peak performance. This demonstration highlights the importance of assessing model performance under various conditions.
- A demonstration of the benefits of in-domain pretraining, resulting in the highest performance on the original test set, the lower subjective image quality test set, and the synthetically created test sets.

This paper offers a comprehensive and systematic evaluation of DNN robustness in endoscopic image analysis, expanding upon our prior research where we first highlighted the impact of lower image quality on DNN performance (Jaspers et al., 2023). The paper develops as follows. In Section 2, the related work on endoscopic analysis and DNN robustness evaluation are discussed. Section 3 presents the experimental setup of this paper. The study makes use of various datasets, which are described in detail in . The training and evaluation metrics are elaborated upon in Section 5. Section 6 explains the robustness evaluation. A comprehensive overview of all the experiments conducted to improve robustness is provided in Section 7. Subsequently, Section 8 presents the results obtained from these experiments. In Sections 9 and 10, these results are further discussed and conclusions are drawn, respectively.

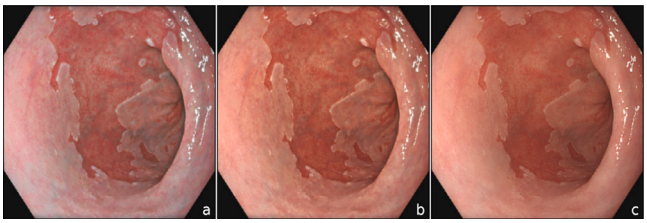

Fig. 2. Examples of different texture and color enhancement imaging (TXI) enhancement settings on the Olympus Evis X1 system. (a) TXI 1, (b) TXI 2, and (c) standard white-light imaging.

2. Related works

2.1. Literature on endoscopic image analysis

The field of GI image analysis has seen a rise in the development of deep learning-based CADe/CADx systems. These systems have been applied to a variety of medical applications, such as the detection of neoplastic lesions in Barrett's esophagus (de Groof et al., 2020; Ebigbo et al., 2019; Hashimoto et al., 2020; Fockens et al., 2023b,a), classification and detection of colorectal polyps (Chen et al., 2018; Byrne et al., 2019; Ozawa et al., 2020) and the detection of gastric lesions (Guimarães et al., 2020; Cho et al., 2019). All these CADe/CADx systems are built upon prior published popular convolution neural network (CNN) architectures. These model architectures include U-net (Ronneberger et al., 2015) and Deeplab (Chen et al., 2016a) with well-known backbone networks, including VGG (Simonyan and Zisserman, 2014), ResNet (He et al., 2016), and EfficientNet (Tan and Le, 2019) variations. However, more recent work on endoscopic image analysis shows state-of-the-art performance on publicly available datasets, such as Kvasir-SEG (Jha et al., 2020), CVC-ColonDB (Tajbakhsh et al., 2016), and CVC-ClinicDB (Bernal et al., 2015), by transformer-based backbone models (Sanderson and Matuszewski, 2022; Chang et al., 2022). Despite the prevalent use of CNNs in CADe/CADx systems for endoscopy, the evolving trends in other computer vision domains indicate the potential for transformer-based models in newer generations, which has motivated us to adopt a transformer-based architecture in the list of selected networks for experiments.

2.2. Robustness evaluation in medical imaging

In the broader scope of medical image analysis, recent studies have addressed the importance of analyzing the robustness of DNNs against lower IQ. Eche et al. (2021) describes stress testing, a form of robustness evaluation, as a strategy to address underspecification in radiology. The authors suggest that robustness evaluation can be designed by modifying medical images or selecting specific testing datasets. Boone et al. (2023) conducted a benchmark study on segmentation models using MRI data that are out-of-distribution and corrupted. They concluded that modern CNN architectures are highly susceptible to distribution shifts, corruptions, and artifacts. Similarly, Maron et al. (2021) demonstrated comparable results in a skin cancer classification task. According to Young et al. (2021), skin cancer-classification models that meet conventional metrics require further validation through computational stress tests to assess clinical readiness. In the field of histopathology, Zhang et al. (2022) evaluated the robustness of DNNs developed for histopathology classification against real-world corruptions, whereas (Zanjani et al., 2019) focused on the robustness against degradations caused by image compression. Islam et al. (2023) evaluated the effect of common image degradation on model performance of both skin cancer-classification as well as chest X-ray scans. They suggest robustness testing should be a standard practice in clinically validating image-based disease detection models. Shen et al. (2020) investigated the impact of image noise for CT lung nodule classification and proposed a training scheme to improve robustness. (Jiang et al., 2023) provided insight into the impact of noise on ultrasound images for breast cancer detection, on the performance of CAD models. All aforementioned studies emphasize the necessity of robustness against reduced image quality prior to clinical implementation of these models.

2.3. Adversarial examples analysis

DNNs have advanced in such a way they can surpass human-level performance on a number of computer vision tasks. Since DNNs are finding their way to real-world applications, the security and integrity of the application pose a great concern. Adversarial examples, i.e., inputs carefully perturbed by small distortions to make a machine learning model fail, show that currently used DNNs are still vulnerable to perturbations, in contrast to the human visual system which can handle such degradations (Szegedy et al., 2013; Papernot et al., 2016; Goodfellow et al., 2015; Kurakin et al., 2017). Adversarial examples are employed to perform worst-case scenario analysis for model robustness. These adversarial examples are specifically modified to foul the attacked model and mostly cannot be transferred to other models (Su et al., 2018). Although such specifically engineered adversarial examples are unlikely to be encountered in real-world use, they have raised concerns about the reliability and robustness of DNNs, especially in safety-critical applications like medical applications (Finlayson et al., 2019; Paschali et al., 2018). According to Ma et al. (2021) these adversarial attacks are easily detectable in medical images. However, several studies demonstrate that applying general, visual distortions can result in a significant performance drop, similar to the effect of adversarial examples. These corruptions are more likely to be encountered in real-world practice, underscoring the importance of enhancing DNNs' resilience to such challenges. (Dodge and Karam, 2017) have shown that the performance of image classification on clean images is comparable between humans and DNNs, but on visually distorted images, humans outperform DNNs by a large margin. These findings are in line with the work of Karahan et al. (2016b), which shows that image degradations can dramatically lower face recognition accuracy. Also, Pei et al. (2021) showed the effect of nine different types of image degradations on multiple popular DNNs and concluded that for all DNNs performance did drop significantly.

2.4. Publicly reported robustness benchmarks

In recent years, improving and understanding model robustness has enjoyed an increasing trend in the computer vision community. To this end, researchers have developed multiple benchmarks to evaluate models on various forms of robustness. The work of Recht et al. (2019), reports new test sets for the CIFAR-10 and ImageNet datasets, containing slightly "harder" images and showing a decrease in performance of up to 15% compared to the general benchmark datasets. Additionally, Hendrycks and Dietterich (2019) created a robustness benchmark for common perturbations and corruptions on ImageNet. Inspired by this, the authors of Michaelis et al. (2020) created similar benchmark datasets, focusing on object detection. Furthermore, Hendrycks et al. (2019a) created a benchmark called ImageNet-A, including natural adversarial examples from other large unlabeled datasets. These benchmarks work great to find trends and suggestions to improve robustness and have been helpful in identifying model weaknesses. However, it is important to note that not all improvements in general computer vision benchmarks, such as ImageNet or Cityscapes (Deng et al., 2009; Cordts et al., 2015), translate well to medical image analysis. Medical images differ from natural images in several important ways, and therefore the general robustness benchmarks may not fully capture the challenges faced in medical image analysis. Previous endoscopic computer vision challenges have addressed the issue of diminished image quality and generalization across different endoscopic modalities (Ali and Ghatwary, 2022).

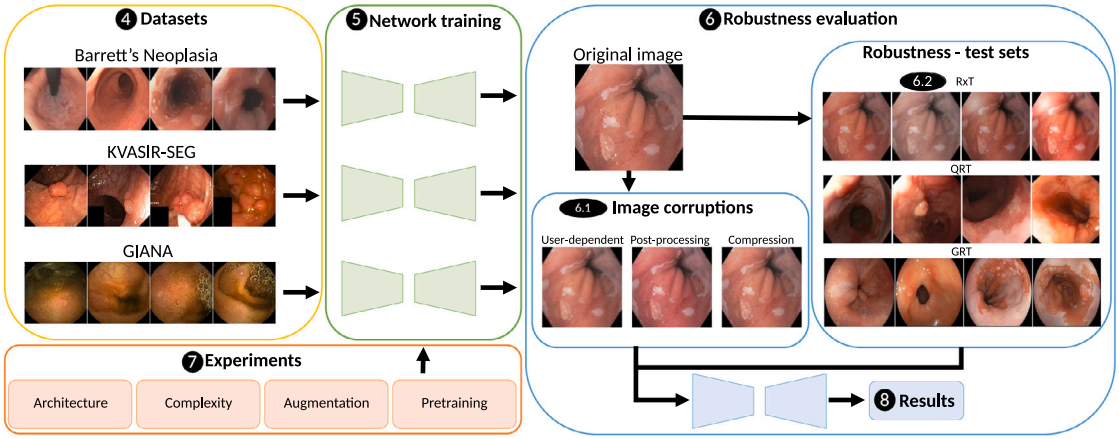

Fig. 3. Visualization of the experimental setup. The numbers in the black dots indicate the corresponding section in the paper. At the left examples of the three different datasets used for model training are shown. Then test images of each of the datasets are perturbed by specific image degradations for evaluation. The robustness capabilities of the different models are measured based on the performance of the synthetic degraded test sets (RxT), the manually selected real test set (QRT), and a generalization test set (GRT). At the bottom, the four different experiments conducted to increase robustness are visualized.

2.5. Work on robustness enhancements

Attempts have been made to increase model robustness and perform better on the aforementioned robustness benchmarks. The work of Hendrycks and Dietterich (2019) and Xie and Yuille (2020) suggest that a larger model size improves robustness on corrupted images and adversarial attacks. However, larger models also tend to need more data to converge. This becomes particularly challenging in the context of medical image analysis, where data scarcity is prevalent as compared to natural image datasets. As a consequence, larger models may not always yield the expected increase in performance. Moreover, the addition of self-attention layers to models has been shown to improve robustness to common corruptions (Hendrycks et al., 2021b; Bhojanapalli et al., 2021; Xie et al., 2021a) and adversarial examples (Benz et al., 2021; Shao et al., 2022; Ghaffari Laleh et al., 2022). Furthermore, some research shows that diverse data augmentation can improve robustness (Geirhos et al., 2019; Yun et al., 2019). However, the research of Vasiljevic et al. (2016) finds that fine-tuning on blurred images can marginally decrease performance on the original dataset. Lastly, both the works of Hendrycks et al. (2019a) and Orhan (2019) conclude that training on larger and more diverse datasets improves model generalization and robustness. In Hendrycks et al. (2019b) the authors show that a simple self-supervised learning framework increased the robustness of DNNs compared to solely supervised training. Notably, these results were echoed by Navarro et al. (2021) and Srinivasan et al. (2021) who showcased enhanced performance of self-supervised learning specifically in handling imperfect data within the medical domain. Despite several strategies proposed in natural image analysis, a comprehensive systematic evaluation encompassing various forms of IQ degradation and mitigation approaches is notably absent, particularly in the realm of endoscopic analysis.

3. Experimental setup

The previous discussion on related work has revealed that although robustness evaluation in natural image processing is a well-known topic, this area is largely unexplored in endoscopic image analysis, where robustness is inherently required for useful clinical applications. Fig. 3 depicts the major steps of this study. Three different datasets have been used for model training as indicated. Up to this point, there is no existing dataset containing both low-quality and their corresponding high-quality images. Therefore, the experimental setup contains a part for generating various types of synthetic image degradations to evaluate the robustness of the models. Additionally, for realistic testing, a specific dataset has been collected with practical clinical examples containing lower subjective image quality. Four additional experiments are carried out to find trends that lead to better robustness.

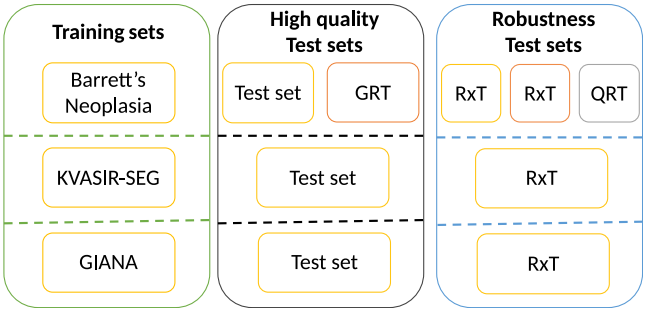

Fig. 4. Overview of all datasets. The applied colors to delineate the GRT and QRT test sets highlight distinctions in scope manufacturer and subjective IQ, respectively, in contrast to the original test set.

4. Datasets

Fig. 4 provides a schematic overview of all datasets in this study. We use a primary dataset comprising images of Barrett's neoplasia for training, validation, and testing. The general test set is employed to evaluate the effect of individual clinically calibrated corruptions described in Section 6.1. This test set will also be degraded to generate the RxT test sets explained in Section 6.2. To ensure comprehensive evaluation, we also incorporate an internal test set comprising manually selected subjective low-quality images. Furthermore, an external Barrett's neoplasia test set with subjective high-quality images is used for generalization testing. This test set is also synthetically degraded, exploiting the scheme described in Section 6.2.

Lastly, we conduct a subset of experiments on two publicly available datasets, specifically designed for other endoscopic-related tasks. The test sets will also be transformed as described in Section 6.2. The following sections provide detailed specifications for each of the applied datasets in this study.

4.1. Barrett's neoplasia: training set

The dataset used for training includes 2752 images of Barrett's esophagus neoplasia (NEO) (628 patients) and 7595 images of non-dysplastic Barrett's esophagus (NDBE) (1095 patients), all histopathology confirmed. Delineations of neoplasia are obtained from 14 Barrett's experts, where at least 2 experts delineated the same image. For each image, the experts delineated the largest area that is suspected to be neoplasia (lower likelihood) and the area within the lower likelihood that stands out more profoundly (high likelihood). To achieve a minimal level of consensus, a third expert endoscopist is invited in case the two high-likelihood delineations obtained less than 30% agreement in terms of Dice Score. Subsequently, the two delineations among the three experts that achieve the highest overlap are used for further ground-truth processing. For training and evaluation purposes, a consensus ground-truth mask is constructed using (1) the union of high-likelihood delineations (2) the intersection of low-likelihood delineations, and (3) the union of the areas (1) and (2). The data was obtained in 17 different clinical centers using the Exera II, Exera III, and EVIS X1 production lines (Olympus Corp., Tokyo, Japan). De-identification of the images is performed in line with the General Data Protection Regulation (EU) 2016/679.

4.2. Barrett's neoplasia: general test set

The general test set is enriched with challenging subtle neoplasia cases based on the clinicians' input. This set comprises 196 NEO images from 108 patients and 400 NDBE images from 161 patients. Notably, this test set is created using patients from the same 17 clinical centers and endoscopes as the training set. All imagery in the general test set is of high subjective IQ. To avoid data leakage and intra-patient bias, a strict split on a patient basis is implemented between the test and training set.

