香港中文大學計算機科學與工程學系
Department of Computer Science and Engineering
The Chinese University of Hong Kong

The Chinese University of Hong Kong

Department of Computer Science and Engineering

Final Year Project

ESTR 4998 Graduation Thesis I

LYU2308 Evaluation of Multimodal Models:
Assessing Performance and Finding Improvements

Authors:
Haoran WU
Yushan WU

Supervised by:
Michael Rung Tsong Lyu

November 2023

1 Introduction

In the United States, medical errors rank as the third most common cause of death, claiming the lives of over 250,000 individuals annually[43]. Among these, diagnostic errors emerge as a particularly critical concern[22, 45]. The integration of medical imaging into computer systems has catalyzed the development of automated analysis tools, designed to bolster the precision of clinical diagnoses [39]. The evolution of Artificial Intelligence (AI) in this realm has significantly elevated the accuracy of these tools, transforming them into industrial-grade products. Techniques ranging from elementary image processing methods [60, 48] to advanced neural networks [28, 3] have contributed to this advancement.

Major technology corporations, recognizing the potential in this domain, have driven the AI healthcare market to a valuation of USD 16.3 billion in 2022, with a predominant focus on medical imaging diagnosis products [15]. IBM Watson Health and Google DeepMind are notable examples, having implemented AI-based tools in top hospitals and demonstrating superior performance in tasks such as breast cancer screening compared to human doctors [44, 49].

Despite the advanced capabilities of modern medical image diagnosis systems, challenges such as the misalignment of IBM Watson for Oncology with clinicians' assessments in gastric cancer cases highlight their fallibility [31]. Given the critical nature of medical diagnosis, the reliability of these AI-driven tools is paramount. Consequently, there is a pressing need for robust testing frameworks [71], akin to those for traditional software and other AI products like autonomous cars. The methodologies for generating test cases in general computer vision software cannot be directly applied to medical image diagnosis systems due to the unique complexities and real-world scenarios in medical contexts [78, 46, 63, 26, 67].

The scarcity of effective testing frameworks for medical image diagnosis software underscores the complexity of this challenge, necessitating specialized knowledge in medical and clinical domains for creating testing oracles. Additionally, existing image generation models and software, predominantly trained on datasets of natural images, face limitations in producing realistic and high-quality medical images vital for accurate testing.

The integration of AI in medical diagnostics has evolved beyond unimodal image analysis to encompass multimodal models that process and interpret diverse data types. These advanced AI tools, capable of integrating data from various modalities, promise to revolutionize medical diagnostics by offering a more comprehensive analysis than single-modality models. However, the testing of these multimodal models requires a nuanced approach that accounts for their complexity and the intricacies of multimodal data interpretation.

This report outlines the development of MedTest, a novel metamorphic testing paradigm crafted for medical image diagnosis software analysis, including both conventional academic SOTA models and large-scale multimocal models. A pilot study involving over 2,500 images from three hospitals has led to the identification of nine metamorphic relations across four artifact categories: lightness, motion, object, and non-object. These relations are integrated into MedTest to generate test cases that mirror real-world clinical scenarios, ensuring relevance and effectiveness in testing medical image diagnosis applications.

We have applied MedTest to both commercial software and state-of-theart (SOTA) algorithms, evaluating cutting-edge tools designed for medical diagnosis tasks, polyp segmentation as an example. The performance of these networks on original images versus images with introduced artifacts highlights significant variances, indicating areas for improvement. As we progress, our focus will shift to the challenges and opportunities presented by multimodal models, aiming to contribute significantly to the field of AI in medical diagnostics.

This paper's primary contributions are as follows:

- We introduce MedTest, the first comprehensive testing framework specifically designed for the validation of medical image diagnosis software. This framework represents a significant advancement in the field of medical imaging software testing.
- We Executed of a pilot study on $2,553$ real-world medical images, deriving 9 metamorphic relations, which are instrumental in the operationalization of MedTest.
- We applied 9 different types of perturbations on 4 datasets, involving more than $2,052$ images and generated $18,468$ images with artifacts.
- We provide an extensive evaluation of MedTest's effectiveness. This includes its application to various commercial medical image analysis software and several SOTA academic models. Our results demonstrate that MedTest can not only reliably trigger errors in these systems but also significantly enhance the robustness of SOTA algorithms, thereby contributing to the advancement of medical imaging technology.

2 Background

2.1 Medical Image Analysis

Over recent decades, a variety of medical imaging techniques, including computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), mammography, ultrasound, and X-ray, have played a crucial role in the early detection, diagnosis, and management of various diseases [7]. Traditionally, the interpretation of these medical images has predominantly been the domain of human experts, such as radiologists and physicians. However, the inherent variability in pathology, combined with the potential for human fatigue, has steered the medical community towards the adoption of computer-assisted interventions. Despite the slower pace of advancements in computational medical image analysis compared to the rapid developments in medical imaging technologies, recent strides have been notable, particularly with the integration of machine learning techniques.

In the realm of machine learning applied to medical image analysis, the identification or creation of informative features that accurately capture the patterns and regularities in the data is paramount. Historically, these features were crafted predominantly by human experts, drawing upon their domain-specific knowledge. This expert-driven approach posed significant challenges for those without specialized knowledge, limiting the broader application of machine learning in various medical studies. Parallel to this, there have been initiatives to establish sparse representations based on either predefined dictionaries or those learned from training samples. The concept of sparse representation, inspired by the principle of parsimony common across multiple scientific disciplines, suggests that simpler explanations for observations should be favored over more complex ones. This principle has been validated through the use of sparsity-inducing penalization and dictionary learning, demonstrating their efficacy in feature representation and selection within the field of medical image analysis [55, 69, 57, 13]. It is important to note, however, that the methods of sparse representation and dictionary learning mentioned in the literature still rely on a relatively shallow architectural approach to uncover patterns or regularities in data, thereby somewhat limiting their representational capacity.

Deep learning, on the other hand, has effectively surpassed these limitations [53]. It revolutionizes the traditional approach by incorporating the feature engineering process directly into the learning phase. In other words, deep learning does not necessitate manual extraction of features; rather, it requires only a dataset—sometimes with minimal preprocessing—and autonomously uncovers informative representations through a self-teaching mechanism [5, 35]. This shift has significantly reduced the burden of feature engineering from human experts to computational systems, enabling those without extensive expertise in machine learning to effectively utilize deep learning in their research and applications, particularly in the field of medical image analysis. This paradigm shift not only democratizes access to advanced analytical techniques but also opens new avenues for innovation and discovery in medical diagnostics and treatment planning.

2.2 Metamorphic Testing

Metamorphic testing, a well-established testing technique, addresses the oracle problem and has gained widespread recognition and application across various software domains [11]. The fundamental principle of metamorphic testing revolves around the identification and examination of MRs across successive executions of the software being tested. An MR essentially delineates the expected relationship between different sets of input-output pairs of a software application. In metamorphic testing, an initial test case is transformed into a subsequent, related test case through a predefined transformation rule. The software's responses to these test cases are then scrutinized to ascertain if they adhere to the anticipated relationship between outputs.

To illustrate, consider a software program designed to compute $\sin x$. One can leverage the well-known mathematical property stating "$\sin(\pi - x) = \sin x$" as a metamorphic relation for the sine function. Here, even if the exact expected value of $\sin x_1$ for a given source test case is unknown, a related test case $x_2 = \pi - x_1$ can be formulated. The equivalence of $\sin x_1$ and $\sin x_2$ can then be tested without prior knowledge of the specific outputs of these sine calculations. Any deviation from this metamorphic relation would suggest a possible flaw in the software's implementation of the sine function [54].