4.3. Barrett's neoplasia: subjective quality robustness test set (QRT)

To evaluate the impact of subjective image quality, We have established an additional internal test set, comprising 277 neoplasia images (115 patients) and 65 non-dysplastic Barrett's esophagus (NDBE) images (53 patients). These images were meticulously selected by two clinical researchers to ensure their inferior subjective image quality. The selection criteria for inferior quality included aspects such as illumination, blur, image resolution, as well as the presence of mucus and bubbles. None of the patients included in the internal QRT test sets was used for model training to avoid data leakage and intra-patient bias.

4.4. Barrett's neoplasia: external generalization test set (GRT)

For external validation, a test set was used consisting of 209 images of NEO (134 patients) and 248 images of NDBE (121 patients). This data was acquired using the ELUXEO 7000 endoscopy system by Fujifilm, based in Tokyo, Japan. It is important to note that this system is produced by a different manufacturer than the one used in the training dataset. The images were collected from four different international medical centers. The dataset was carefully curated to ensure high subjective image quality, with a significant portion of the data featuring subtle neoplastic lesions. As a result, the test set is comparable in terms of quality and composition to the main test set. Neoplasia delineations are obtained from 8 Barrett's experts, where at least 3 experts delineated the same image. The ground truth for this dataset was based on the area of neoplasia that the majority of experts delineated.

4.5. Public datasets

To warrant the generalizability of our findings, we use two public datasets that are intended for other gastrointestinal applications. The first dataset is the Kvasir-Seg dataset (Jha et al., 2020), which allows us to examine the performance of colonic polyp localization under image degradations. Since the dataset was not originally partitioned by the authors, we decided to randomly divide it as follows: 75% of the images (750 images) were allocated for training, and 25% (250 images) were set aside for testing. We ensured that the split maintained the separation of images belonging to different patients.

For the second additional endoscopic task, we utilize the Gastrointestinal Image ANAlysis (GIANA) Grand Challenge dataset from the Endoscopic Vision Challenge at the MICCAI 2017 conference, which involves joint classification and localization. This dataset comprises 1812 wireless capsule images from the small bowel, categorized as normal, inflammatory, and angiodysplasia images. Each image containing inflammation or angiodysplasia is accompanied by a segmentation mask that indicates the lesion's location. Consequently, this dataset is suitable for tackling the multi-task problem of 3-class classification and binary segmentation. The dataset is randomly divided into training (75%) and test (25%) sets, resulting in 1354 and 452 images, respectively. Throughout the splitting process, we maintain a strict separation at patient level, while the class distributions remain similar in each set.

5. Network training and evaluation metrics

This section provides an overview of the general training approach, used for all models in this study. More specific details on the architectures, data augmentation techniques, and pretrained weights are described in Section 7. Training parameters such as learning rate and number of epochs are kept fixed for all experiments and chosen based on experimental findings for all models to converge. The batch size is maximized for the available memory to fit the largest model.

The training process involves a five-fold cross-validation at patient level of the training data. First, the encoder with ImageNet-pretrained weights is frozen and the decoder is trained on all data (Deng et al., 2009). After the initialization of the decoder in stage one, all weights of the encoder are unfrozen. During the second stage, the complete network is trained on all data. During training, the Adam optimizer is used with a learning rate of 1×10^{-3} and 1×10^{-5} for the first and second stage, respectively, while $(\beta_1, \beta_2) = (0.9, 0.99)$ remain constant during both stages (Kingma and Ba, 2014). The learning rate is reduced by half after 10 consecutive epochs without a decrease in validation loss, while the early stopping criteria is applied if there is no decrease in validation loss over 20 consecutive epochs. All input images are resized to a resolution of 256×256 pixels. A batch size of 16 images and the binary cross-entropy loss are used to train the network. The proposed methods are implemented in Python using the PyTorch framework and experiments are executed on a TITAN RTX GPU (NVIDIA Corp., CA, USA).

The models are evaluated using a detection-by-segmentation approach for the Barrett's Neoplasia datasets, where a detection label is assigned if the network's segmentation prediction is above the default 0.5 threshold. To assess the classification accuracy, we use the Area under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve as the evaluation metric, which provides a comprehensive assessment of the model's sensitivity and specificity. In clinical practice, lesions are typically defined by expert endoscopists using optical chromoscopy with a magnified view, after initial detection. However, this is not the intended use of this CAD system and thus, pixel-precise delineation metrics were not evaluated and are not relevant in this study.

For evaluating the models trained on the Kvasir dataset, we employ the Dice similarity metric. This metric quantifies the overlap

Table 1
Overview of the synthetic image perturbations evaluated in this study.

User-dependent corruptions	Image acquisition - processing changes	Compression artifacts
Motion blur	Contrast	Reducing resolution
Local defocus blur	Saturation	JPEG compression
Overexposure	Hue	JPEG2000 compression
	Brightness	
	Sharpness	

between the predicted segmentation mask (thresholded at 0.5) and the ground-truth mask. The evaluation of the models trained on the GIANA dataset encompassed both classification and segmentation levels. The employed metrics were the AUC for classification assessment and the DICE coefficient for segmentation evaluation.

6. Robustness evaluation

As mentioned, the evaluation of robustness is particularly extensive for the Barrett's neoplasia task. The models undergo evaluation using single-image perturbations, encompassing all levels of severity. This evaluation incorporates both quantitative and qualitative assessment to ensure a comprehensive evaluation. Furthermore, the models are evaluated on synthetic robustness test sets, which involve combining multiple perturbations per image. Additionally, the findings from these tests are compared with the real-world QRT test set to provide a comprehensive analysis. In contrast, for the other two endoscopic tasks, the evaluation of robustness is limited to the synthetic robustness test sets.

6.1. Synthetic image corruptions

This study assesses model robustness by detailing perturbations to 11 types of synthesized corruptions. We grouped the distortions into 3 categories: (1) user-dependent corruptions, (2) image acquisition and processing changes, and (3) artifacts induced by compression. Table 1 presents an overview of the perturbations, a more detailed table including the parameters of each corruption can be found in Appendix A.3 in the supplementary materials. Additionally, visual examples of each corruption and corresponding severity levels, are presented in Figs. 15, 16, and 17.

To ensure clinical relevance, 10 levels of severity were calibrated by two clinical experts. The first 2 levels simulate variations commonly found in high-quality endoscopic image datasets. As long as critical features such as mucosal pattern and vasculature are clearly identifiable, the image is considered expert quality. When this is no longer the case, but the image quality still falls within the spectrum of quality variations one might encounter in everyday clinical practice, the image quality is considered 'non-expert quality'. This quality level is simulated within the severity Levels 3 to 5. The remaining Levels 6 to 10 are designed to determine the model's breaking point. They may be classified as extreme, but since the severity levels are subjective, in some cases Levels 6 to 8 could still be clinically relevant.

To create consistent perturbation behavior within the range, the severity of the corruptions with respect to hue, saturation, brightness, and contrast is categorized into 10 ascending and 10 descending levels. Additionally, to facilitate a reasonable comparison between JPEG and JPEG2000 compression, the severity factor is based on the same compression factor.

6.2. Robustness test sets (RxT)

To further evaluate model robustness against lower IQ, we exploit the QRT dataset described in Section 4.2 and we create three new, more heterogeneous test sets (R2T, R5T, and R10T) using the 11 corruptions introduced in the previous section. The same variations of robustness test sets were created for the GRT test set.

The synthetic robustness test sets, R2T, R5T, and R10T, encompass a range of randomly applied corruptions compounded on individual images. The test set R2T comprises 1 to 5 randomly applied corruptions with severity levels 1 or 2. R5T consists of a comparable number of compounded corruptions with severity levels 3 to 5. Likewise, set R10T involves 1 to 5 random corruptions with severity levels 5 to 10. This iteration is performed five times for all images in the test set. Fig. 5 shows several illustrative examples from the newly created synthetic robustness test sets and real test sets. The code for generating the synthetic test sets is publicly available at https://github.com/BONS-AI-VCA-AMC/Robustness.

7. Experiments

To enhance the robustness of the models, we have explored four different strategies in this study. In particular, we evaluate the impact of varying (1) model architectures, (2) model complexity, (2) data augmentation techniques, and (4) pretraining on the performance of the models. The experiments on model architecture and model complexity are purely based on the encoder. All of the encoders are incorporated into a U-Net architecture, using weights derived from initializing with ImageNet-1k (Deng et al., 2009; Ronneberger et al., 2015). The models are created and trained using PyTorch and the library PyTorch segmentation models (Paszke et al., 2019; Iakubovskii, 2019).

7.1. Architecture

In terms of architecture, we evaluated five different encoders: VGG-19, ResNet-50, DenseNet-161, EfficientNet-B5, and MiT-B2 (Simonyan and Zisserman, 2014; He et al., 2016; Huang et al., 2017; Tan and Le, 2019; Xie et al., 2021b). These encoders have been selected because they all have a comparable complexity in terms of the number of parameters. Moreover, at the moment of their publication, each encoder was producing state-of-the-art performance on multiple computer vision benchmarks. These two features enable us to derive a trend from the past years with respect to robustness, depending on the applied model architecture.

7.2. Complexity

Similarly, to find a trend in dependence on model complexity, the impact of encoder depth is evaluated by using various sizes of the ResNet encoder: ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152 (He et al., 2016). These encoders have 11, 21, 23, 42, and 58 million parameters, respectively. Table 2 shows an overview of the encoders used for the architecture and complexity experiments, including their number of parameters, top-1% accuracy on ImageNet, and year of publication.

7.3. Augmentation

Data augmentation is mostly considered to improve the generalization of a network to unseen data. In our case, we evaluate *extended* data augmentation with relation to model robustness, in three different ways. We first compare training with several levels of standard data augmentation: no data augmentation, light data augmentation (including rotations, flipping, and cropping), and heavy data augmentation (light data augmentation with additional changes in contrast, brightness, hue, saturation, and addition of Gaussian noise). Secondly, we explore two advanced augmentation methods, Cutout and Cutmix (DeVries and Taylor, 2017; Yun et al., 2019), which are reported to improve model robustness on natural images. Thirdly, to further assess the effectiveness of data augmentation, we compare models trained with adversarial generated images based on the approach proposed by Gu et al. (2022).

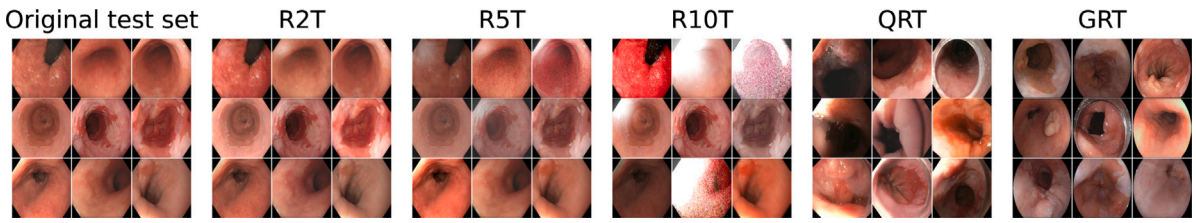

Fig. 5. Random visual examples of test sets used for evaluation. From left to right: the original, R2T, R5T, R10T, QRT, and GRT sets. Notably, the similarity between the clean test set and the R2T set is caused by the subtle nature of these degradation levels.

Table 2
Overview of the evaluated encoders in this study, including their parameter count, ImageNet Top-1% accuracy, and publication year.

Encoder	Parameters [M]	ImageNet Top-1% ACC	Year
VGG-19	20	72.6	2014
ResNet-50	23	76.1	2015
DenseNet-16	26	77.7	2016
EfficientNet-B5	28	83.6	2019
MiT-B2	24	81.6	2021
ResNet-18	11	69.8	2015
ResNet-34	21	73.3	2015
ResNet-50	23	76.2	2015
ResNet-101	42	77.4	2015
ResNet-152	58	78.3	2015

7.4. Pretraining

Pretraining models on large data sets of natural images, such as ImageNet, has become the standard for endoscopic image analysis. In terms of performance, this method generally shows superior performance compared to training from scratch. In this study, we explore the impact of pretraining on both ImageNet and GastroNet, a large-scale in-house endoscopic dataset containing over 5 million images from the gastrointestinal tract, on model robustness. We have used the ImageNet-1K supervised pretrained weights since these are still the most generally used and often considered the gold standard. Pretraining on GastroNet is performed in a self-supervised manner, using the novel "Self-Distillation with No Labels" (DINO) approach (Caron et al., 2021). The training details of the GastroNet pretraining are provided in the Supplementary Materials, in Table 7.

8. Results

The empirical results of user-dependent, image acquisition and processing, and compression artifact corruptions, are all shown on the Barrett's neoplasia dataset. The quantitative results on the GRT, QRT, and synthetic robustness test sets (RxT) are provided in Section 8.4. Additionally, the results obtained from the two public datasets are presented in Section 8.5.

8.1. User-dependent degradations

Fig. 6 displays the mean performance of different models and training strategies across 10 levels of severity for defocus-blur, motion-blur, and overexposure. The dotted line represents the performance of the ResNet-50 encoder trained with ImageNet-1K-pretrained weights and light data augmentation, which serves as a reference in all four experiments. The results indicate that all models, regardless of architecture and complexity (Fig. 6, Row (a) and (b)), show similar performance degradation under increasing levels of defocus blur, with a decrease of approximately 1%–2% at severity Level 2 and 4%–5% at severity Level 5. Including adversarial examples during training results in the worst performance, with a decline of 2.6% and 6.5% at severity Levels 2 and 5, respectively, for defocus blur (Row (c)). Pretraining on GastroNet seems to make the models more robust against small levels of defocus blur, with a decrease in performance of only 2% below the dotted reference line at severity Level 5 (Row (d)). However, compared to the performance on the original test set, even pretraining with GastroNet results in a 5.0% decrease.

Overexposure (Fig. 6, 2nd column) does not significantly affect performance up to severity Level 2, but results in a comparable decrease of about 3%–4% for all encoders at severity Level 5 (Row (a) and (b)). However, using CutOut for augmentation during the training procedure limits the performance degradation to 2.0% only (Row (c)). Additionally, overexposure does not seem to affect the performance of the encoder pretrained on GastroNet, with the performance at severity Level 5 still being above the reference performance and showing a minor 1.6% decrease only, compared to the performance on the original test set (Row (d)).

Regarding motion-blur (Fig. 6, 3rd column), the VGG-19 encoder experiences the largest decline in performance, with an initial AUC of 0.95, decreasing to 0.84 (−11.6%) and 0.65 (−31.5%) at severity Levels 2 and 5, respectively (Row (a)). The MiT-B2 encoder degrades with the least amount of performance, with an initial AUC of 0.93, decreasing to 0.90 (−2.8%) and 0.82 (−11.8%), at severity Levels 2 and 5. In terms of complexity (Row (b)), ResNet-18 shows the largest decrease in performance. However, the difference with respect to the other complexities is not significant. Furthermore, the inclusion of heavy data augmentation results in the best performance at Levels 2 and 5 (Row (c)). The same accounts for pretraining with GastroNet (Row (d)), which shows the best results between Levels 2 and 5. However, the trend in performance decrease is comparable to the decrease with ImageNet-1-pretrained weights.

In addition, Fig. 7 provides a visual representation of the reference model's prediction up to severity Level 5 for the user-dependent degradations. The heatmap represents the raw segmentation output of the decoder. The figure reveals that both defocus-blur and overexposure lead to a vanishing certainty of the prediction, and at severity Level 5, the neoplasia is nearly missed. Motion-blur affects the prediction even earlier, with a once-pronounced neoplastic case, even not recognized at severity Level 2. These findings are consistent with the overall numerical results, indicating that motion-blur has a significant impact on model performance. Further examples in Fig. 13 in the Supplementary Materials illustrate the counter effect of not missing lesions but creating undesirable false positives.

8.2. Image acquisition and processing changes

We investigated the impact of various corruptions related to image acquisition and processing changes on model performance, of which the results are shown in Fig. 8. At severity Level 5, applying a sharpness filter causes a considerable decrease in the initial AUC of the ResNet-50 encoder from 0.92 to 0.84 (−8.2%, 1st column). The same trend is observed for all ResNet models (Row (b)), without considerable decrease at severity Level 2 and a decrease of 7%–8% at severity Level 5. Increasing contrast has no significant effect on any of the models up to severity Level 5, while decreasing contrast up to Level −5 results in a decrease of 10.7% for the VGG-19 encoder and 3.4% for

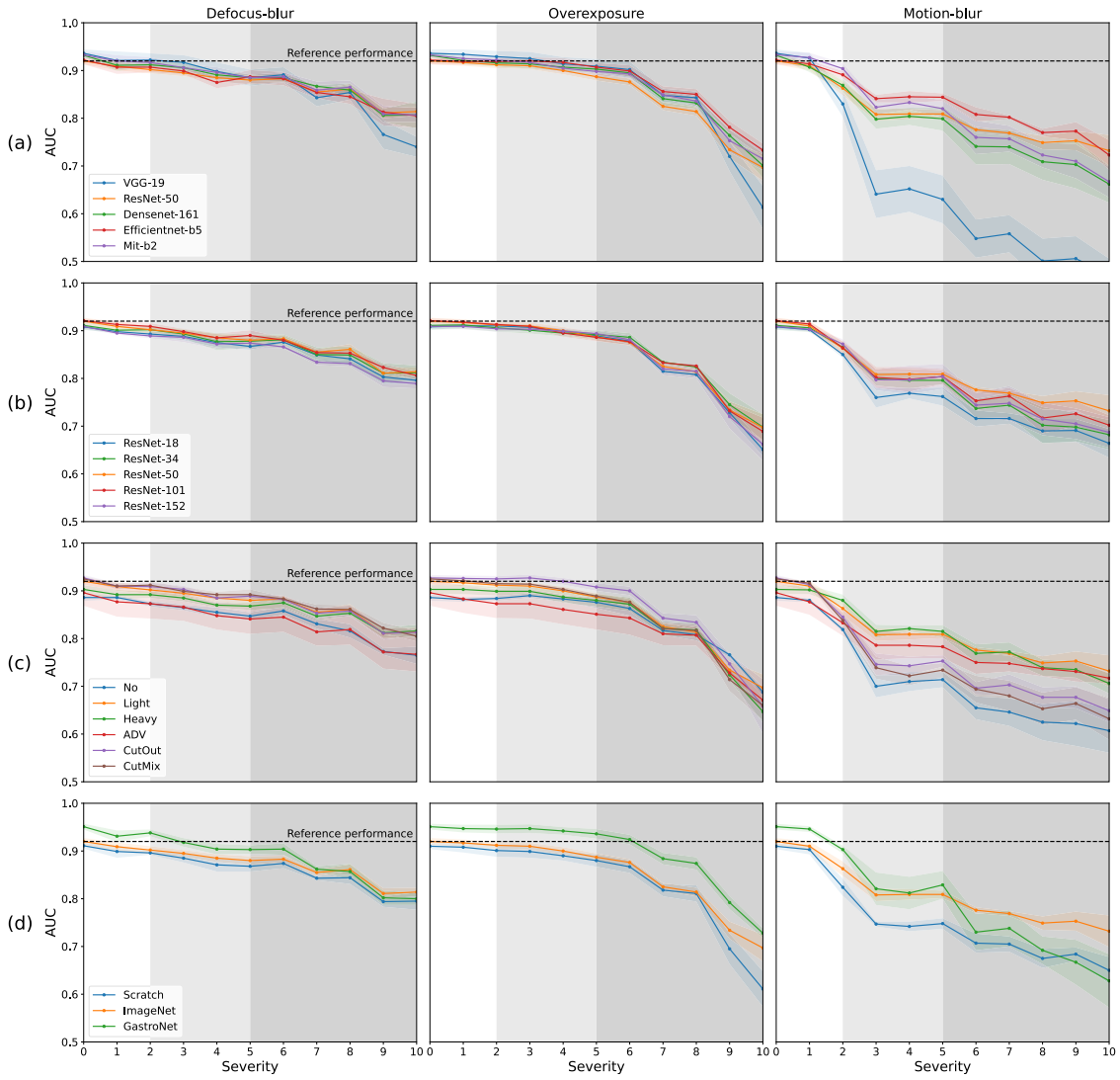

Fig. 6. Model performance evaluation for user-dependent degradations for all four classes of experiments. From top to bottom, (a) shows the experiment focusing on the network architecture. (b) portrays the experiments on encoder complexity. (c) and (d) depict the experiments on data augmentation and pretraining, respectively. The white zone in each figure corresponds to severity Levels 1 and 2, which represents the amount of limited degradation typically present in high-quality endoscopic image datasets. The light gray area in each figure consists of Levels 3–5, which corresponds to the amount of degradation expected in clinical practice. The dark gray area includes Levels 5–10, representing increasingly unrealistic levels of image degradation. The black dotted horizontal line in all figures depicts the performance of the reference model on the original test set.

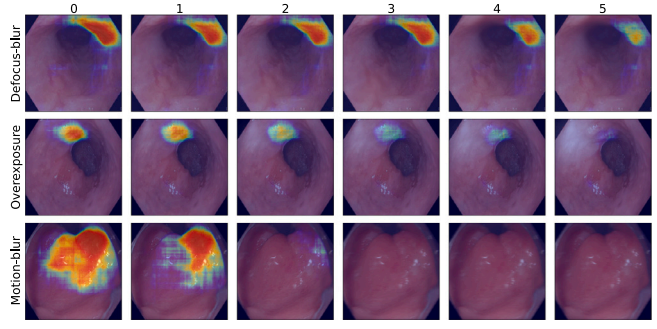

Fig. 7. Visual examples of the impact of user-dependent corruptions on the reference model's prediction, ranging from severity Level 0 (clean image) to Level 5 (clinically relevant). Severe degradations lead to missing neoplastic lesions. The heatmap indicates the raw segmentation output.

the ResNet-50 encoder, respectively. Notably, pretraining on GastroNet proves to be robust against changes in contrast, without a decrease in performance observed from severity Levels −5 to +5.

Decreasing brightness has less impact on model performance compared to increasing brightness (Fig. 8, 3rd column). Most models show a decrease of 7%–10% in performance at brightness severity Level 5. The lowest decrease is observed for the GastroNet-pretrained encoder with 6.2% at Level 5. Regarding saturation (4th column), most models are found to be robust against changes within clinically relevant ranges, except at severity Level −5, where a decrease in saturation results in a decline of 3.7% in the performance of the reference model. On the other hand, changes in hue have a noteworthy impact on all encoders between Levels −5 and −1, except for the encoder pretrained on GastroNet. At Levels −2 and −5, the decrease is between 1%–2% and 4%–8% for all encoders, however, pretraining on GastroNet results in a 0.5% and 1.1% decrease (5th column), respectively. Interestingly, including changes in hue during training (heavy data augmentation) does not make the models more robust against this form of corruption, resulting in a decrease of 9.7% at severity Level −5 (Row c).

Additionally, Fig. 9 presents visual examples of the impact of image acquisition and processing changes. The examples demonstrate that

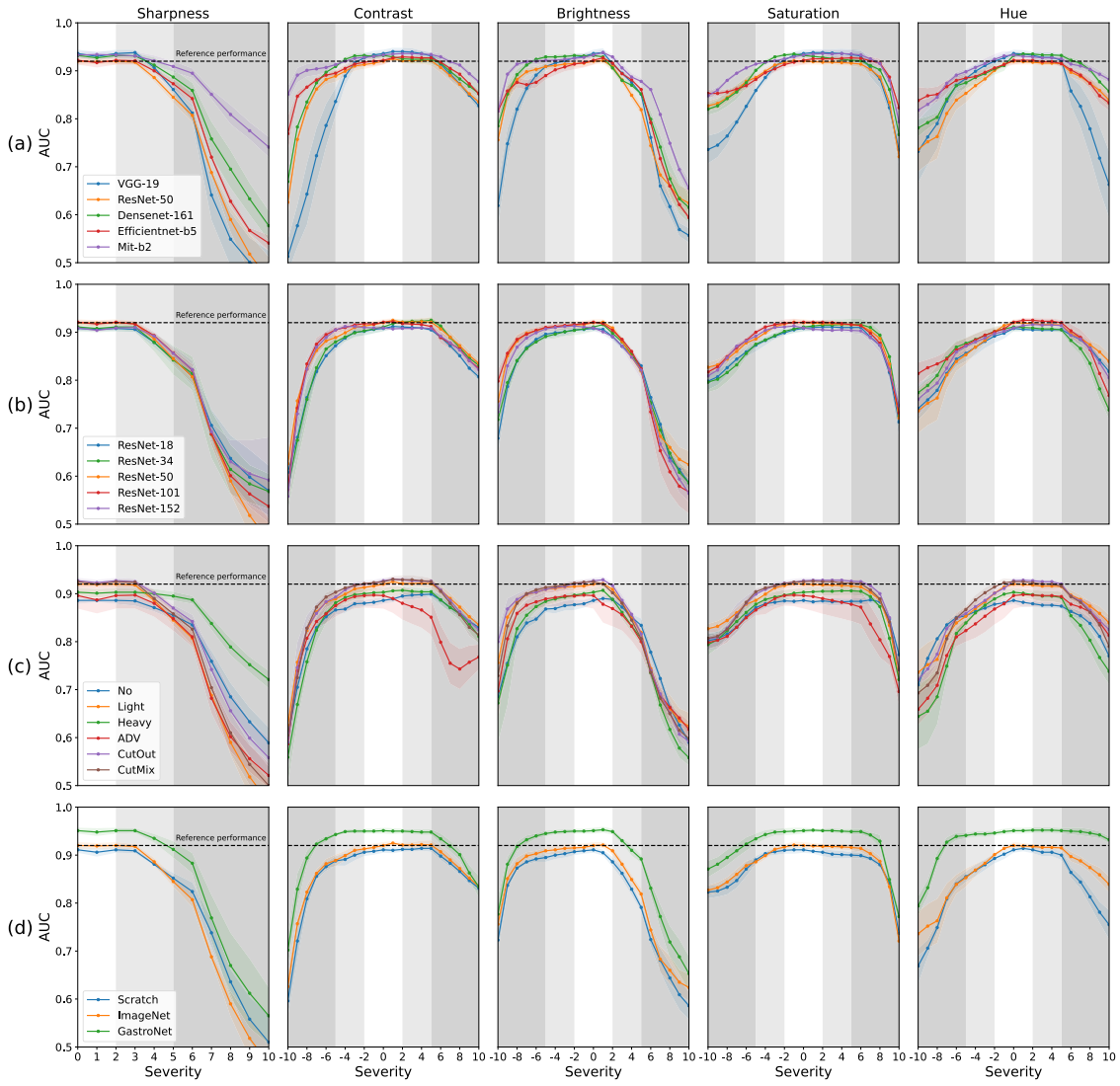

Fig. 8. Model performance evaluation for image acquisition and processing changes for all four classes of experiments. From top to bottom, (a) shows the experiment focusing on the network architecture. (b) portrays the experiments on encoder complexity. (c) and (d) depict the experiments on data augmentation and pretraining, respectively. The white zone in each figure corresponds to severity Levels +1 and +2 and Levels −1 and −2, which represents the amount of limited degradation typically present in high-quality endoscopic image datasets. The light gray area in each figure consists of Levels +3 to +5 and Levels −3 to −5, which corresponds to the amount of degradation expected in clinical practice. The dark gray area includes Levels +5 to +10 and Levels −5 to −10, representing increasingly unrealistic levels of image degradation. The black dotted horizontal line in all figures depicts the performance of the reference model on the original test set.

changes in saturation, brightness, and contrast can have a significant effect on model predictions, potentially leading to missed neoplastic cases in subtle instances. Conversely, an increase in contrast and saturation may cause more false positives. In the Supplementary Materials, Fig. 14 provides further examples of such false positives. Furthermore, Fig. 10 shows that sharpness filters, although they may increase the perceived IQ, can result in more false positives.

8.3. Compression corruptions

The impact of image compression on model performance can be observed in Fig. 11. We have found that a reduction in resolution does not significantly impact performance up to Level 5 (1st column). However, a more extreme reduction results in a decreased performance, with the VGG-19 encoder dropping to 0.64 (−32.0%) at severity Level 8, compared to 0.83 (−10.1%) and 0.78 (−16.3%) for the Efficientnet-B5 and MiT-B2 encoders.

The use of heavy data augmentation improves performance at severity levels higher than 5, with the reference model achieving 0.74 (−19.3%) at Level 8, compared to 0.84 (−7.1%) with training including heavy data augmentation (Row c). The same effect is observed with JPEG compression, however, JPEG2000 compression leads to severe performance degradation at Levels 2 and 5. At Level 2, all models experience a drop of 4%–9%. At Level 5, this drop increases to 13%–17% for all models except for the VGG-19 encoder, which drops to 0.62 (−34.0%). Interestingly, increasing complexity does not make the models more robust against JPEG2000 compression. Adding heavy data augmentation results in the least decrease in performance at both Levels 2 and 5 (2.1% and 9.6%). Pretraining with GastroNet is effective in mitigating degradation at Level 2 (2.3%), but at higher levels, the performance decreases more rapidly than with the encoder pretrained on ImageNet (23.1% compared to 16.6% at severity Level 5, Row d).

Visual examples of the effect of image compression on model predictions can be observed in Fig. 12. Although both reducing resolution and JPEG compression do not lead to a significant decrease in AUC up to Level 5, the examples illustrate that both corruptions do have a considerable effect on model prediction. The effects of JPEG2000 compression are more severe since the examples show an obvious neoplastic region being missed at severity Levels 4 and 5. These results emphasize the

Table 3
Performance on the Barrett's esophagus test sets, including the original test set, R2T, R5T, R10T, and QRT test set. AUC values are presented as mean ± std.