In recent years, metamorphic testing has been increasingly adapted to evaluate Artificial Intelligence (AI) software, with the goal of automatically identifying erroneous outcomes produced by AI applications through the development of innovative MRs. Notably, Chen et al. [12] explored the application of metamorphic testing in bioinformatics, demonstrating its potential in this specialized field. Xie et al. [73] established eleven MRs specifically designed to assess the performance of k-Nearest Neighbors and Naive Bayes algorithms. In a similar vein, Dwarakanath et al. [16] introduced eight MRs for the testing of SVM-based and ResNet-based image classifiers, showcasing the versatility of metamorphic testing in handling various AI models. Furthermore, Zhang et al. [78] applied this technique to autonomous driving systems, utilizing Generative Adversarial Networks (GANs) to generate diverse driving scenes under varying weather conditions, subsequently evaluating the consistency and reliability of the systems' outputs in these simulated environments. This broadening scope of metamorphic testing, particularly in AI software validation, underscores its growing importance and utility in ensuring the robustness and reliability of increasingly complex software systems.

3 MedTest

In this section, we commence with an insightful pilot study, which delves into an analysis of authentic medical images that have been sourced directly from hospital environments (as detailed in Section 3.1). This preliminary exploration sets the stage for the subsequent introduction of nine metamorphic relations (MRs). These relations, derived and inspired by the findings of the pilot study, represent a significant step in understanding and evaluating medical image analysis processes.

We have meticulously categorized these nine MRs into four distinct groups, each based on the type of perturbation they involve. The first category focuses on lightness perturbations, where we examine how variations in image brightness and contrast can impact medical image analysis (discussed in Section 3.2). The second category, motion perturbations, explores the effects of simulated motion artifacts such as blurring, which can occur during image capture in dynamic clinical settings (covered in Section 3.3).

The third category revolves around object perturbations (Section 3.4), where the emphasis is on alterations related to the objects within the medical images. This includes changes in the size, shape, or position of clinically relevant features within the image. The final category, non-object perturbations (Section 3.5), addresses modifications that do not directly involve the primary objects of interest in the images. This could include alterations to background elements or other aspects of the image that, while not directly related to the primary diagnostic features, may still influence the overall analysis process.

Each of these categories plays a pivotal role in understanding how different types of perturbations can affect the accuracy and reliability of medical image analysis, thereby contributing to the enhancement of diagnostic processes and tools in the healthcare sector. This structured approach allows for a comprehensive exploration of the complexities involved in medical image analysis and paves the way for developing more robust and reliable diagnostic methodologies.

3.1 Pilot Study

In our research, we set out with the ambitious goal of developing a set of MRs tailored to the field of medical imaging. These MRs are designed on the premise that a 'seed' test case (an original medical image) and its 'perturbed' counterpart (the same image but with added artifacts) should yield consistent classification labels or similar segmentation masks when analyzed by medical image analysis software. To ensure that these test cases are both effective and relevant, we have established a set of criteria for the perturbations incorporated in our MRs, which include:

- *Clinical-semantic-preserving*: This criterion ensures that the perturbed test cases should maintain the integrity of the analysis results, matching those of the original seed image.

- *Realistic*: The perturbations should closely mirror the types of artifacts encountered in actual clinical settings.
- *Unambiguous*: Clarity and precision in definition are key, ensuring that the perturbations are well-defined and easily interpretable.

To establish a foundation for designing these perturbations, we embarked on a pilot study focusing on the types of artifacts typically encountered in medical images used in real-world clinical scenarios. This involved an extensive review of 103 endoscopic videos sourced from three hospitals. From these videos, we extracted $2,553$ individual images. We then engaged ten highly qualified annotators, each holding at least a postgraduate degree in medicine, to meticulously label these images. These annotators underwent thorough training, including guidelines, test tasks, and sessions specifically focused on endoscopic images and the identification of artifacts. During the annotation process, each image was evaluated to determine the presence of any artifact. The consensus among the annotators was used to establish the final human label, resulting in a dataset of $1,199$ endoscopic images identified as containing artifacts.

Upon detailed examination of these artifact-laden images, we identified and summarized 9 distinct methods of perturbation, commonly encountered in clinical settings. These methods are categorized from different perspectives: 1) those related to endoscopic imaging cameras, including lightness and motion perturbations; and 2) those pertaining to the visual content within the endoscopic images, such as object and non-object perturbations. Building on these insights, we formulated nine corresponding MRs, each based on a specific perturbation method. As shown in Table 1, we introduce 4 different perturbation groups, i.e. lightness, motion, objects, and non-objects, where each group includes at least one

Table 1: Categorization of Perturbation Types in Medical Images: A Pilot Study.

Perturbation Group	Type	Description
Lightness	Saturation	Over-saturation caused by excessive lighting
	Contrast	Resulting from underexposure or obstructions in the field of view
	White Balance	Color distortions due to presence of white objects
	Specularity	Reflections resembling a mirror-like surface
Motion	Blur	Blurring from hand movements or rapid camera motion
Objects	Instrument	Presence of surgical instruments in the image frame
	Feces	Incomplete colon cleansing in patients
	Blood	Visible bleeding from wounds
Non-objects	Text	Embedded clinical information related to patients

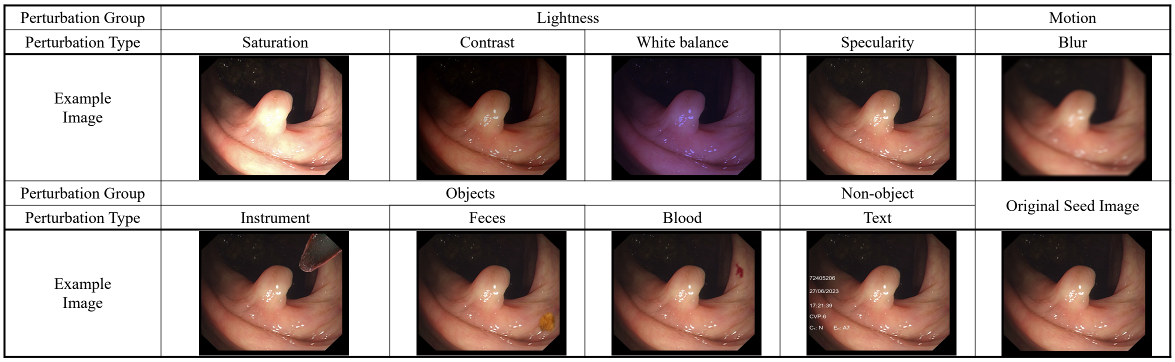


Figure 1: The visualization of the different perturbations groups.

perturbation type. Fig. 1 demonstrates the visual perturbed images of different perturbation types. According to these MRs, the diagnostic label assigned by the medical analysis software to a perturbed endoscopic image (i.e., the generated test case) should align with the label given to the original, unperturbed seed image. Through this approach, we aim to rigorously test and validate the robustness and reliability of medical image analysis software, ensuring its effectiveness even in the presence of common clinical artifacts.

3.2 MRs with Lightness Perturbations

These MRs leverage the lightness perturbations that imitate the various illumination conditions during the endoscopic camera imaging.