Encoder	Augmentation	Pretraining	Test set $AUC \uparrow$	R2T $AUC \uparrow$	R5T $AUC \uparrow$	R10T $AUC \uparrow$	QR Set $AUC \uparrow$	Parameters	ImageNet $Top1ACC$
VGG-19	Light	ImageNet	**0.93 ± 0.01**	0.90 ± 0.01	0.78 ± 0.01	0.58 ± 0.01	0.83 ± 0.03	20M	72.6
ResNet-50			0.91 ± 0.01	0.89 ± 0.01	0.80 ± 0.01	0.61 ± 0.01	0.86 ± 0.02	23M	76.1
DenseNet-16			**0.93 ± 0.01**	0.90 ± 0.01	0.83 ± 0.01	0.64 ± 0.01	0.88 ± 0.01	26M	77.7
EfficientNet-B5			0.92 ± 0.01	0.90 ± 0.01	**0.84 ± 0.01**	**0.66 ± 0.01**	0.87 ± 0.00	28M	83.6
MiT-B2			**0.93 ± 0.01**	**0.91 ± 0.01**	**0.84 ± 0.01**	0.65 ± 0.01	**0.89 ± 0.02**	24M	81.6
ResNet-18	Light	ImageNet	0.90 ± 0.01	0.87 ± 0.01	0.80 ± 0.01	0.61 ± 0.01	0.86 ± 0.02	11M	69.8
ResNet-34			0.91 ± 0.01	0.88 ± 0.01	**0.81 ± 0.01**	**0.63 ± 0.01**	**0.88 ± 0.01**	21M	73.3
ResNet-50			0.91 ± 0.01	**0.89 ± 0.01**	0.80 ± 0.01	0.61 ± 0.01	0.86 ± 0.02	23M	76.2
ResNet-101			**0.92 ± 0.00**	**0.89 ± 0.01**	**0.81 ± 0.01**	0.62 ± 0.02	0.86 ± 0.04	42M	77.4
ResNet-152			0.90 ± 0.01	0.88 ± 0.01	**0.81 ± 0.01**	0.62 ± 0.01	0.87 ± 0.03	58M	78.3
ResNet-50	No	ImageNet	0.89 ± 0.01	0.85 ± 0.01	0.76 ± 0.01	0.59 ± 0.01	0.85 ± 0.02	23M	76.1
	Light		0.92 ± 0.01	0.88 ± 0.01	0.80 ± 0.01	0.61 ± 0.01	**0.86 ± 0.02**		
	Heavy		0.90 ± 0.01	0.88 ± 0.01	**0.82 ± 0.01**	**0.62 ± 0.01**	0.82 ± 0.02		
	ADV		0.90 ± 0.03	0.84 ± 0.01	0.75 ± 0.04	0.57 ± 0.01	0.84 ± 0.02		
	CutOut		**0.93 ± 0.01**	**0.89 ± 0.01**	0.79 ± 0.01	0.60 ± 0.01	**0.86 ± 0.02**		
	CutMix		**0.93 ± 0.01**	0.88 ± 0.01	0.79 ± 0.01	0.59 ± 0.01	0.85 ± 0.01		
ResNet-50	Light	No	0.91 ± 0.01	0.87 ± 0.01	0.77 ± 0.01	0.58 ± 0.01	0.85 ± 0.01	23M	76.1
		ImageNet	0.91 ± 0.01	0.89 ± 0.01	0.80 ± 0.01	0.61 ± 0.01	0.86 ± 0.02		
		GastroNet	**0.95 ± 0.01**	**0.93 ± 0.01**	**0.86 ± 0.02**	**0.65 ± 0.01**	**0.88 ± 0.02**		

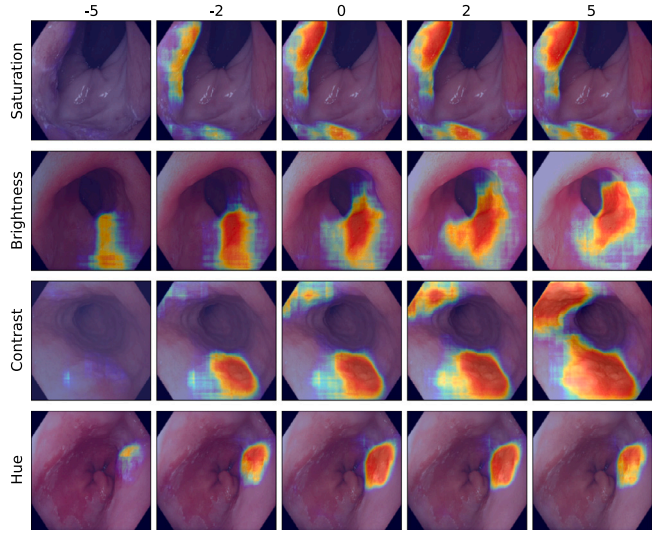

Fig. 9. Visual examples of the impact of image acquisition and processing corruptions on the reference model's prediction. The severity levels (−5, −2, 0, 2, 5) are shown with corresponding heatmaps representing the raw segmentation output.

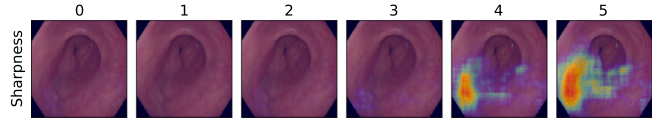

Fig. 10. Visual example of a false positive prediction resulted from sharpening post-processing step. The heatmap overlay represents the predicted segmentation model probabilities.

importance of considering the impact of lossy compression on model performance in the context of medical image analysis.

8.4. Robustness test sets

The following section presents the robustness performance of models on the Barrett's esophagus test sets, as shown in Table 3. The performance is evaluated on the robustness variations of the test set: R2T, R5T, R10T, and the QRT test set. The best-performing models are highlighted in bold.

On the original test set, the VGG-19, MiT-B2, and DenseNet-16 encoders achieve the highest AUC of 0.93 ± 0.01. However, when evaluated on the R2T and R5T sets, the MiT-B2 encoder outperforms others with AUC scores of 0.91 ± 0.01 and 0.84 ± 0.01, respectively. Notably, the VGG-19 encoder has the lowest AUC on both the R5T and R10T sets, with scores of 0.78 ± 0.01 and 0.58 ± 0.01. Moreover, there are no significant differences larger than 0.02 in complexity among all models, across all levels of corruption.

Furthermore, incorporating CutOut as data augmentation yields the best results on the original test set and R2T, with AUC values of 0.93 ± 0.01 and 0.89 ± 0.01. However, using heavy data augmentation leads to the highest performance on the R5T and R10T sets, with AUC scores of 0.82 ± 0.01 and 0.62 ± 0.01, respectively. Most notably, the largest improvement across all test sets is achieved by pretraining on GastroNet, outperforming the use of pretrained weights from ImageNet or training from scratch. Pretraining on GastroNet results in AUC scores of 0.95 ± 0.01, 0.93 ± 0.01, 0.86 ± 0.02, and 0.65 ± 0.01 for the original test set, R2T, R5T, and R10T, respectively.

On the QRT test set, the same trends are visible as on the synthetic corrupted test sets. The MiT-B2 encoder achieves the highest AUC (0.89 ± 0.02), and the VGG-19 encoder the lowest (0.83 ± 0.03). Additionally, GastroNet pretraining results in an AUC of 0.88 ± 0.02 compared to pretraining on ImageNet 0.86 ± 0.02.

The GRT test set, recorded with endoscopes of a different manufacturer, is used to perform the same experiment as the previously mentioned test set. The results are presented in Table 4. Overall all the results are in line with the results on the first test set. The MiT-B2 encoder achieves the highest AUC on all four test sets with an AUC of 0.94 ± 0.01, 0.93 ± 0.01, 0.86 ± 0.01, and 0.68 ± 0.01, respectively. Although the ResNet-34 encoder shows better performance compared to other ResNet encoders on this test set, the difference never exceeded more than 0.02. Pretraining on GastroNet resulted in the best performance on all four test sets, with an AUC of 0.97 ± 0.00, 0.95 ± 0.00, 0.88 ± 0.01, and 0.68 ± 0.01.

Table 5 shows the percentage decrease with respect to performance on original test sets for both datasets. The reference model exhibits a decrease of 2.8%, 12.3%, and 33.2% on R2T, R5T, and R10T for the Olympus dataset. Using the EfficientNet-B5 encoder results in improved robustness and shows a percentage decrease on all robustness sets of 2.2%, 8.3%, and 28.1%. The VGG-19 encoder reports the highest percentage decrease on all three robustness sets (3.9%, 16.0%, 38.4% down). Pretraining with GastroNet yields overall the highest performance (Table 3), and additionally also has the lowest percentage

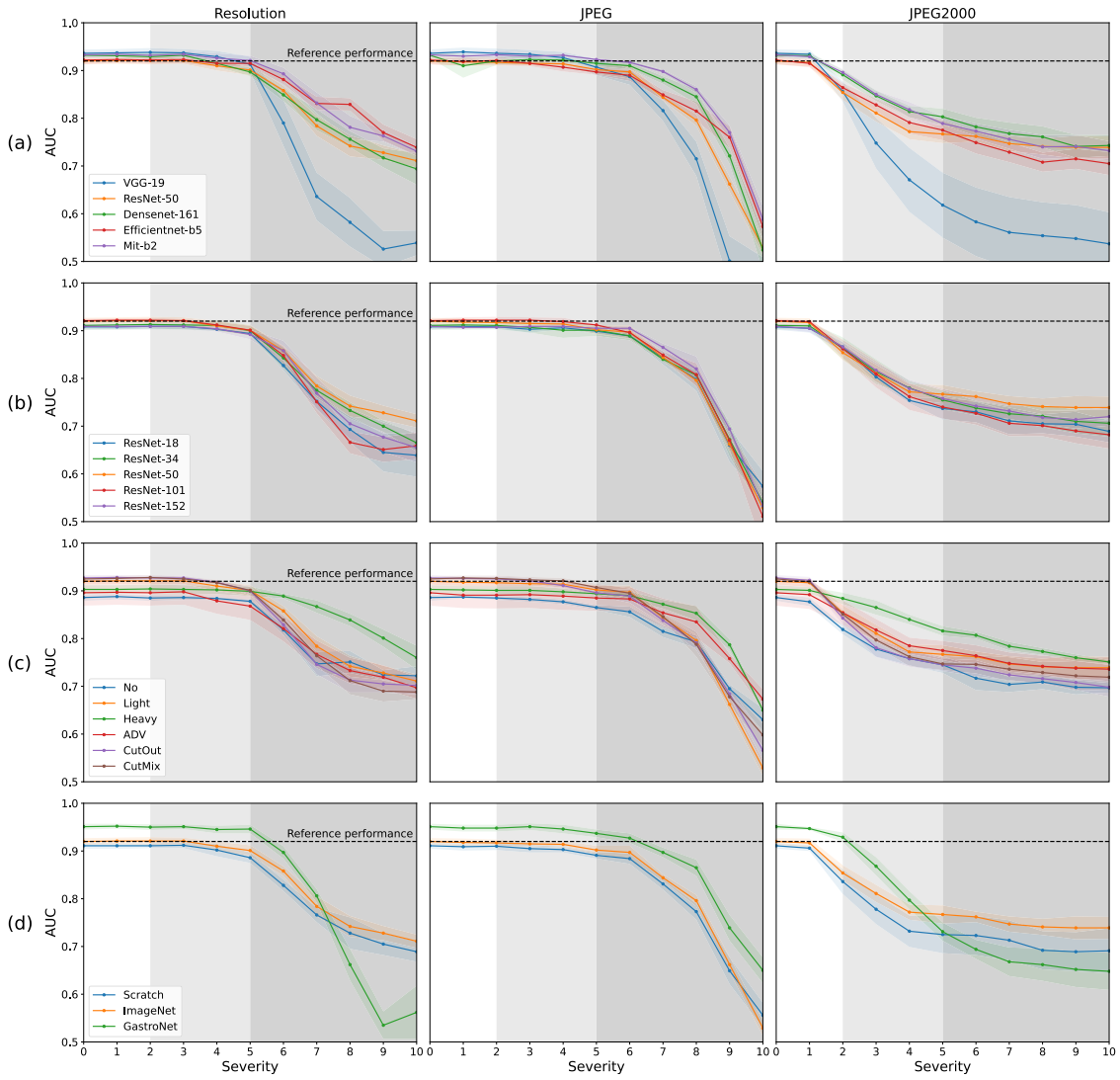

Fig. 11. Model performance evaluation for compression corruptions for all four classes of experiments. From top to bottom, (a) shows the experiment focusing on the network architecture. (b) portrays the experiments on encoder complexity. (c) and (d) depict the experiments on data augmentation and pretraining, respectively. The white zone in each figure corresponds to severity Levels 1 and 2, which represents the amount of limited degradation typically present in high-quality endoscopic image datasets. The light gray area in each figure consists of Levels 3–5, corresponding to the amount of degradation expected in clinical practice. The dark gray area includes Levels 5–10, representing increasingly unrealistic levels of image degradation. The black dotted horizontal line in all figures depicts the performance of the reference model on the original test set.

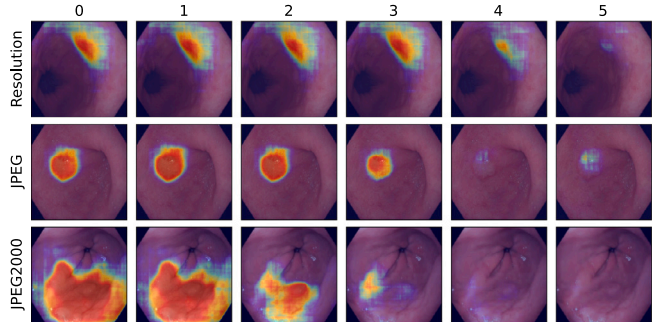

Fig. 12. Examples illustrating the negative impact of compression artifacts on model predictions. The heatmap represents the raw segmentation output.

decrease in performance on all robustness sets (2.2%, 9.4%, 31.8% down, see Table 5), thereby showing the best robustness.

Additionally, on the GRT test set, the EfficientNet-B5 encoder shows the lowest decrease on the R5T and R10T (8.7% and 26.6% down), and the DenseNet-16 encoder on the R2T (1.3%). Furthermore, similar to the results on the first dataset, GastroNet pretraining has the lowest decrease in performance on all robustness test sets (1.5% and 9.2%, 29.4% down).

8.5. Public datasets

A subset of previous experiments is conducted on two public datasets, with the results presented in Table 6. Similar trends to those observed in the Barrett's neoplasia datasets discussed earlier are appearing in both public datasets. The most significant differences are observed in terms of different architectures and pretraining methods.

On the KVASIR-SEG dataset, the MiT-B2 encoder demonstrates the best performance with a Dice coefficient of 0.87 ± 0.00, 0.86 ± 0.01, 0.82 ± 0.01, and 0.60 ± 0.01 for the original test set, R2T, R5T, and R10T, respectively. Notably, utilizing GastroNet pretraining results in a Dice of 0.84 ± 0.02 on the R5T set, exceeding the performance of ImageNet-pretrained weights, which achieves a Dice coefficient of 0.77 ± 0.01. Regarding the GIANA dataset, larger differences are

Table 4
Performance on the generalization Barrett's esophagus test sets, including the original test set, R2T, R5T, and R10T. AUC values are presented as mean ± std.

Encoder	Augmentation	Pretraining	Test set $AUC \uparrow$	R2T set $AUC \uparrow$	R5T set $AUC \uparrow$	R10T set $AUC \uparrow$	Parameters	ImageNet $Top1ACC$
VGG-19	Light	ImageNet	**0.94 ± 0.01**	0.92 ± 0.01	0.82 ± 0.01	0.61 ± 0.01	20M	72.6
ResNet-50			0.93 ± 0.01	0.91 ± 0.01	0.83 ± 0.01	0.64 ± 0.00	23M	76.1
DenseNet-16			0.93 ± 0.01	0.91 ± 0.01	0.84 ± 0.01	0.67 ± 0.01	26M	77.7
EfficientNet-B5			0.92 ± 0.01	0.90 ± 0.0.01	0.84 ± 0.01	0.67 ± 0.01	28M	83.6
MiT-B2			**0.94 ± 0.01**	**0.93 ± 0.01**	**0.86 ± 0.01**	**0.68 ± 0.01**	24M	81.6
ResNet-18	Light	ImageNet	0.93 ± 0.01	0.90 ± 0.01	0.80 ± 0.02	0.63 ± 0.01	11M	69.8
ResNet-34			**0.94 ± 0.01**	**0.91 ± 0.01**	**0.83 ± 0.02**	**0.65 ± 0.01**	21M	73.3
ResNet-50			0.93 ± 0.01	**0.91 ± 0.01**	**0.83 ± 0.01**	0.64 ± 0.01	23M	76.2
ResNet-101			0.92 ± 0.02	**0.91 ± 0.01**	0.82 ± 0.02	0.64 ± 0.02	42M	77.4
ResNet-152			0.92 ± 0.01	**0.91 ± 0.01**	**0.83 ± 0.01**	0.64 ± 0.01	58M	78.3
ResNet-50	No	ImageNet	0.91 ± 0.00	0.87 ± 0.00	0.76 ± 0.01	0.61 ± 0.01	23M	76.1
	Light		**0.93 ± 0.06**	**0.91 ± 0.01**	**0.83 ± 0.01**	**0.64 ± 0.01**		
	Heavy		0.90 ± 0.00	0.88 ± 0.01	0.81 ± 0.00	**0.64 ± 0.01**		
	ADV		0.91 ± 0.03	0.88 ± 0.03	0.76 ± 0.03	0.60 ± 0.01		
	CutOut		0.92 ± 0.01	0.89 ± 0.01	0.79 ± 0.00	0.63 ± 0.01		
	CutMix		**0.93 ± 0.01**	0.90 ± 0.01	0.79 ± 0.01	0.62 ± 0.01		
ResNet-50	Light	No	0.91 ± 0.01	0.89 ± 0.01	0.77 ± 0.01	0.62 ± 0.01	23M	76.1
		ImageNet	0.93 ± 0.01	0.91 ± 0.01	0.83 ± 0.01	0.64 ± 0.00		
		GastroNet	**0.97 ± 0.00**	**0.95 ± 0.00**	**0.88 ± 0.01**	**0.68 ± 0.01**		

Table 5
Performance decrease in terms of percentage on the robustness sets, compared to the AUC score on both Barrett's esophagus test sets.

Encoder	Pretraining	Barrett's esophagus			Generalization Barrett's esophagus		
		$R2T \downarrow$	$R5T \downarrow$	$R10T \downarrow$	$R2T \downarrow$	$R5T \downarrow$	$R10T \downarrow$
VGG-19	ImageNet	−3.9%	−16.0%	−38.4%	−2.1%	−12.0%	−34.5%
ResNet-50		−2.8%	−12.3%	−33.2%	−2.6%	−11.2%	−31.0%
DenseNet-16		−3.0%	−11.1%	−31.6%	**−1.3%**	−9.0%	−27.9%
EfficientNet-B5		**−2.2%**	**−8.3%**	**−28.1%**	−1.9%	**−8.7%**	**−26.6%**
MiT-B2		−2.7	−9.2%	−30.3%	−1.9%	−9.4%	−27.5%
ResNet-50	No	−4.0%	−15.0%	−35.7%	−2.7%	−15.2%	−31.9%
	ImageNet	−2.8%	−12.3%	−33.2%	−2.6%	−11.2%	−31.0%
	GastroNet	**−2.2%**	**−9.4%**	**−31.8%**	**−1.5%**	**−9.2%**	**−29.4%**

Table 6
Performance on the public test sets, including the original test set, R2T, R5T, and R10T. For the KVASIR-SEG test sets the DICE values are shown as mean ± std. For the GIANA test set AUC and Dice values are presented as mean ± std.

Encoder	Augmentation	Pretraining	KVASIR-SEG	R2T	R5T	R10T	GIANA		R2T		R5T		R10T	
			$DICE \uparrow$	$DICE \uparrow$	$DICE \uparrow$	$DICE \uparrow$	$DICE \uparrow$	$AUC \uparrow$	$DICE \uparrow$	$AUC \uparrow$	$DICE \uparrow$	$AUC \uparrow$	$DICE \uparrow$	$AUC \uparrow$
VGG-19	Light	ImageNet	0.84 ± 0.01	0.83 ± 0.01	0.75 ± 0.01	0.48 ± 0.02	0.66 ± 0.03	0.96 ± 0.01	0.60 ± 0.03	0.95 ± 0.01	0.47 ± 0.02	0.87 ± 0.02	0.28 ± 0.03	0.70 ± 0.02
ResNet-50			0.85 ± 0.01	0.83 ± 0.01	0.77 ± 0.01	0.51 ± 0.01	**0.70 ± 0.02**	**0.99 ± 0.00**	**0.65 ± 0.02**	0.97 ± 0.01	**0.53 ± 0.02**	0.86 ± 0.02	**0.34 ± 0.02**	0.67 ± 0.01
MiT-B2			**0.87 ± 0.00**	**0.86 ± 0.01**	**0.82 ± 0.01**	**0.60 ± 0.01**	0.69 ± 0.01	0.98 ± 0.01	0.64 ± 0.01	**0.98 ± 0.001**	**0.53 ± 0.02**	**0.93 ± 0.01**	**0.34 ± 0.01**	**0.79 ± 0.01**
ResNet-18	Light	ImageNet	0.82 ± 0.01	0.80 ± 0.01	0.74 ± 0.00	0.42 ± 0.01	0.69 ± 0.01	0.99 ± 0.00	0.62 ± 0.01	**0.98 ± 0.03**	0.48 ± 0.01	0.88 ± 0.01	0.30 ± 0.01	0.69 ± 0.00
ResNet-50			**0.85 ± 0.01**	0.83 ± 0.01	0.77 ± 0.01	0.51 ± 0.01	**0.70 ± 0.02**	0.99 ± 0.00	**0.65 ± 0.02**	0.97 ± 0.01	0.53 ± 0.02	0.86 ± 0.02	**0.34 ± 0.02**	0.67 ± 0.01
ResNet-152			**0.85 ± 0.01**	**0.84 ± 0.01**	**0.78 ± 0.01**	**0.52 ± 0.03**	0.69 ± 0.01	0.99 ± 0.01	**0.65 ± 0.01**	0.97 ± 0.01	**0.54 ± 0.01**	**0.90 ± 0.02**	0.33 ± 0.03	**0.71 ± 0.01**
ResNet-50	No	ImageNet	0.82 ± 0.01	0.80 ± 0.01	0.72 ± 0.02	0.47 ± 0.03	0.65 ± 0.03	0.98 ± 0.00	0.59 ± 0.02	**0.97 ± 0.00**	0.48 ± 0.01	**0.88 ± 0.00**	0.31 ± 0.03	0.55 ± 0.01
	Light		**0.85 ± 0.01**	**0.83 ± 0.01**	**0.77 ± 0.01**	**0.51 ± 0.01**	**0.70 ± 0.02**	**0.99 ± 0.00**	**0.65 ± 0.02**	**0.97 ± 0.01**	**0.53 ± 0.02**	0.86 ± 0.02	0.34 ± 0.02	0.67 ± 0.01
	Heavy		0.84 ± 0.01	**0.83 ± 0.01**	0.75 ± 0.01	0.49 ± 0.02	0.66 ± 0.01	0.97 ± 0.01	0.62 ± 0.01	0.96 ± 0.02	0.51 ± 0.01	0.87 ± 0.02	**0.35 ± 0.01**	**0.71 ± 0.01**
ResNet-50	Light	No	0.63 ± 0.01	0.58 ± 0.01	0.48 ± 0.02	0.23 ± 0.03	0.48 ± 0.04	0.69 ± 0.06	0.41 ± 0.04	0.66 ± 0.05	0.30 ± 0.02	0.61 ± 0.02	0.18 ± 0.02	0.55 ± 0.01
		ImageNet	0.85 ± 0.01	0.83 ± 0.01	0.77 ± 0.01	0.51 ± 0.01	0.70 ± 0.02	0.99 ± 0.00	0.65 ± 0.02	0.97 ± 0.01	**0.53 ± 0.02**	0.86 ± 0.02	**0.34 ± 0.02**	0.67 ± 0.01
		GastroNet	**0.88 ± 0.01**	**0.88 ± 0.01**	**0.84 ± 0.02**	**0.58 ± 0.01**	**0.73 ± 0.01**	**0.99 ± 0.00**	**0.67 ± 0.02**	**0.98 ± 0.01**	0.52 ± 0.02	**0.90 ± 0.02**	0.32 ± 0.01	**0.71 ± 0.02**

observed in classification performance compared to localization performance. Once again, GastroNet pretraining and the MiT-B2 encoder yield an AUC of 0.90 ± 0.02 and 0.93 ± 0.01, respectively, at R5T. In contrast, the reference encoder achieves AUC values of 0.86 ± 0.02 at R5T.

9. Discussion

The quality of the images in endoscopy relies heavily on the expertise and experience of the endoscopist, as well as the specifications of the imaging system. This leads to a more diverse range of data compared to typical datasets used for the training and validation of CADe/CADx systems. If endoscopic CADe/CADx systems cannot perform reliably under these conditions, they are unreliable in clinical practice. Therefore, the objective of this study is to conduct a comprehensive evaluation of the robustness of popular DNN architectures and training strategies across multiple datasets, using clinically relevant and calibrated levels of IQ degradation. Additionally, a manually selected dataset including lower subjective image-quality frames is also used for this evaluation. This assessment offers valuable insights into the performance of these models and their limitations for endoscopic applications. Besides the insights and limitations, the conducted experiments also show trends on how to improve robustness.

Overall findings: First, the obtained results indicate that DNN performance can decrease up to 31%, 10%, and 34% for user-dependent degradation, image acquisition and processing changes, and compression corruptions up to severity level 5. Note that this decrease in performance happens within the clinically relevant boundaries, from a human visual perspective these images remain to appear realistic. Additionally, on the robustness datasets R2T, R5T, and R10T we have observed a decrease in performance of ±2%, ±10%, and ±30%, respectively, from the reference model, on both Barrett's Neoplasia and the two public datasets. This finding aligns with literature on robustness in the broader scope of medical image analysis that shows that modern

DNN architectures are highly susceptible to distribution shifts, corruptions, and artifacts (Boone et al., 2023; Maron et al., 2021; Young et al., 2021; Jiang et al., 2023).

Robustness strategies: To improve robustness, we have analyzed various strategies, including increasing network complexity, extensive data augmentation, in-domain pretraining, and more advanced network architectures. Our findings on both the corrupted and the QRT test sets suggest that the most effective methods for improving robustness against reduced image quality are (1) in-domain pretraining and (2) more advanced encoder architectures. Both increasing network complexity (in terms of the number of parameters due to more layers) and extensive data augmentation do not result in a considerable increase in robustness.

Our findings align with previous research conducted by Hendrycks et al. (2019b), who empirically showed that a simple self-supervised learning framework improves model robustness compared to solely fully supervised training. According to Navarro et al. (2021), the advantages of self-supervised in-domain pretraining in medical image analysis become apparent when evaluating networks against imperfect data. These findings may be explained by the observation that the representations learned by in-domain self-supervised training contain explicit information about the semantic information of an image that does not emerge as clearly with supervised training (Caron et al., 2021). Additionally, research by Hendrycks et al. (2021b), Bhojanapalli et al. (2021), Kusters et al. (2024), Boone et al. (2023), Zhou et al. (2022) and Bai et al. (2021), highlights the increased robustness exhibited by transformers. In their investigation, Bai et al. (2021) delved into the transformer architecture's components contributing to this enhancement, pinpointing the self-attention mechanism as the primary factor. Furthermore, according to findings from (Wang et al., 2023), CNNs demonstrate the potential to match transformers in robustness through tailored design optimizations. This resonates with our results, demonstrating that modern CNN models such as EfficientNet-B5 exhibit a comparable level of robustness to the MiT-B2 network.

User-dependent degradations: Specifically, the performed analysis of user-dependent corruptions reveals that even small corruptions (at Level 2) can have a substantial impact on model performance. For example, the VGG-19 encoder experiences an 11.6% decrease in performance on the test set corrupted with motion blur. However, using a more recent encoder architecture such as the MiT-B2 limits the percentage performance decrease to 2.8% only at severity Level 2. Nevertheless, it should be noted that even the MiT-B2 encoder still exhibits a performance drop of 11.8% at severity Level 5, which falls within the clinically relevant boundaries. Furthermore, the obtained results of the augmentation experiments indicate that adding Gaussian blur as an extra augmentation technique does not improve the model's inherent robustness against defocus blur or motion blur. This suggests that more targeted approaches may be necessary to address the specific challenges presented by individual types of degradation.

Post-processing and acquisition changes: The presented investigation delves into how alterations in contrast, brightness, saturation, and hue, which may arise from virtual chromoendoscopy technologies, affect the performance of DNNs. We have found that these changes lead to performance loss for all encoder architectures and data augmentation techniques between severity Levels 2 and 5. However, pretraining on GastroNet makes the model robust to these types of changes (Row d of Fig. 8). This can be explained by the fact that in-domain data is more likely to contain relevant variations for the target domain, resulting in models that are better equipped to handle subtle and clinically relevant corruptions. Moreover, it is worth noting that although applying a synthetic filter can improve the perceived image quality for the user, it actually results in an 8.2% performance decrease for the reference model. This highlights the importance of carefully considering the impact of different image processing techniques on DNNs' performances.

Compression corruptions: In clinically relevant areas, JPEG compression has a less substantial effect on model performance compared to JPEG2000 compression. This can be attributed to the fact that the training data includes JPEG-compressed images, but not JPEG2000-compressed images. Previous research has shown that training and predicting the same compression type of images can lead to better results (Zanjani et al., 2019). However, from our findings, it is worth noting that at higher levels of compression (Level 10), all encoders except for VGG-19 maintain their performances within the AUC range of 0.71–0.78 for JPEG2000 compression. In contrast, for the obtained JPEG compression results, as the levels of severity are increased to Level 10, the performance of all encoders drops to an AUC close to 0.5, indicating random classification output.