3.2.1 MR1-1 Saturation

To address saturation issues in endoscopic imaging, a key concern is the proximity of the light source to colon tissue. Overexposure can occur when the light source is too close, leading to saturation artifacts. Our method for simulating this effect involves applying variable levels of saturation to endoscopic images to mimic different degrees of overexposure. This is achieved by adjusting the saturation of an image using a random factor selected from a predefined range $[s_1, s_2]$.

The process utilizes the torchvision library functions that control brightness, contrast, and saturation. We define a fluctuation range and randomly select a saturation factor within this range, with a bias towards values greater than 1 to replicate the overexposure effect. This factor is then used to modulate the saturation level, where a value of 1 indicates no change, values less than 1 decrease saturation, contrast, and brightness, and values greater than 1 increase them, thereby simulating the impact of light source proximity on the colon tissue.

3.2.2 MR1-2 Contrast

In the context of endoscopic examinations, the distance between the colon tissue and the light source, or obstructions, can result in underexposure. To simulate this scenario, our method focuses on altering the contrast of endoscopic images. Beginning with a seed endoscopic image, we establish a contrast range denoted as $[c_1, c_2]$. A value is then randomly selected from this range, which is used to adjust the image's contrast level.

This technique parallels the approach used for saturation adjustments, but with an inclination towards lower levels of contrast, brightness, and saturation, corresponding to the underexposed nature of the images. By carefully modulating the contrast in this manner, we aim to authentically replicate the conditions of underexposure commonly encountered during endoscopic procedures.

3.2.3 MR1-3 White Balance

In endoscopic imaging, we often observe color biases, predominantly manifesting as green or purple hues. The likely reason for these color biases could be attributed to the white balance settings of the endoscopic camera or the lighting conditions within the endoscopic environment, which may not always accurately represent the true colors of the tissue.

To simulate these white balance discrepancies in endoscopic images, we selectively modify the RGB channels. For images with a green bias, we reduce the red and blue channels by approximately half of their original values, maintaining their proportional relationship. Similarly, for images exhibiting a purple color bias, we decrease the values of both the red and green channels proportionately. This method allows us to realistically replicate the color distortions that might occur due to white balance issues in endoscopic imaging.

3.2.4 MR1-4 Specularity

The observed phenomena indicate that the manifestation of spots, attributable to specular reflection, predominantly occurs in a compact region as opposed to being dispersed throughout the entire image. Our initial approach involved identifying clusters as potential sites for spot generation. Subsequently, we introduced circle, ellipse, and distorted circle as the potential shape for generating white spots. Spots are generated at random locations near the cluster centers, with randomly chosen radius bounded by λ times image height in order to control the size of spots comparing with the image size. After trials and error, we found that the elliptical spots can achieve a best effect, creating most realistic specularity on the endoscopy images. The elliptical spots are decided by the following formula:

$$\frac{(x-\bar{x})^2}{(a+\epsilon)^2} + \frac{(y-\bar{y})^2}{(b+\epsilon)^2} = 1 \tag{1}$$

where $\bar{x}$ and $\bar{y}$ denote the center of the ellipse, and a and b are the two axes of the ellipse respectively. The additional ϵ acts as a term to avoid zero denominator due to the randomized selection of parameters. The following process involves the application of Gaussian blur to facilitate their seamless integration into the image. Additionally, we integrated these spots with a gray mask, derived from our algorithm, to modulate their intensity, particularly ensuring they do not exhibit excessive brightness in the darker regions of the image.

3.3 MRs with Motion Perturbations

3.3.1 MR2-1 Blur

We have noted that possible camera movement and tissue movement when the image is captured can often cause motion blur in images. To replicate this phenomenon, we employed Gaussian Blur, a technique that involves convolving each pixel of the image with a Gaussian function. The blurring degradation is defined as following:

$$x' = x \cdot G_B(r_B, \sigma_B) + n \tag{2}$$

where G_B is a Gaussian filter with a radius r_B and a spatial constant σ_B, and n is the random Gaussian noise added to the image. [56] In our implementation, we first generate a random number in the range of

(0, 15] as the σ value. Based on the chosen σ, we randomly chose odd integers within the range of $\frac{\sigma}{3}$ to $\frac{\sigma}{2}$ for the width and height of the kernel of the Gaussian filter, respectively. This process effectively blends each pixel with the information from its neighboring pixels, creating a blurring effect reminiscent of a weighted average of the surrounding area. By using Gaussian Blur, we can simulate the kind of blur typically introduced by camera motion, enhancing the realism of our simulated images.

3.4 MRs with Object Perturbations

3.4.1 MR3-1 Instrument

In this study, we utilized the Kvasir-Instrument dataset [27], which comprises 590 images featuring medical instruments and their corresponding segmentation masks. Our initial step involved extracting these instruments from the original images and documenting their positions. Subsequently, we employed our algorithm to identify an optimal target area for each instrument, ensuring it met the following criteria:

- Avoidance of overlap with the Polyp.
- Preservation of a position and orientation akin to those in the original image.
- Maintenance of an appropriate size, neither excessively large nor small.

Finally, we repositioned the extracted instruments into these target areas. To enhance realism, we applied Gaussian blur and integrated our blending algorithm, ensuring a natural appearance of the instruments in their new context.

3.4.2 MR3-2 Feces

In this section, we employed fecal matter images extracted from the Kvasir dataset [29] using Meta's Segment Anything algorithm [33]. Similar to the aforementioned method used for instruments, we replicated this approach, albeit without constraints on the positioning and orientation of the feces. Crucially, we calculated a brightness ratio by comparing the fecal matter with the target image, enabling us to adjust the feces' brightness for a more coherent integration. Furthermore, to prevent excessive brightness in particularly dark areas of the target image, we again utilized the gray mask previously mentioned in the context of specular reflections, providing an additional layer of realism to the adjusted fecal images.

3.4.3 MR3-3 Blood

In this phase, we focused on the blood images and their associated masks from the EAD2020 dataset [52, 2, 1]. Our methodology mirrored the approach previously described for pasting feces, with an emphasis on modifying various lighting parameters. This adjustment was crucial to enhance the natural appearance of the blood when integrated into the target images, ensuring a realistic representation in the context of the dataset.

3.5 MRs with Non-Object Perturbations

3.5.1 MR4-1 Text

Our analysis revealed a consistent pattern in the text displayed on endoscopic images, as illustrated in figure 2. Although the specific position and content of the text varied across images, it predominantly comprised

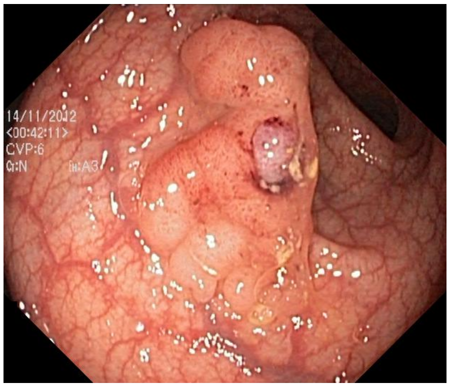


Figure 2: Pattern of the text in Kvasir dataset

temporal data and device parameters. To replicate this characteristic, we employed the ImageDraw method, generating text that adhered to the observed pattern through random generation, thereby maintaining consistency with the original text format in the images.

4 Evaluation

To rigorously assess the efficacy of MedTest, our methodology has been applied to four SOTA algorithms specifically designed for medical image diagnosis focusing on polyps. Additionally, we have plans to extend this evaluation to include two commercial software products and other SOTA algorithms across a variety of diagnostic tasks. This section is dedicated to exploring and providing insights into four critical Research Questions (RQs), which are as follows:

- RQ1: Does MedTest generate test cases that are diagnostically consistent with the original seed images and maintain a realistic appearance?
- RQ2: Is MedTest effective in identifying incorrect outputs produced by medical image diagnosis software and algorithms?
- RQ3: Can the test cases generated by MedTest be utilized to enhance

the performance of medical image diagnosis software?

- RQ4: What are the various factors that influence MedTest's performance and how do they do so?

MedTest is designed to create test cases that are not only clinically equivalent to the original seed images but also replicate the realistic nature of artifacts found in actual clinical scenarios. Thus, in addressing RQ1, we aim to validate whether the perturbations introduced in the test cases preserve the clinical diagnosis and realism, as assessed by human annotation.

In RQ2, our objective is to determine the capacity of MedTest to consistently and effectively trigger errors in medical image diagnosis systems, encompassing both the chosen commercial software and SOTA algorithms.

Moving forward, the discovery of errors naturally leads to their rectification. Hence, RQ3 focuses on exploring how the test cases generated by MedTest can be leveraged to improve the functionality and accuracy of medical image diagnosis tools.

Given that MedTest represents a pioneering approach in testing medical image diagnosis products, RQ4 is dedicated to examining the impact of various factors present in medical images on the performance of MedTest. This analysis aims to provide a comprehensive understanding of how these diverse elements can affect the efficacy of MedTest in evaluating the tested software and algorithms. By delving into these research questions, we endeavor to not only validate the effectiveness of MedTest but also to contribute significantly to the advancement of medical image diagnosis technology.

4.1 Experimental Settings

4.1.1 Datasets

In our endeavor to thoroughly validate MedTest, we have utilized a diverse array of datasets as seed data, drawing upon the extensive work of previous researchers who have meticulously collected, labeled, and made available various types of data for research applications. For the purposes of this paper, we have specifically chosen the most widely used datasets in the field of polyp segmentation, all of which are publicly accessible. These include CVC-300 [66], CVC-ClinicDB [6], CVC-ColonDB [59], and Kvasir [29]. There are in total 2052 images combined.

CVC-300, a subset of the larger EndoScene dataset, is a relatively compact dataset comprising 60 images, each with dimensions of 578×500 pixels. In addition to CVC-300, the EndoScene dataset also encompasses images from the CVC-ClinicDB dataset. To maintain clarity and precision in our analysis, we have treated these two datasets as distinct entities in separate experiments, meticulously recording and analyzing their respective results [38].

CVC-ClinicDB, also known as CVC-612, is a more extensive collection, featuring 612 publicly available polyp images sourced from 25 different colonoscopy videos. The images in this dataset are of the size 384×288 pixels, offering a distinct set of characteristics for analysis.

The CVC-ColonDB dataset is composed of 15 different endoscopy sequences, totaling 380 polyp images. Each image in this dataset shares the same resolution as the CVC-300 dataset, specifically 578×500 pixels.

Lastly, the Kvasir dataset, a more recent addition to the field, stands out due to its large scale, diverse endoscopy scenes, and varied polyp shapes.

This diversity renders the segmentation task particularly challenging. The images in Kvasir vary considerably in size, ranging from 332 × 487 to 1920 × 1072 pixels. This variability not only presents a significant challenge for medical diagnosis software but also adds complexity to our method of generating simulated artifacts [29] [38]. The selection of these datasets for our validation process reflects our commitment to ensuring that MedTest is rigorously tested against a wide spectrum of real-world scenarios, thereby ensuring its robustness and applicability in diverse clinical settings.

4.1.2 Pre-process of the Datasets

To ensure a uniform approach in our analysis, we initially undertook the task of standardizing the dataset. This involved pre-processing both the images and their corresponding segmentation masks to a consistent size of 512 × 512 pixels, a dimension commonly accepted and utilized by various medical diagnosis algorithms. This standardization is crucial for maintaining consistency across different datasets and ensuring that the input to the medical diagnosis algorithms is uniform, thus allowing for more accurate comparisons and evaluations.

A notable characteristic of most endoscopy images is the presence of a black frame around the edges, which typically lacks a consistent pattern. This irregularity renders traditional image processing techniques, such as thresholding and region growing, ineffective for their extraction. To address this challenge, we developed a specialized model specifically designed to extract these black frames from the images. This extraction is vital, as it enables us to mask out any potential synthesized artifacts that may appear on these black edges, thereby ensuring the integrity and realism of the images used in our tests.

In addition to frame extraction, our pre-processing model plays a crucial role in assessing the brightness levels across different areas within the images. By employing a sigmoid function, we generate gray masks that reflect these brightness variations. These gray masks are then strategically used in the process of artifact addition. They allow for precise adjustments in color and brightness of the synthesized components, effectively preventing the creation of images with overtly unnatural or artificial effects. This meticulous approach to pre-processing and artifact integration is fundamental to our objective of producing realistic test cases that accurately mimic real-world clinical scenarios. It ensures that our testing environment closely replicates the conditions under which medical diagnosis software is typically employed, thus providing a robust and reliable framework for evaluating the performance and efficacy of these algorithms.

4.1.3 Software and Models Under Test

We use MedTest to test commercial medical image diagnosis software products and SOTA academic models. Commercial software products include ChatGPT and Bard, on which we want to test the performance on visual question answering (VQA) based on polyp diagnosis on endoscpoy images. SOTA academic models consist of PraNet [17], SANet [72], TGANet [64] and SSFormer [68], all targeting the polyp segmentation task on endoscopy images.

PraNet PraNet model is a well-established model on polyp segmentation task and paved the path for the later ones. The special design on PraNet is that it uniquely combines high-level feature aggregation via a Parallel Partial Decoder (PPD) and detailed segmentation through Reverse Attention (RA) modules. This approach can enable the model to effectively handle the variability in polyps' size, color, and texture, as well as the often-blurred boundaries in colonoscopy images. PraNet has demonstrated excessive performance over existing methods in terms of segmentation accuracy, generalizability, and real-time efficiency. Because it is one of the initial influential model, we decided to investigate deep into it and evaluate its robustness.

SANet Besides, we explored the Shallow Attention Network (SANet), also designed to address key challenges in polyp segmentation but with more unique designs and implementation to handle detailed problems in previous studies. SANet innovatively tackles issues like inconsistent color distributions in samples, degradation of small polyps due to repeated downsampling, and imbalance between foreground and background pixels. Based on this idea, the model employs a color exchange operation to reduce overfitting by decoupling image content from color, enhancing focus on shape and structure. It also introduces a shallow attention module to filter background noise in shallow features, which helps preserve small polyps more effectively. Additionally, the probability correction strategy during inference improves model performance, especially for small polyps. SANet's extensive testing across five benchmarks shows its outstanding capability in polyp segmentation task, suitable for us to evaluate.

TGANet We also investigated the TGANet model, which focuses on enhanced polyp segmentation in colonoscopy images using the auxiliary text input as additional information. The model aimed at the challenges posed by the variability in polyp size and number, which can impact the effectiveness of segmentation models. Targeting this, TGANet innovatively employs text-guided attention mechanisms, leveraging attributes like polyp size and count through additional text input to improve segmentation accuracy. The text input containing the polyp information serves as an auxiliary classification task and further enhance model's learned representations of important features within the image. After that, a feature enhancement module and multi-scale feature aggregation within the network are present to allow for more precise adaptation to varying polyp characteristics. With these implementation, especially the module to incorporate text description information, the model is expected to have better performance because of the excessive learning of image feature representations. As is describe in their inference part, text input is unnecessary, so we leverage this novel design to test our evaluation framework MedTest.