Robustness test sets: The results on the robustness with Barrett's esophagus test sets demonstrate that encoder architectures that perform similarly on the original test set exhibit varying performances on the R2T, R5T, and QRT test sets. These observations emphasize the significance of including robustness evaluation for DNNs utilized in endoscopy since other architectural designs may exhibit improved performance under heterogeneous, real-world imaging conditions. Furthermore, the influence of model complexity appears to be limited, as all models exhibit comparable sensitivity to IQ degradation. This finding contrasts with previous research that suggested an increase in model complexity could potentially enhance robustness (Hendrycks and Dietterich, 2019; Xie and Yuille, 2020). This inconsistency might be attributed to the fact that more complex models tend to require larger amounts of data to perform optimally. The previously mentioned studies base their conclusions on models trained on natural image datasets, which are significantly larger than medical datasets. This suggests that with larger endoscopic datasets, more complex models could potentially yield more robust solutions as well.

Furthermore, the implementation of CutMix and CutOut as augmentation techniques does not yield more robust solutions. While these methods have demonstrated success in the robustness on natural image datasets (Yun et al., 2019; DeVries and Taylor, 2017), their applicability to medical image semantic segmentation remains limited. Medical image content often possesses non-rigid morphological characteristics that demand careful preservation, making cut-and-paste-based augmentation methods less applicable. These techniques can potentially compromise vital objects within the images, making meticulously annotated pixel-level data unusable. This observation aligns with findings in other endoscopy research, where the utilization of CutMix or CutOut demonstrated negligible to marginal performance improvements (Cho et al., 2023).

Additionally, the results in Table 5 indicate that using GastroNet-pretrained weights not only results in better performance on all 5 test sets but also reduces the percentage of performance decrease on the robustness test set compared to the peak performance on the original test sets.

The generalization of the findings has been further supported by evaluating the performance on two public datasets, indicating that the observed trends extend to other endoscopic tasks. The percentage decrease in localization performance on the KVASIR-SEG robustness sets aligns with the performance decrease observed using the Barrett's neoplasia datasets. Similarly, comparable results have been obtained for the GIANA robustness sets.

Robustness testing standardization: Thorough understanding of the weaknesses and potential impact of IQ degradation on DNN performance is crucial in medical applications. Here, similar challenges of IQ degradation are widespread across various modalities, albeit with varying corruptions. Comprehensive robustness evaluation is necessary to gain insight into the severity and consequences of these image degradations. Some image corruptions may be managed by protocol standardization, while others are induced by less controllable factors. For instance, in endoscopy, compression corruptions and image acquisition and processing changes can be controlled during tuning of the

Table 7
Training details of the self-supervised pertaining on GastroNet.

Hyperparameter	Value
Input size	224
Optimizer	SGD
Gradient clip	3.0
LR decay schedule	Cosine schedule
Train steps	200
Train batch size	320
learning rate	5×10^{-4}
Warming-up epochs	10
Weight decay	0.04
Mixed precision	Yes

system setup. However, user-dependent corruptions can only partly be controlled by live feedback during acquisition. This makes endoscopic imaging different than e.g. CT or MRI, where image quality may not be controlled in such a live fashion and factors such as noise are expressed differently. This renders it important that factors of quality degradation are explicitly investigated for each modality, and combined in a standard set of robustness tests, using for example the methods and tools presented in this work.

Limitations: One of the main contributions of this study is that it provides a comprehensive evaluation of the robustness of popular CNN architectures and data augmentation methods, using clinically relevant and calibrated levels of IQ degradation. However, it is important to note that the synthetic corruptions used in this study are relatively simple and only partly encompass the heterogeneous and diverse data encountered in clinical practice. Future research endeavors can focus on incorporating even more realistic corruptions. For instance, depth estimation techniques can be included to better simulate over-exposure scenarios. Moreover, in this work, we have calibrated the IQ level to ensure that frames remain classifiable by an endoscopist while still maintaining a realistic appearance.

Furthermore, it may be argued that in the case of the compression methods, the high compression factors applied in JPEG are not realistic for medical applications. However, the obtained results address that lossy compression methods can lead to information loss that may not impact the endoscopists' diagnosis but can still negatively affect DNNs, particularly if these compression methods are absent from the training data.

Finally, the objective of this study was not to present a single method to improve robustness against image corruptions or domain shift, like some works in general computer vision e.g. Rusak et al. (2020) and Dai et al. (2023) focuses on. However, this study presents a thorough systematic approach to characterize and quantify the problem, providing guidance and insights for further research in this area. In conjunction with the development of endoscopic robustness benchmarks or challenges, such as the one proposed in Ali and Ghatwary (2022), this study has the potential to significantly advance the field, similar to the advancements observed in natural image analysis (Recht et al., 2019; Hendrycks et al., 2019a, 2021a). By addressing the challenges of robustness in endoscopic image analysis and proposing evaluation standards, this study contributes to further progress and the development of reliable deep-learning models for clinical applications.

10. Conclusion

This study has conducted an extensive empirical study to explore the effects of image degradation on DNNs' model performance for cancer detection in endoscopy. This work provides valuable insights into the robustness of popular DNN architectures on clinically relevant and calibrated levels of image degradation. Overall, the conducted study highlights the importance of understanding the impact of a decrease in image quality and the need to include robustness evaluation for DNNs used in endoscopy, as other architectural designs may exhibit better performance under heterogeneous, real-world imaging conditions. Our findings also suggest that in-domain pretraining and more advanced encoder architectures are effective methods for improving robustness against reduced image quality. Future research in this area will be crucial to ensure reliable and robust deep-learning models for endoscopic clinical applications.

CRediT authorship contribution statement

Tim J.M. Jaspers: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Tim G.W. Boers:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Carolus H.J. Kusters:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Martijn R. Jong:** Writing – review & editing, Methodology, Formal analysis, Data curation, Conceptualization. **Jelmer B. Jukema:** Writing – review & editing, Methodology, Formal analysis, Data curation, Conceptualization. **Albert J. de Groof:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Jacques J. Bergman:** Writing – review & editing, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Peter H.N. de With:** Writing – review & editing, Validation, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Fons van der Sommen:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jacques J. Bergman reports financial support was provided by Olympus Corporation, Tokyo, Japan. Peter H.N. de With reports financial support was provided by Olympus Corporation, Tokyo, Japan. Fons van der Sommen reports financial support was provided by Olympus Corporation, Tokyo, Japan. Jacques J. Bergman reports financial support was provided by PENTAX Medical. Jacques J. Bergman reports financial support was provided by Medtronic. Jacques J. Bergman reports financial support was provided by Aqua Medical.

Data availability

Besides the publicly available Kvasir-SEG and Giana datasets, the image data is this research will not be made public. The code will be made public on GitHub as specified in the manuscript.

Acknowledgments

This publication is part of the project "You won't find what you don't image: Exposing blind spots in endoscopic cancer screening" (project number: 19091) of the research program Talentprogramma Veni-TTW which is financed by the Dutch Research Council (NWO).

Appendix. Supplementary material

A.1. GastroNet pertaining details

The proposed framework by Caron et al. (2021) has introduced a novel approach called "Self-Distillation" with No Labels"(DINO). This method utilizes distillation-based techniques to facilitate efficient learning with smaller batch sizes, consequently reducing the demand for extensive compute resources. Full details of the training parameters on GastroNet can be found in Table 7. This experiment was executed on a high-performance computing platform utilizing a single custom compute node with a dual-socket Xeon 4216 CPU (Intel Corp, USA) with 196-GB RAM and 8 Geforce RTX2080 Ti's (Nvidia, USA).

Table 8
Overview of the synthetic image corruptions with corresponding severity levels. In red severity Level 0 is indicated.

	Synthetic image corruption	Parameters	Severity 0	1	2	3	4	5	6	7	8	9	10
User-dependent degradations	Motion blur	r, σ	(0,0)	(10, 3)	(24, 8)	(40, 16)	(70, 16)	(100, 16)	(100, 24)	(120, 24)	(120, 30)	(150, 30)	(200, 40)
	Local defocus blur	r, t, σ	(0, 1, 0)	(12, 0.4, 31)	(12, 0.4, 41)	(10, 0.3, 51)	(8, 0.3, 61)	(8, 0.2, 71)	(8, 0.2, 81)	(6, 0.2, 91)	(6, 0.2, 101)	(4, 0.1, 111)	(4, 0.1, 121)
	Overexposure	r, σ	(0, 1)	(12, 0.9)	(12, 0.8)	(10, 0.8)	(8, 0.75)	(8, 0.65)	(8, 0.5)	(6,0.3)	(6,0.2)	(4,0.1)	(2,0.1)
Compression corruptions	Reducing resolution	f	1	2	3	4	5	6	8	10	12	16	20
	JPEG compression	c	1	4	12	20	28	34	40	46	50	54	58
	JPEG2000 compression	c	1	4	12	20	28	34	40	46	50	54	58

			Severity −10	−9	−8	−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8	9	10
Post-processing and acquisition changes	Sharpness	f	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	2.0	4.0	5.0	10	15	20	40	60	80	100
	Contrast	f	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.9	0.95	1	1.1	1.2	1.25	1.3	1.4,	1.7	2.0	2.3	2.7	3.1
	Saturation	f	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.05	1.1	0.15	1.2	1.25	1.3	1.7	2.0	2.6	3.5
	Hue	f	−0.1	−0.08	−0.065	−0.04	−0.03	−0.025	−0.02	−0.016	−0.013	−0.01	0	0.01	0.013	0.016	0.018	0.02	0.04	0.05	0.06	0.08	0.1
	Brightness	f	0.2	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.85	0.9	1	1.2	1.4	1.6	1.8	2.0	2.4	2.8	3.2	3.6	4.0

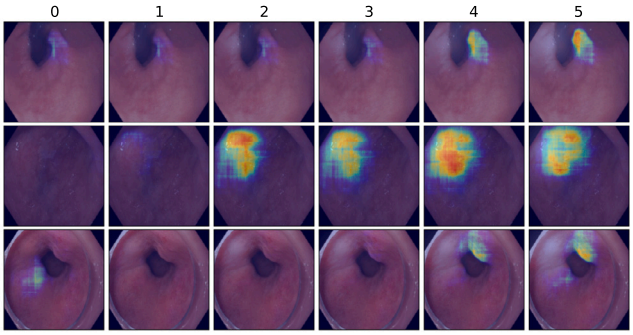

Fig. 13. Visual example of how synthetic overexposure leads to false positives. The examples range from severity Levels 0–5.

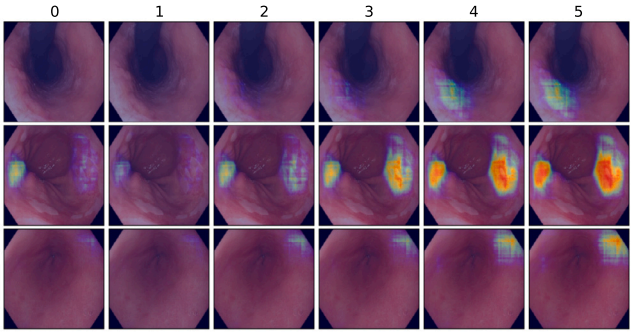

Fig. 14. Visual example of how an increase in saturation leads to false positives. The examples range from severity Levels 0–5.

A.2. Additional visual analysis

Visual examples in Figs. 13 and 14 show instances where synthetic over-exposure and saturation corruptions result in false positives within the clinically relevant areas (severity Levels 0–5). These effects are comparable to the false positives induced by the use of synthetic sharpness filters.

A.3. Synthetic image corruptions

Table 8 presents a comprehensive overview of parameters and their respective severity levels for all types of corruptions. Severity Level 0 is highlighted in red for easy reference. To achieve the motion blur corruption, we utilize a Gaussian operator with a specified radius (r) and standard deviation (σ), applying it at a fixed angle to convolve the image. For both local defocus blur and overexposure, we generate a Gaussian mask with a given radius (r), standard deviation (σ), and transparency (t) at a randomly selected point. Adjustments in sharpness, contrast, saturation, hue, brightness, and resolution reduction are uniformly adapted using a single factor (f). The severity levels for JPEG and JPEG2000 are determined based on the compression factor (c), calculated as the ratio between bytes before compression and bytes after compression. The code to regenerate all corruptions can be found on our GitHub page https://github.com/BONS-AI-VCA-AMC/Robustness.

A.4. Visual examples severity levels

The examples in Figs. 15, 16, and 17 illustrate different severity levels for all corruptions. The background colors of white, light gray, and dark gray represent the section's variation in expert datasets, the amount of degradation that can be faced in clinical use, and

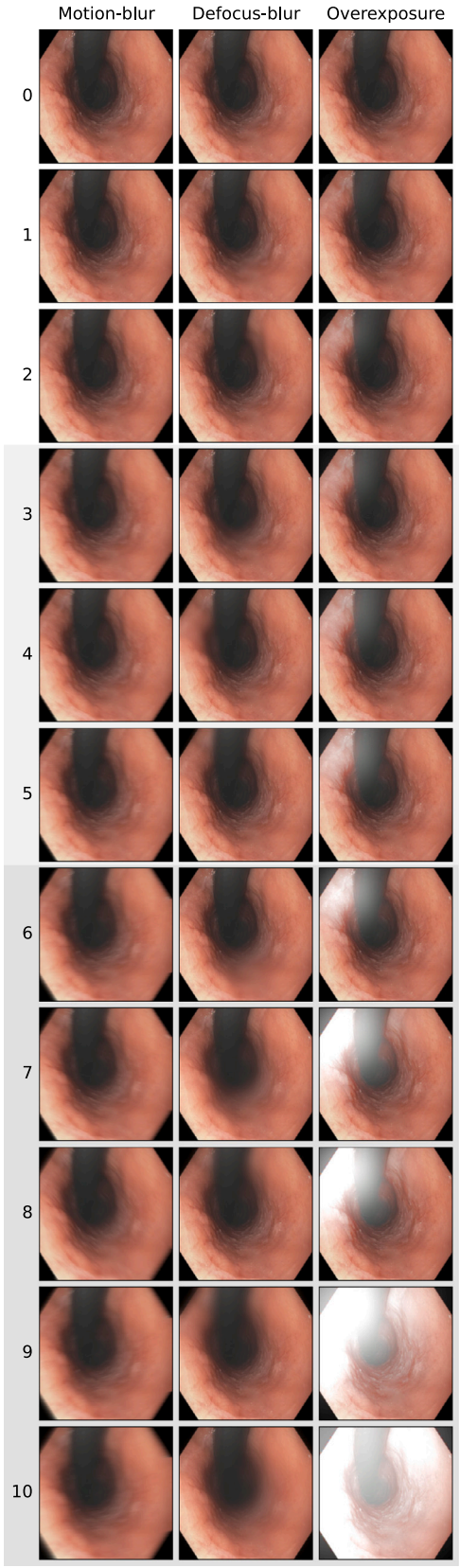

Fig. 15. Visual examples illustrating the ten different levels of severity for user-dependent corruptions. The light gray area indicates the clinically relevant amount of degradation and the dark gray area the unrealistic amount of image degradation.