SSFormer We also explored the SSFormer model, which also targeted the challenges imposed by the complex and diverse structure of polyps image and the varying shapes of poly. These problems, together with the indistinctive bound between polyp and other categories, make the whole segmentation task difficult and the learning on existing dataset prone to over-fitting. This model stands out by incorporating a pyramid Transformer encoder, significantly enhancing the model's generalization capabilities. The Progressive Locality Decoder (PLD) in it emphasizes local features while integrating them into global features. This can effectively address the common issue of attention dispersion in Transformer models. Such delicate design improves the detail processing ability of the neural network and allows the establishment of its SOTA performance in polyp segmentation tasks. Because this model demonstrates exceptional learning and generalization abilities on unseen datasets, we want to test whether its performance is robust enough on our MedTest.

4.2 RQ1: Are the test cases generated by MedTest diagnosis-identical to seed images and realistic?

In this study, MedTest is designed with the specific objective of creating test cases that not only yield identical diagnostic results compared to their corresponding seed images but also closely resemble the types of artifacts encountered by medical professionals in real-world clinical settings. To assess the effectiveness of MedTest in achieving these goals, we conducted an evaluation based on human annotations.

For this purpose, we generated a sample set of 100 images for each perturbation method. This resulted in a total of 900 uniquely generated images. To ensure a thorough and expert evaluation, we enlisted the help of three annotators. Each of these annotators possesses postgraduate qualifications in fields related to medical or radiology and is proficient in English. Prior to the annotation process, these annotators were provided with comprehensive guidelines and training sessions to familiarize them with the specific requirements of the task.

The annotators were then tasked with examining each pair of images, consisting of an original and its corresponding perturbed counterpart. Their evaluation was guided by two key questions for each image pair:

- On a scale from ”1 (strongly disagree)” to ”5 (strongly agree)”, to what extent do you believe that the perturbed image maintains the identical diagnostic outcome as the seed image?
- On the same scale, how realistically do you think the perturbation reflects what might occur in actual clinical scenarios?

Any test cases that elicited disagreements among the annotators or were flagged as unrealistic were subjected to further review. The results of this annotation process were quite revealing. On average, the images scored 4.5 for maintaining identical diagnostic results and 4.79 for the realism of the perturbations. To quantify the level of agreement among our annotators, we employed Randolph's Kappa, a statistical measure commonly used for assessing inter-rater reliability. The resultant kappa value indicative of "almost perfect agreement", as defined in the established guidelines of the field [34].

This rigorous evaluation underscores the high degree of fidelity and realism that MedTest achieves in simulating clinical artifacts, as well as its effectiveness in maintaining diagnostic consistency. This validation is crucial, as it demonstrates the potential of MedTest to serve as a reliable tool for testing and improving medical image diagnosis software, ensuring its readiness for practical application in clinical environments.

Answer to RQ1: The test cases generated by MedTest are diagnosis-identical to seed images and realistic.

4.3 RQ2: Can MedTest find erroneous outputs returned by medical image diagnosis software?

MedTest aims to automatically generate test cases to find potential errors in current medical image diagnosis software. Hence, in this section, we evaluate the number of errors that MedTest can find in the outputs of commercial software and academic models.

We first input all the original seed images and obtain the original output for each software product or model under test. Then we conduct perturbations in MedTest's MRs described in section 3 on the seed image to generate test cases. Finally, we use the generated test cases to validate the software products and academic models. In particular, we check whether the generated test cases have identical diagnosis results as the corresponding seed images. If not, the diagnosis-identical perturbation affects the diagnosis of the software products or academic models, indicating erroneous outputs.

To evaluate how well MedTest does on generating test cases that trigger errors, we calculate Error Finding Rate (EFR), which is defined as follows:

$$\mathrm{EFR} = \frac{\text{Number of misclassified test cases}}{\text{Number of generated test cases}} * 100\%. \tag{3}$$

Since we are currently testing on polyp segmentation task, here we apply two similarity coefficients, Dice score and IoU score, which are proved to be simple and useful summary measures of spatial overlap and can measure the accuracy in image segmentation [80]. A test case is considered misclassified when its scores, both Dice and IoU, are 50% less than the scores tested on the seed image. The Dice score is given by

$$\begin{aligned} \mathrm{Dice}(\hat{Y}, Y) &= \frac{2 \times |\hat{Y} \cap Y|}{|\hat{Y}| + |Y|} \\ &= \frac{2 \times TP}{(TP + FP) + (TP + FN)}. \end{aligned} \tag{4}$$

And the IoU score is given by

$$\begin{aligned} \mathrm{IoU}(\hat{Y}, Y) &= \frac{|\hat{Y} \cap Y|}{|\hat{Y} \cup Y|} \\ &= \frac{TP}{TP + FP + FN}. \end{aligned} \tag{5}$$

In both equations 4 and 5, the $\hat{Y}$ stands for the predicted segmentation mask output by the models, while Y is the ground truth segmentation mask. Here, the TP, FP, and FN are all calculated pixel-wise on the masks.

Based on these two similarity coefficients, we define a segmentation output case as "missclassified" or "error" when difference between the performance of the seed image and that of the synthesize image to be larger than a proportion of the performance of the seed image. We set the proportion to be a threshold t. Our definition is therefore by the following:

$$\frac{OriginalScore - ArtifactScore}{OriginalScore} > t \tag{6}$$

where $OriginalScore$ represents the Dice/IoU score calculated from the seed image and $ArtifactScore$ represents the Dice/IoU score calculated from the synthesized image with specific artifact. In the table showed in later sections, we recorded the statistics for choosing both 50% and 25%.

Regarding the EFR, we found that it can be analyzed in different dimension, regarding different artifacts, datasets and models. The EFR varies in different experiment settings and we will use the following sections to illustrate the influence of above factors, with our proposed explanations for the situations.

4.3.1 Artifact Types

Based on our observations on the experiment results, we found evident discrepancy between the degree to which the models are influenced based on different artifact types.

Generally, light-related perturbations, including white balance, saturation and contrast, may affect the model performance more severely, that is, leading to a higher EFR. Especially, we discovered that models ability to properly segment polyps reduce more dramatically when the boundary of the polyp is unclear. A common case is when synthesized images in contrast (underexposure) category are input to the model, the model cannot distinguish the polyps with their surrounding, thus outputting unsatisfying prediction masks. White balance may also affect the model performance in a relatively similar way. Because we adjust the RGB channel to imbalance by reducing the undesired channel values to make the value of the vital channel outweigh other, this at the same time make the image darker (as the channel value reduces). We suspected that this is one reason why white balance biases can sometimes lead to most dramatic drop in model performance. Besides, when using the test cases with generated saturation, the rise in EFR may result from the scenario that polyps are present at the area with overexposure. In this case, the color in the overexposed areas will tend to white, making the model unable to distinguish the polyp boundary.

Blurring effect may sometimes lead to higher EFR, especially when testing on PraNet. We think the mechanism in confusing the model may be similar to the light perturbations, that is, making the edge of the polyp unclear so that the model will regard it as normal tissue as its surroundings or segmenting a much larger area with the unimportant tissues.

Surprisingly, the EFR on object perturbations do not affect models as seriously as we expected, though constantly triggering a small amount of error. The major reason for such object-based artifacts to influence model performance is mislead the model to misinterpret them as ”polyps” that should be segmented out. Indeed, based on our observations on the model output predictions, this was usually the case when model performance decreased. However, because of the relatively obvious difference between these objects and polyps, their ability to fool the models is limited.

4.3.2 Dataset

The performance of different models on our customized dataset is also different. In most cases, models tend to perform better on images generated on CVC-ClinicDB and CVC-300. One possible explanation is that these two datasets are usually included in the training data of models targeting the polyp segmentation tasks. Therefore, even though with our synthesized artifacts, models have the sufficient ability to comprehend and transform the image features into proper representations for further predictions, resulting a relatively more robust performance on such datasets. In this case, our EFR are also lower as a consequence.

On the contrary, models tend to be less robust on CVC-ColonDB and Kvasir. For CVC-ColonDB, it is less common to be used in training, as for the models we have surveyed and evaluated. It is a reasonable phenomenon as the models have not encountered similar images and may not learn sufficient features regarding images within this dataset. As for Kvasir, the images within this dataset have fairly different features in the overall image layout and visual effect from the other CVC datasets, imposing huger difficulties for models to learn how to extract suitable features from these images. With the synthesized effects on our customized dataset, this problem always turned out to be more severe. Therefore, the overall model performance on our synthesized images with artifacts is worse than other datasets, producing higher EFR on such settings.

4.3.3 Models for Evaluation

Though the models selected for evaluation are targeting the same task, i.e. polyp segmentation, and there are all accepted by top conferences or have high citations, vary in their emphasis on different components in the network designs. Therefore, the difference between their robustness should be discussed.

PraNet: PraNet, renowned in the domain of Polyp segmentation, exhibits commendable performance on the original dataset. Its Error Finding Rate (EFR) on the Dice score, with a threshold of 0.25, stands at 4.38%, indicative of its relative robustness. However, the inclusion of CVC-ClinicDB and Kvasir datasets in PraNet's training set may predispose the results to bias. A more critical examination using the CVC-ColonDB dataset reveals a heightened EFR of 8.56%. Focusing on the CVC-ColonDB analysis, PraNet demonstrates proficiency in handling Specularity and Blood artifacts, but shows vulnerability to White Balance and Blur. This suggests a higher resilience to object-based distortions as opposed to those induced by lighting variations. 4.3.3 4.3.3

SANet: SANet, introduced a year subsequent to PraNet, is evaluated for its enhanced robustness. The EFR of SANet on the Dice score, with the threshold set at 0.25, is recorded at 1.7%. To mitigate the potential bias from images in the training set, we scrutinized its performance on the CVC-ColonDB, where the EFR is noted to be 5.43%. SANet exhibits a markedly reduced EFR across most artifact categories in the CVC-ColonDB. It is predominantly impacted by White Balance and Blur, while demonstrating greater resistance to artifacts related to Blood, Saturation, and Contrast. 4.3.3 4.3.3

TGANet TGANet has a special design of incorporating the text embedding to provide additional information to enhance feature representations. We found that its performance is unsatisfying on both the original seed images and the synthesized ones, and our synthesized image inputs have triggered even more errors. When threshold t is set to 0.25, the network produced an EFR up to 15.70% on the overall 4 datasets. Among all artifacts, EFR was much higher on White Balance, Instrument, and Blur. Contrast and Blood artifacts also sometimes caused severe problems on specific datasets. Our conjecture on the explanation is that the model were trained using insufficient data and the presence of auxiliary text inputs may result in the overfitting problem when learning feature representations. 4.3.3 4.3.3

SSFormer SSFormer is the most robust models when tested using the synthesized images. When we relax the threshold to $t = 0.25$, the EFR is only 1.47% for the whole 4 datasets when calculating using Dice Score. As the newly released model, SSFormer showed its robustness and strong capability in addressing the task even when faced with bad image conditions. As can be found in table 4.3.3 4.3.3, many of the artifacts only trigger a small number of errors in specific datasets. As is the common case in other experiments, artifacts synthesized on seed images in CVC-ColonDB, as a dataset seldom used in training, can confuse the model the most and generate more errors consequently. Light-reltaed perturbations, including White Balance, Contrast, and Saturation, are able to find corner cases most often, which exactly aligns with our previous conjectures. Also, SSFormer's performance on images synthesized with Blood also decreased, which may suggest that blood in medical images has the potential to fool the model into misclassification even on a relatively robust model.

Using the statistics present in tables 4.3.3 4.3.3, we calculated the EFR for all four academic models we have tested, including PraNet, SANet, SSFormer, TGANet, and ther EFR are 4.38%, 1.70%, 1.47%, and 15.70%,

PraNet	CVC-300		CVC-ClinicDB		CVC-ColonDB		Kvasir	
t=0.5	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
Blood	0.0	0.0	0.0	0.0	1.1	1.3	0.1	0.3
Feces	0.0	0.0	0.0	0.0	3.7	5.0	0.0	0.2
Instrument	3.3	5.0	1.1	2.6	7.9	9.8	0.1	0.3
Spot	0.0	0.0	0.2	0.2	1.8	1.8	0.1	0.1
Saturation	3.3	6.7	0.7	0.7	4.2	4.5	2.3	3.7
Contrast	0.0	0.0	0.3	0.3	4.0	4.2	0.3	0.6
White Balance	3.3	3.3	7.4	10.9	15.0	16.9	2.9	4.9
Blur	1.7	1.7	5.2	8.0	7.9	11.3	7.4	11.7
Text	0.0	0.0	0.5	0.5	2.4	2.6	0.0	0.1

Table 2: EFR(%) of PraNet Model on Various Datasets with $t = 0.5$

SANet	CVC-300		CVC-ClinicDB		CVC-ColonDB		Kvasir	
t=0.5	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
Blood	0.0	0.0	0.0	0.0	1.6	1.8	0.0	0.0
Feces	0.0	1.7	0.0	0.2	5.0	5.5	0.1	0.1
Instrument	0.0	0.0	0.0	0.0	3.4	3.7	0.0	0.0
Spot	0.0	0.0	0.2	0.2	2.1	2.6	0.0	0.0
Saturation	3.3	3.3	0.2	0.5	2.4	2.4	0.9	1.3
Contrast	0.0	0.0	0.0	0.0	2.4	2.4	0.0	0.0
White Balance	0.0	0.0	1.5	2.3	8.2	9.0	1.0	3.4
Blur	0.0	0.0	0.2	0.2	2.6	3.4	0.3	0.6
Text	0.0	0.0	0.2	0.2	4.0	4.0	0.0	0.0

Table 3: EFR(%) of SANet Model on Various Datasets with $t = 0.5$

respectively. We can clearly see that SANet and SSFomer are relatively more robust with much lower EFR on our synthesized images, While PraNet and TGANet performed worse and more errors are triggered using synthesized images generated from our framework MedTest.

Detailed visualization of the artifacts and the corresponding output are illustrated in Table10 and Table11, where PraNet and SANet are used as example models.