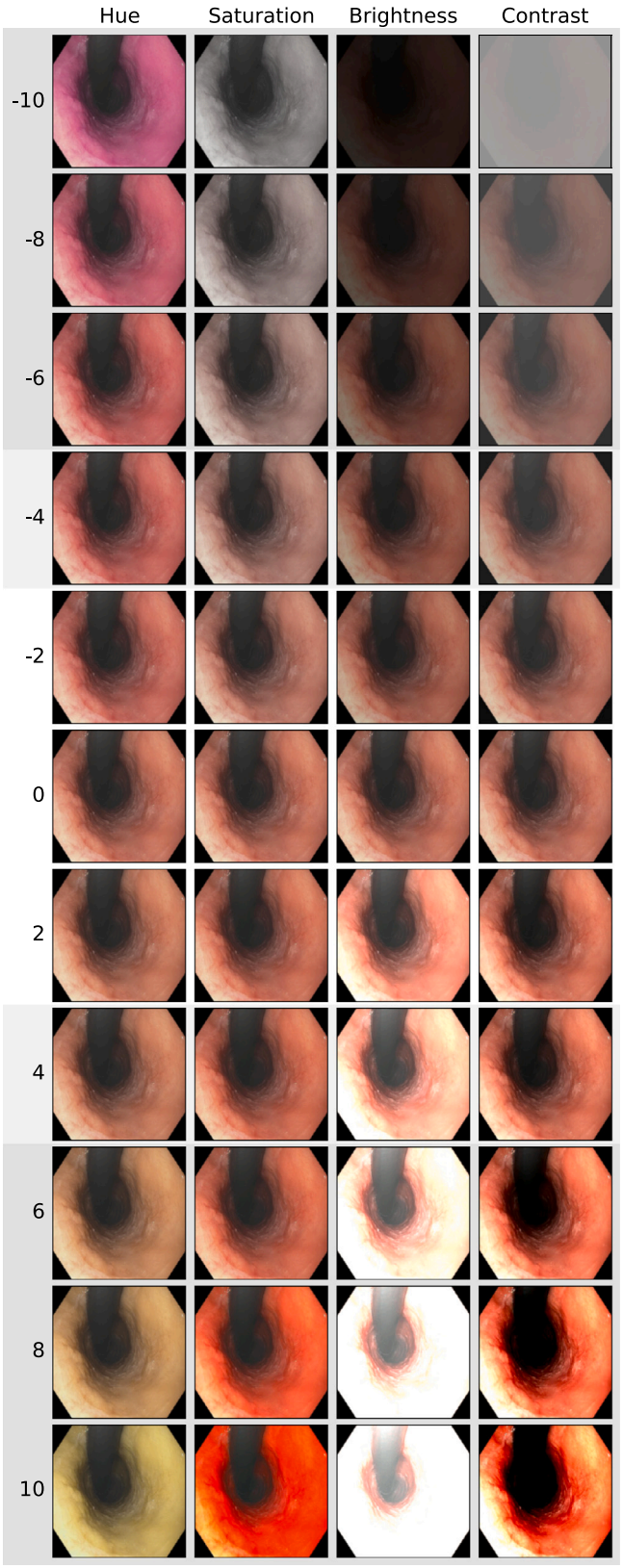

Fig. 16. Visual examples illustrating the ten ascending and descending levels of severity for image acquisition and processing corruptions. The light gray area indicates the clinically relevant amount of degradation and the dark gray area the unrealistic amount of image degradation. The examples of the sharpness filter are included in the next figure.

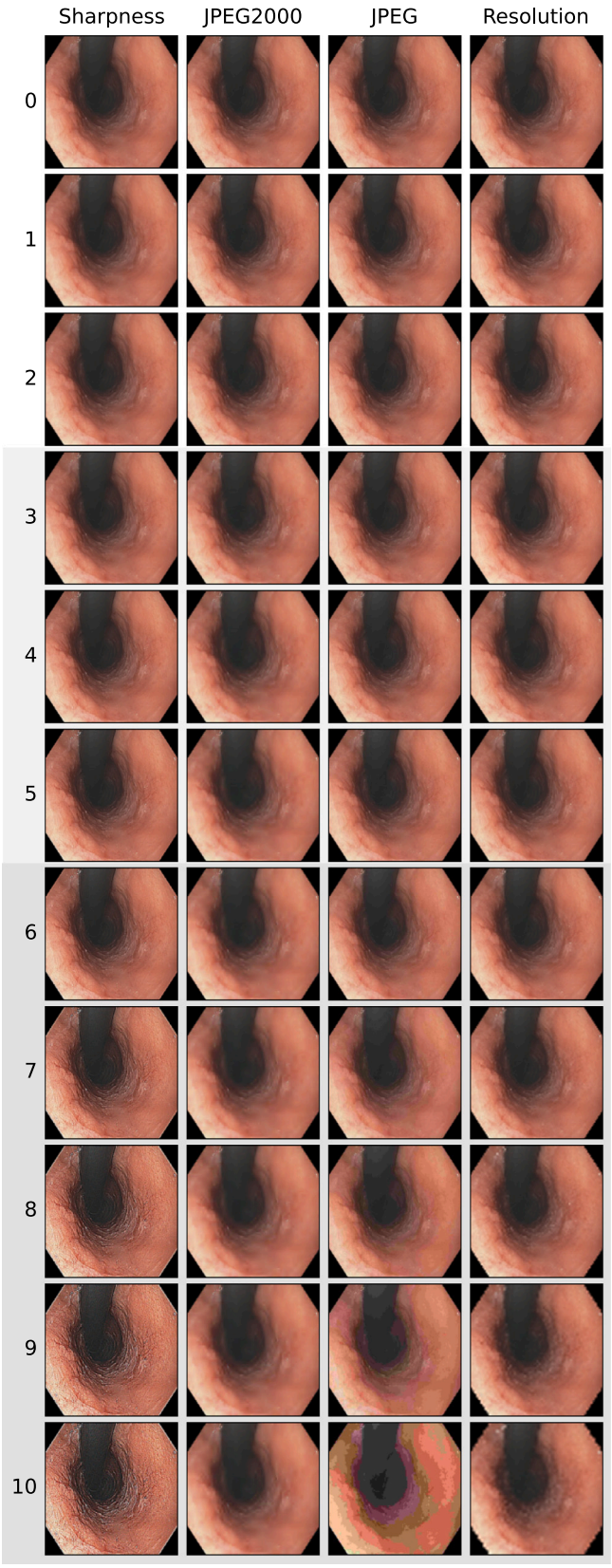

Fig. 17. Visual examples illustrating the ten different levels of severity for compression artifacts and the sharpness filter. The light gray area indicates the clinically relevant amount of degradation and the dark gray area the unrealistic amount of image degradation.

an unrealistic proportion of degradation, respectively. Notably, the corruption factors related to image acquisition and processing – Hue, Saturation, Brightness, and Contrast – span from −10 to 10, whereas other corruptions have a narrower range, specifically 0 to 10.

References

Ali, S., Ghatwary, N., 2022. Endoscopic computer vision challenges 2.0. In: 4th International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV2022) in Conjunction with the 19th IEEE International Symposium on Biomedical Imaging (ISBI2022). Kolkata, India.

Arnold, M., Abnet, C.C., Neale, R.E., Vignat, J., Giovannucci, E.L., McGlynn, K.A., Bray, F., 2020. Global burden of 5 major types of gastrointestinal cancer. Gastroenterology 159 (1), 335–349.e15. http://dx.doi.org/10.1053/j.gastro.2020.02.068.

Bai, Y., Mei, J., Yuille, A., Xie, C., 2021. Are transformers more robust than CNNs? In: Thirty-Fifth Conference on Neural Information Processing Systems.

Benz, P., Ham, S., Zhang, C., Karjauv, A., Kweon, I.S., 2021. Adversarial robustness comparison of vision transformer and MLP-Mixer to CNNs. arXiv:2110.02797.

Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F., 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Comput. Med. Imaging Graph. 43, 99–111. http://dx.doi.org/10.1016/j.compmedimag.2015.02.007, URL: https://www.sciencedirect.com/science/article/pii/S0895611115000567.

Beveridge, C.A., Mittal, C., Muthusamy, V.R., Rastogi, A., Kushnir, V., Wood, M., Wani, S., Komanduri, S., 2023. Identification of visible lesions during surveillance endoscopy for Barrett's esophagus: a video-based survey study. Gastrointest. Endosc. 97 (2), 241–247.e2. http://dx.doi.org/10.1016/j.gie.2022.08.024, URL: https://www.sciencedirect.com/science/article/pii/S0016510722019320.

Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., Veit, A., 2021. Understanding robustness of transformers for image classification. In: 2021 IEEE/CVF International Conference on Computer Vision. ICCV, IEEE Computer Society, Los Alamitos, CA, USA, pp. 10211–10221. http://dx.doi.org/10.1109/ICCV48922.2021.01007, URL: https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.01007.

Boone, L., Biparva, M., Mojiri Forooshani, P., Ramirez, J., Masellis, M., Bartha, R., Symons, S., Strother, S., Black, S.E., Heyn, C., Martel, A.L., Swartz, R.H., Goubran, M., 2023. ROOD-MRI: Benchmarking the robustness of deep learning segmentation models to out-of-distribution and corrupted data in MRI. NeuroImage 278, 120289. http://dx.doi.org/10.1016/j.neuroimage.2023.120289, URL: https://www.sciencedirect.com/science/article/pii/S1053811923004408.

Byrne, M.F., Chapados, N., Soudan, F., Oertel, C., Pérez, M.L., Kelly, R., Iqbal, N., Chandelier, F., Rex, D.K., 2019. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. Gut 68, 94–100. http://dx.doi.org/10.1136/gutjnl-2017-314547.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. In: Proceedings of the International Conference on Computer Vision. ICCV.

Chang, Q., Ahmad, D., Toth, J., Bascom, R., Higgins, W.E., 2022. ESFPNet: efficient deep learning architecture for real-time lesion segmentation in autofluorescence bronchoscopic video. arXiv preprint arXiv:2207.07759.

Chen, P.-J., Lin, M.-C., Lai, M.-J., Lin, J.-C., Lu, H.H.-S., Tseng, V.S., 2018. Accurate classification of diminutive colorectal polyps using computer-aided analysis. Gastroenterology 154 (3), 568–575. http://dx.doi.org/10.1053/j.gastro.2017.10.010, URL: https://www.sciencedirect.com/science/article/pii/S0016508517362510.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2016a. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs.. CoRR abs/1606.00915 URL: http://dblp.uni-trier.de/db/journals/corr/corr1606.html#ChenPK0Y16.

Chen, W., Zheng, R., Baade, P.D., Zhang, S., Zeng, H., Bray, F., Jemal, A., Yu, X.Q., He, J., 2016b. Cancer statistics in China, 2015. CA: Cancer J. Clin. 66 (2), 115–132. http://dx.doi.org/10.3322/caac.21338.

Cho, B.-J., Bang, C.S., Park, S.W., Yang, Y.J., Seo, S.I., Lim, H., Shin, W.G., Hong, J.T., Yoo, Y.T., Hong, S.H., Choi, J.H., Lee, J.J., Baik, G.H., 2019. Automated classification of gastric neoplasms in endoscopic images using a convolutional neural network. Endoscopy 51 (12), 1121–1129. http://dx.doi.org/10.1055/a-0981-6133.

Cho, H., Han, Y., Kim, W.H., 2023. Anti-adversarial consistency regularization for data augmentation: Applications to robust medical image segmentation. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Springer Nature Switzerland, Cham, pp. 555–566.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. URL: http://arxiv.org/abs/1604.01685.

Cordts, M., Omran, M., Ramos, S., Scharwächter, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2015. The cityscapes dataset. In: CVPR Workshop on the Future of Datasets in Vision.

Dai, Y., Qian, Y., Lu, F., Wang, B., Gu, Z., Wang, W., Wan, J., Zhang, Y., 2023. Improving adversarial robustness of medical imaging systems via adding global attention noise. Comput. Biol. Med. 164, 107251. http://dx.doi.org/10.1016/j.compbiomed.2023.107251, URL: https://www.sciencedirect.com/science/article/pii/S0010482523007163.

de Groof, A.J., Struyvenberg, M.R., van der Putten, J., van der Sommen, F., Fockens, K.N., Curvers, W.L., Zinger, S., Pouw, R.E., Coron, E., Baldaque-Silva, F., Pech, O., Weusten, B., Meining, A., Neuhaus, H., Bisschops, R., Dent, J., Schoon, E.J., de With, P.H., Bergman, J.J., 2020. Deep-learning system detects neoplasia in patients with Barrett's esophagus with higher accuracy than endoscopists in a multistep training and validation study with benchmarking. Gastroenterology 158 (4), 915–929.e4. http://dx.doi.org/10.1053/j.gastro.2019.11.030, URL: https://www.sciencedirect.com/science/article/pii/S0016508519415862.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR, Ieee, pp. 248–255.

DeVries, T., Taylor, G.W., 2017. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552.

Dodge, S.F., Karam, L.J., 2016. Understanding how image quality affects deep neural networks. CoRR abs/1604.04004 arXiv:1604.04004 URL: http://arxiv.org/abs/1604.04004.

Dodge, S.F., Karam, L.J., 2017. A study and comparison of human and deep learning recognition performance under visual distortions. CoRR abs/1705.02498 arXiv:1705.02498 URL: http://arxiv.org/abs/1705.02498.

Ebigbo, A., Mendel, R., Probst, A., Manzeneder, J., Souza, L.A.D., Papa, J.P., Palm, C., Messmann, H., 2019. Computer-aided diagnosis using deep learning in the evaluation of early oesophageal adenocarcinoma. Gut 68, 1143–1145. http://dx.doi.org/10.1136/gutjnl-2018-317573.

Eche, T., Schwartz, L.H., Mokrane, F.-Z., Dercle, L., 2021. Toward generalizability in the deployment of artificial intelligence in radiology: Role of computation stress testing to overcome underspecification. Radiol.: Artif. Intell. 3 (6), e210097. http://dx.doi.org/10.1148/ryai.2021210097.

Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S., 2019. Adversarial attacks on medical machine learning. Science 363 (6433), 1287–1289. http://dx.doi.org/10.1126/science.aaw4399.

Fockens, K.N., Jong, M.R., Jukema, J.B., Boers, T.G.W., Kusters, C.H.J., van der Putten, J.A., Pouw, R.E., Duits, L.C., Montazeri, N.S.M., van Munster, S.N., Weusten, B.L.A.M., Alvarez Herrero, L., Houben, M.H.M.G., Nagengast, W.B., Westerhof, J., Alkhalaf, A., Mallant-Hent, R.C., Scholten, P., Ragunath, K., Seewald, S., Elbe, P., Baldaque-Silva, F., Barret, M., Ortiz Fernández-Sordo, J., Villarejo, G.M., Pech, O., Beyna, T., van der Sommen, F., de With, P.H., de Groof, A.J., Bergman, J.J., Alkhalaf, A., Alvarez Herrero, L., Baldaque-Silva, F., Barret, M., Bergman, J.J., Beyna, T., Bisschops, R., Boers, T.G., Curvers, W., Deprez, P.H., Duits, L.C., Elbe, P., Esteban, J.M., Falk, G.W., Fockens, K.N., Ganguly, E., Ginsberg, G.G., de Groof, A.J., Haidry, R., Houben, M.H., Infantolino, A., Iyer, P.G., Jong, M.R., De Jonge, P.-J., Jukema, J.B., Koch, A.K., Komanduri, S., Konda, V., Kusters, C.H.J., Leclercq, P., Leggett, C.L., Lemmers, A., Lightdale, C.J., Mallant-Hent, R.C., Moral Villarejo, G., Muthusamy, V.R., Nagengast, W., Ortiz Fernández-Sordo, J., Pech, O., Penman, I., Pleskow, D.K., Pouw, R.E., van der Putten, J.A., Ragunath, K., Scholten, P., Seewald, S., Sethi, A., Smith, M.S., Van der Sommen, F., Trindade, A., Wani, S., Waxman, I., Westerhof, J., Weusten, B.L., de With, P.H.N., Wolfsen, H.C., 2023a. A deep learning system for detection of early Barrett's neoplasia: a model development and validation study. Lancet Digit. Health 5 (12), e905–e916. http://dx.doi.org/10.1016/S2589-7500(23)00199-1, URL: https://www.sciencedirect.com/science/article/pii/S2589750023001991.