SSFormer	CVC-300		CVC-ClinicDB		CVC-ColonDB		Kvasir	
t=0.5	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
Blood	0.0	1.7	0.2	0.2	3.2	3.9	0.0	0.0
Feces	0.0	0.0	0.0	0.0	4.7	5.8	0.0	0.0
Instrument	0.0	1.7	0.5	1.1	5.8	5.8	0.0	0.0
Spot	0.0	0.0	0.2	0.2	1.6	1.8	0.0	0.0
Saturation	6.7	6.7	0.8	1.8	1.6	2.1	0.2	0.4
Contrast	0.0	0.0	0.2	0.2	3.4	3.4	0.1	0.2
White Balance	0.0	1.7	2.5	3.9	9.5	10.5	0.9	1.5
Blur	0.0	0.0	0.2	0.2	2.4	2.6	0.1	0.2
Text	0.0	0.0	0.2	0.2	0.8	1.6	0.0	0.0

Table 4: EFR(%) of SSFormer Model on Various Datasets with $t = 0.5$

TGANet	CVC-300		CVC-ClinicDB		CVC-ColonDB		Kvasir	
t=0.5	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
Blood	10.0	10.0	9.2	11.3	12.9	15.3	6.5	7.8
Feces	1.7	8.3	1.8	2.5	5.8	7.4	1.1	1.4
Instrument	11.7	16.7	2.6	4.2	8.2	10.8	1.5	1.8
Spot	3.3	3.3	0.2	0.2	3.2	3.4	0.1	0.3
Saturation	8.3	10.0	9.3	15.2	10.8	15.0	22.4	30.4
Contrast	0.0	0.0	6.7	8.7	15.3	17.6	5.5	7.3
White Balance	23.3	26.7	30.2	37.7	21.3	40.8	25.4	32.6
Blur	15.0	21.7	4.7	6.5	6.6	7.1	5.8	7.5
Text	5.0	6.7	1.1	1.8	4.2	5.0	0.6	1.0

Table 5: EFR(%) of TGANet Model on Various Datasets with $t = 0.5$

Answer to RQ2: MedTest obtains up to 15.70% EFR when testing the SOTA academic models, which indicates that MedTest can effectively discover corner cases and used for further testing the robustness on other models.

4.4 RQ3: Enhancing Medical Image Diagnosis Performance Using MedTest-Generated Test Cases

Our research has substantiated that MedTest is adept at creating diagnostically consistent and realistic test cases, which are proficient in identifying errors in both commercial software and SOTA academic mod-

PraNet	CVC-300		CVC-ClinicDB		CVC-ColonDB		Kvasir	
t=0.25	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
Blood	3.3	6.7	0.5	1.0	4.0	5.0	0.5	0.6
Feces	0.0	1.7	0.8	2.0	7.4	9.2	0.5	1.5
Instrument	6.7	11.7	4.1	5.6	12.1	14.0	0.4	1.1
Spot	1.7	1.7	0.5	0.5	3.2	4.2	0.1	0.5
Saturation	8.3	13.3	1.6	3.4	6.6	8.4	5.8	9.6
Contrast	1.7	5.0	0.3	0.8	4.7	6.1	1.3	2.2
White Balance	8.3	13.3	12.7	18.0	19.8	22.7	7.5	12.3
Blur	8.3	8.3	9.6	13.6	14.2	17.2	14.2	18.8
Text	0.0	0.0	0.7	0.8	5.0	5.8	0.2	0.3

Table 6: EFR(%) of PraNet Model on Various Datasets with $t = 0.25$

SANet	CVC-300		CVC-ClinicDB		CVC-ColonDB		Kvasir	
t=0.25	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
Blood	0.0	0.0	0.0	0.0	2.9	3.4	0.1	0.1
Feces	1.7	1.7	0.3	0.5	6.9	7.4	0.1	0.2
Instrument	1.7	1.7	0.2	0.7	5.5	5.8	0.0	0.0
Spot	0.0	0.0	0.2	0.3	4.2	4.5	0.0	0.0
Saturation	5.0	6.7	1.0	1.8	3.4	5.5	1.7	3.1
Contrast	0.0	0.0	0.2	0.2	3.4	4.0	0.0	0.2
White Balance	0.0	0.0	3.4	6.2	10.8	14.0	5.4	9.2
Blur	3.3	5.0	0.3	0.5	6.3	8.7	1.1	2.0
Text	0.0	0.0	0.5	1.0	5.5	5.8	0.0	0.1

Table 7: EFR(%) of SANet Model on Various Datasets with $t = 0.25$

els. This leads to an imperative query: Can the test cases generated by MedTest be leveraged to augment the performance of medical image diagnosis systems? Essentially, the objective is to enhance the robustness of diagnostic models.

A logical approach to achieve this enhancement is through the retraining of models with test cases synthesized by MedTest, to assess if such retrained models exhibit increased resilience to a variety of perturbations. While we have already curated a dataset for this retraining purpose, the optimization of training outcomes is still an ongoing endeavor. Our future research efforts will focus on selecting images that trigger signifi-

SSFormer	CVC-300		CVC-ClinicDB		CVC-ColonDB		Kvasir	
t=0.25	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
Blood	3.3	3.3	0.2	0.2	5.0	5.3	0.1	0.1
Feces	0.0	0.0	0.3	0.5	7.6	8.2	0.0	0.0
Instrument	3.3	6.7	1.8	2.5	7.1	7.6	0.0	0.0
Spot	0.0	0.0	0.3	0.3	2.4	2.4	0.0	0.0
Saturation	6.7	10.0	1.0	1.3	2.6	4.5	0.4	0.8
Contrast	1.7	3.3	0.2	0.2	3.9	4.7	0.3	0.5
White Balance	3.3	5.0	4.7	7.5	11.8	13.9	2.0	4.0
Blur	0.0	1.7	0.2	0.2	3.4	3.4	0.3	0.4
Text	0.0	0.0	0.3	0.3	2.1	2.6	0.0	0.1

Table 8: EFR(%) of SSFormer Model on Various Datasets with $t = 0.25$

TGANet	CVC-300		CVC-ClinicDB		CVC-ColonDB		Kvasir	
t=0.25	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
Blood	16.7	20.0	15.8	22.1	23.9	29.2	12.9	15.7
Feces	13.3	25.0	4.4	7.0	13.9	18.2	2.7	3.7
Instrument	30.0	46.7	9.2	14.9	18.9	24.2	4.4	6.9
Spot	3.3	3.3	1.5	2.1	5.5	6.6	0.8	1.0
Saturation	16.7	18.3	21.2	28.9	21.8	24.7	46.1	53.7
Contrast	0.0	1.7	12.9	17.3	26.8	29.2	14.3	18.0
White Balance	31.7	38.3	47.5	59.5	35.3	40.8	43.0	49.8
Blur	28.3	31.7	4.7	6.5	9.7	11.8	15.3	18.3
Text	8.3	8.3	3.9	5.1	10.8	13.4	3.9	5.6

Table 9: EFR(%) of TGANet Model on Various Datasets with $t = 0.25$

cant errors when subjected to synthesized effects, and constructing a new training dataset. This dataset will be an amalgamation of original and synthesized diagnostically consistent images.

However, a critical consideration in this process is the judicious selection of perturbations for each seed image. Excessive repetition of segmentation label masks, targeting specific features, might inadvertently introduce biases in the retraining process. Consequently, careful curation of the training dataset is paramount to ensure a balanced representation of features and avoid skewed learning outcomes.