Fockens, K.N., Jukema, J.B., Boers, T., Jong, M.R., van der Putten, J.A., Pouw, R.E., Weusten, B.L.A.M., Alvarez Herrero, L., Houben, M.H.M.G., Nagengast, W.B., Westerhof, J., Alkhalaf, A., Mallant, R., Ragunath, K., Seewald, S., Elbe, P., Barret, M., Ortiz Fernández-Sordo, J., Pech, O., Beyna, T., van der Sommen, F., de With, P.H., de Groof, A.J., Bergman, J.J., 2023b. Towards a robust and compact deep learning system for primary detection of early Barrett's neoplasia: Initial image-based results of training on a multi-center retrospectively collected data set. United Eur. Gastroenterol. J. 11 (4), 324–336. http://dx.doi.org/10.1002/ueg2.12363, PMID: 37095718.

Food, U., Administration, D., 2021. FDA authorizes marketing of first device that uses artificial intelligence to help detect potential signs of colon cancer. URL: https://www.fda.gov/news-events/press-announcements/fda-authorizes-marketing-first-device-uses-artificial-intelligence-help-detect-potential-signs-colon.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W., 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=Bygh9j09KX.

Ghaffari Laleh, N., Truhn, D., Veldhuizen, G.P., Han, T., van Treeck, M., Buelow, R.D., Langer, R., Dislich, B., Boor, P., Schulz, V., Kather, J.N., 2022. Adversarial attacks and adversarial robustness in computational pathology. Nature Commun. 13 (1), 5711. http://dx.doi.org/10.1038/s41467-022-33266-0.

Goodfellow, I.J., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples. arXiv:1412.6572.

Gu, J., Zhao, H., Tresp, V., Torr, P.H.S., 2022. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (Eds.), Computer Vision – ECCV 2022. Springer Nature Switzerland, Cham, pp. 308–325.

Guimarães, P., Keller, A., Fehlmann, T., Lammert, F., Casper, M., 2020. Deep-learning based detection of gastric precancerous conditions. Gut 69, 4–6. http://dx.doi.org/10.1136/gutjnl-2019-319347.

Hashimoto, R., Requa, J., Dao, T., Ninh, A., Tran, E., Mai, D., Lugo, M., El-Hage Chehade, N., Chang, K.J., Karnes, W.E., Samarasena, J.B., 2020. Artificial intelligence using convolutional neural networks for real-time detection of early esophageal neoplasia in Barrett's esophagus (with video). Gastrointest. Endosc. 91 (6), 1264–1271.e1. http://dx.doi.org/10.1016/j.gie.2019.12.049.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR '16, IEEE, pp. 770–778. http://dx.doi.org/10.1109/CVPR.2016.90, URL: http://ieeexplore.ieee.org/document/7780459.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., Gilmer, J., 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. ICCV.

Hendrycks, D., Dietterich, T., 2019. Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=HJz6tiCqYm.

Hendrycks, D., Lee, K., Mazeika, M., 2019a. Using pre-training can improve model robustness and uncertainty. In: Proceedings of the International Conference on Machine Learning.

Hendrycks, D., Mazeika, M., Kadavath, S., Song, D., 2019b. Using self-supervised learning can improve model robustness and uncertainty. arXiv:1906.12340.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D., 2021b. Natural adversarial examples. CVPR.

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: CVPR. IEEE Computer Society, pp. 2261–2269, URL: http://dblp.uni-trier.de/db/conf/cvpr/cvpr2017.html#HuangLMW17.

Iakubovskii, P., 2019. Segmentation models Pytorch. https://github.com/qubvel/segmentation_models.pytorch.

Islam, M., Li, Z., Glocker, B., 2023. Robustness stress testing in medical image classification. In: MICCAI-MedAGI Workshop.

Jaspers, T.J.M., Boers, T.G.W., Kusters, C.H.J., Jong, M.R., Jukema, J.B., de Groof, A.J., Bergman, J.J., de With, P.H.N., van der Sommen, F., 2023. Investigating the impact of image quality on endoscopic AI model performance. In: Applications of Medical Artificial Intelligence: Second International Workshop, AMAI 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 8, 2023, Proceedings. Springer-Verlag, Berlin, Heidelberg, pp. 32–41. http://dx.doi.org/10.1007/978-3-031-47076-9_4.

Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D., 2020. Kvasir-SEG: A segmented polyp dataset. In: Ro, Y.M., Cheng, W.-H., Kim, J., Chu, W.-T., Cui, P., Choi, J.-W., Hu, M.-C., De Neve, W. (Eds.), MultiMedia Modeling. Springer International Publishing, Cham, pp. 451–462.

Jiang, J., Jiang, X., Xu, L., Zhang, Y., Zheng, Y., Kong, D., 2023. Noise-robustness test for ultrasound breast nodule neural network models as medical devices. Front. Oncol. 13, http://dx.doi.org/10.3389/fonc.2023.1177225, URL: https://www.frontiersin.org/articles/10.3389/fonc.2023.1177225.

Karahan, Ş., Yildirim, M.K., Kirtaç, K., Rende, F.Ş., Bütün, G., Ekenel, H.K., 2016a. How Image Degradations Affect Deep CNN-Based Face Recognition?. Vol. P-260, Gesellschaft fur Informatik (GI), http://dx.doi.org/10.1109/BIOSIG.2016.7736924.

Karahan, S., Yildirum, M.K., Kirtac, K., Rende, F.S., Butun, G., Ekenel, H.K., 2016b. How image degradations affect deep CNN-based face recognition? In: 2016 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, http://dx.doi.org/10.1109/biosig.2016.7736924.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. URL: http://arxiv.org/abs/1412.6980, cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Kurakin, A., Goodfellow, I., Bengio, S., 2017. Adversarial machine learning at scale. arXiv:1611.01236.

Kusters, C.H.J., Boers, T.G.W., Jaspers, T.J.M., Jukema, J.B., Jong, M.R., Fockens, K.N., de Groof, A.J., Bergman, J.J., van der Sommen, F., de With, P.H.N., 2024. CNNs vs. Transformers: Performance and robustness in endoscopic image analysis. In: Wu, S., Shabestari, B., Xing, L. (Eds.), Applications of Medical Artificial Intelligence: Second International Workshop, AMAI 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 8, 2023, Proceedings. Springer Nature Switzerland, Cham, pp. 21–31.

Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., Lu, F., 2021. Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognit. 110, 107332. http://dx.doi.org/10.1016/j.patcog.2020.107332, URL: https://www.sciencedirect.com/science/article/pii/S0031320320301357.

Maron, R.C., Schlager, J.G., Haggenmüller, S., von Kalle, C., Utikal, J.S., Meier, F., Gellrich, F.F., Hobelsberger, S., Hauschild, A., French, L., Heinzerling, L., Schlaak, M., Ghoreschi, K., Hilke, F.J., Poch, G., Heppt, M.V., Berking, C., Haferkamp, S., Sondermann, W., Schadendorf, D., Schilling, B., Goebeler, M., Krieghoff-Henning, E., Hekler, A., Fröhling, S., Lipka, D.B., Kather, J.N., Brinker, T.J., 2021. A benchmark for neural network robustness in skin cancer classification. Eur. J. Cancer 155, 191–199. http://dx.doi.org/10.1016/j.ejca.2021.06.047, URL: https://www.sciencedirect.com/science/article/pii/S0959804921004421.

Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A.S., Bethge, M., Brendel, W., 2020. Benchmarking robustness in object detection: Autonomous driving when winter is coming. arXiv:1907.07484.

Navarro, F., Watanabe, C., Shit, S., Sekuboyina, A., Peeken, J.C., Combs, S.E., Menze, B.H., 2021. Evaluating the robustness of self-supervised learning in medical imaging. arXiv:2105.06986.

Orhan, A.E., 2019. Robustness properties of facebook's ResNeXt WSL models. CoRR abs/1907.07640 arXiv:1907.07640 URL: http://arxiv.org/abs/1907.07640.

Ozawa, T., Ishihara, S., Fujishiro, M., Kumagai, Y., Shichijo, S., Tada, T., 2020. Automated endoscopic detection and classification of colorectal polyps using convolutional neural networks. Therapeutic Adv. Gastroenterol. 13, 1756284820910659. http://dx.doi.org/10.1177/1756284820910659, arXiv:https://doi.org/10.1177/1756284820910659 PMID: 32231710.

Papernot, N., McDaniel, P.D., Goodfellow, I.J., 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. CoRR abs/1605.07277 arXiv:1605.07277 URL: http://arxiv.org/abs/1605.07277.

Paschali, M., Conjeti, S., Navarro, F., Navab, N., 2018. Generalizability vs. Robustness: Investigating medical imaging networks using adversarial examples. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Springer International Publishing, Cham, pp. 493–501.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library. URL: http://arxiv.org/abs/1912.01703.

Pei, Y., Huang, Y., Zou, Q., Zhang, X., Wang, S., 2021. Effects of image degradation and degradation removal to CNN-based image classification. IEEE Trans. Pattern Anal. Mach. Intell. 43, 1239–1253. http://dx.doi.org/10.1109/TPAMI.2019.2950923.

Recht, B., Roelofs, R., Schmidt, L., Shankar, V., 2019. Do ImageNet classifiers generalize to ImageNet? In: International Conference on Machine Learning.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. URL: http://arxiv.org/abs/1505.04597.

Rusak, E., Schott, L., Zimmermann, R.S., Bitterwolf, J., Bringmann, O., Bethge, M., Brendel, W., 2020. A simple way to make neural networks robust against diverse image corruptions. arXiv:2001.06057.

Sanderson, E., Matuszewski, B.J., 2022. FCN-transformer feature fusion for polyp segmentation. In: Annual Conference on Medical Image Understanding and Analysis. Springer, pp. 892–907.

Schölvinck, D.W., van der Meulen, K., Bergman, J.J.G.H.M., Weusten, B.L.A.M., 2017. Detection of lesions in dysplastic Barrett's esophagus by community and expert endoscopists. Endoscopy 49 (02), 113–120. http://dx.doi.org/10.1055/s-0042-118312, URL: http://www.thieme-connect.de/products/ejournals/abstract/10.1055/s-0042-118312.

Shao, R., Shi, Z., Yi, J., Chen, P.-Y., Hsieh, C.-J., 2022. On the adversarial robustness of vision transformers. arXiv:2103.15670.

Shen, C., Tsai, M.-Y., Chen, L., Li, S., Nguyen, D., Wang, J., Jiang, S.B., Jia, X., 2020. On the robustness of deep learning-based lung-nodule classification for CT images with respect to image noise. Phys. Med. Biol. 65 (24), 245037. http://dx.doi.org/10.1088/1361-6560/abc812.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 URL: http://arxiv.org/abs/1409.1556.

Srinivasan, V., Strodthoff, N., Ma, J., Binder, A., Müller, K.-R., Samek, W., 2021. On the robustness of pretraining and self-supervision for a deep learning-based analysis of diabetic retinopathy. arXiv:2106.13497.

Su, D., Zhang, H., Chen, H., Yi, J., Chen, P.-Y., Gao, Y., 2018. Is Robustness the Cost of Accuracy? – A Comprehensive Study on the Robustness of 18 Deep Image Classification Models: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XII. pp. 644–661. http://dx.doi.org/10.1007/978-3-030-01258-8_39.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R., 2013. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

Tajbakhsh, N., Gurudu, S.R., Liang, J., 2016. Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans. Med. Imaging 35 (2), 630–644. http://dx.doi.org/10.1109/TMI.2015.2487997.

Tan, M., Le, Q.V., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. CoRR abs/1905.11946 arXiv:1905.11946 URL: http://arxiv.org/abs/1905.11946.

Vasiljevic, I., Chakrabarti, A., Shakhnarovich, G., 2016. Examining the impact of blur on recognition by convolutional networks.

Wang, Z., Bai, Y., Zhou, Y., Xie, C., 2023. Can CNNs be more robust than transformers? In: International Conference on Learning Representations.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021a. SegFormer: Simple and efficient design for semantic segmentation with transformers. In: Neural Information Processing Systems (NeurIPS).

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021b. SegFormer: Simple and efficient design for semantic segmentation with transformers. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (Eds.), Advances in Neural Information Processing Systems. Vol. 34, Curran Associates, Inc., pp. 12077–12090, URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf.

Xie, C., Yuille, A.L., 2020. Intriguing properties of adversarial training at scale. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, URL: https://openreview.net/forum?id=HyxJhCEFDS.

Young, A.T., Fernandez, K., Pfau, J., Reddy, R., Cao, N.A., von Franque, M.Y., Johal, A., Wu, B.V., Wu, R.R., Chen, J.Y., Fadadu, R.P., Vasquez, J.A., Tam, A., Keiser, M.J., Wei, M.L., 2021. Stress testing reveals gaps in clinic readiness of image-based diagnostic artificial intelligence models. npj Digit. Med. 4 (1), 10. http://dx.doi.org/10.1038/s41746-020-00380-6.

Yuba, M., Iwasaki, K., 2022. Systematic analysis of the test design and performance of AI/ML-based medical devices approved for triage/detection/diagnosis in the USA and Japan. Sci. Rep. 12, http://dx.doi.org/10.1038/s41598-022-21426-7.

Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y., Choe, J., 2019. CutMix: Regularization strategy to train strong classifiers with localizable features. pp. 6022–6031. http://dx.doi.org/10.1109/ICCV.2019.00612.

Zanjani, F.G., Zinger, S., Piepers, B., Mahmoudpour, S., Schelkens, P., 2019. Impact of JPEG 2000 compression on deep convolutional neural networks for metastatic cancer detection in histopathological images. J. Med. Imaging 6, 1. http://dx.doi.org/10.1117/1.jmi.6.2.027501.

Zhang, Y., Sun, Y., Li, H., Zheng, S., Zhu, C., Yang, L., 2022. Benchmarking the robustness of deep neural networks to common corruptions in digital pathology. http://dx.doi.org/10.48550/ARXIV.2206.14973, URL: https://arxiv.org/abs/2206.14973.

Zhou, D., Yu, Z., Xie, E., Xiao, C., Anandkumar, A., Feng, J., Alvarez, J.M., 2022. Understanding the robustness in vision transformers. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (Eds.), Proceedings of the 39th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 162, PMLR, pp. 27378–27394, URL: https://proceedings.mlr.press/v162/zhou22m.html.