Table 10: Comparison of PraNet Model Outputs with Different Artifacts

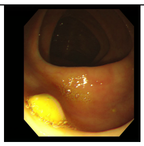
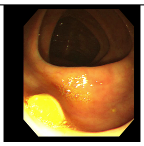
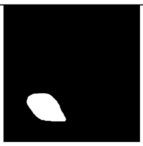
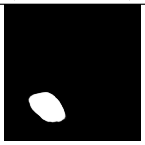
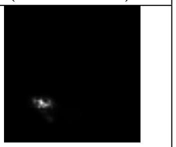
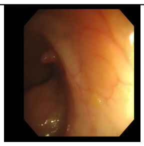
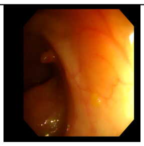
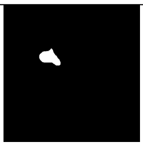
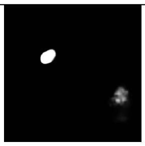
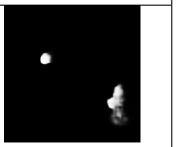
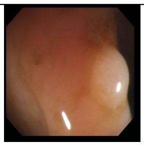
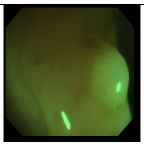
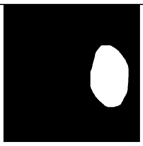
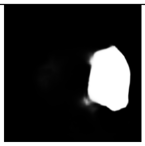
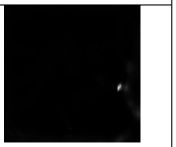
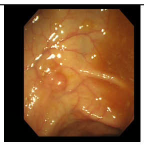
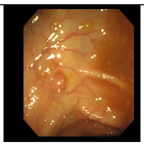
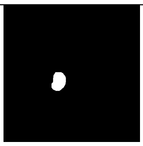
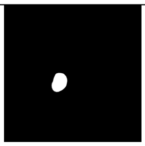
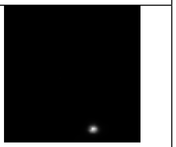
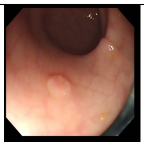
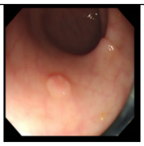
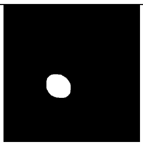
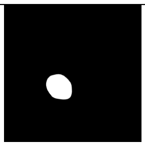
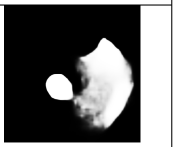
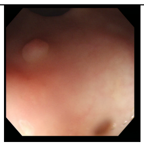
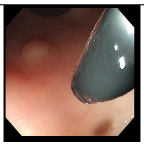
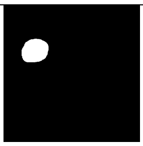
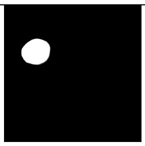
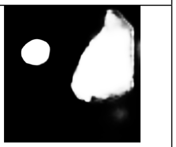
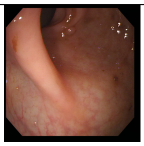
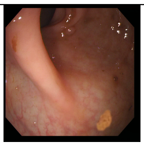
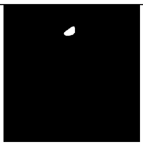
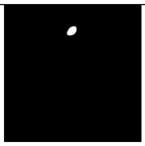
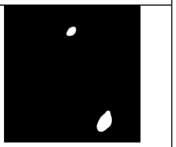
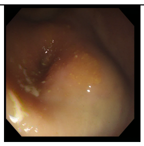
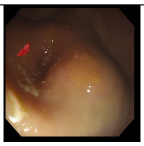
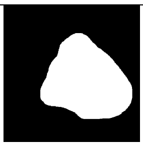
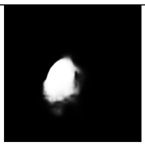
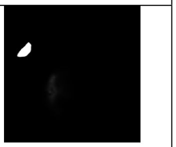
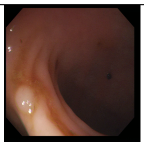
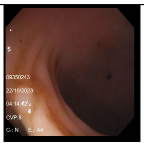
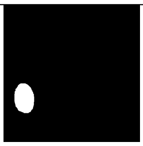
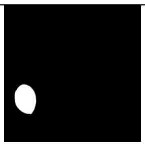
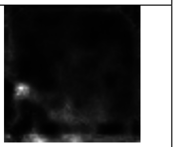
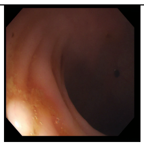
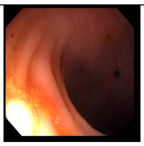
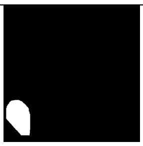
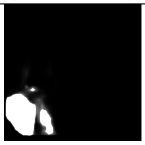
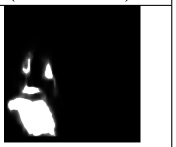
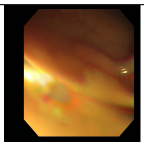
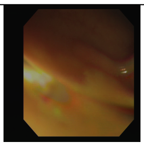
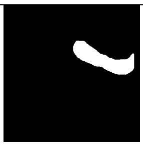
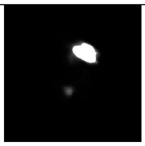
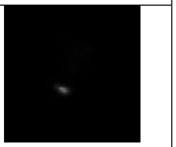
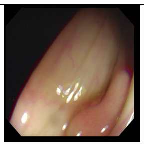
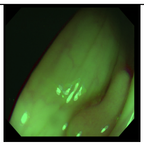
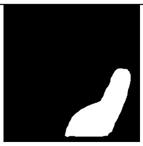
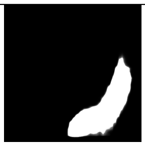
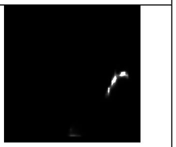
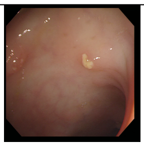
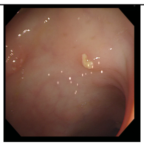
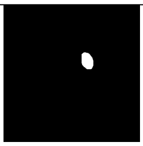
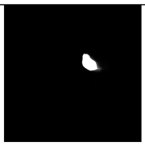
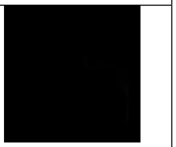
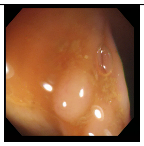
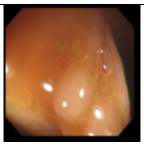
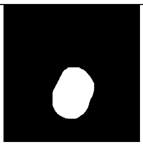
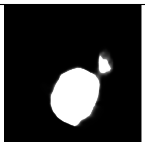
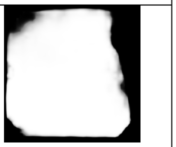
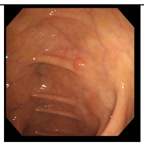
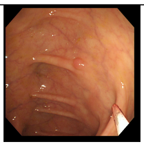
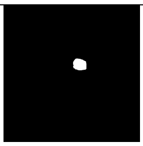
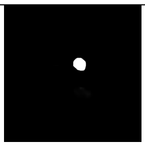
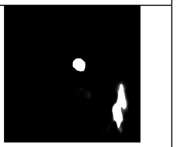
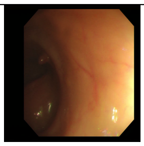
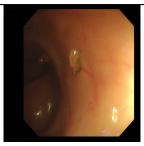
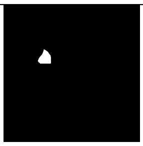
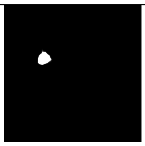
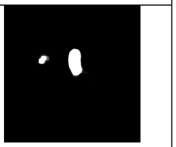
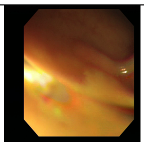
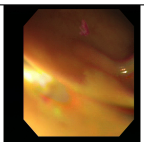
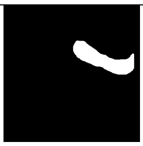
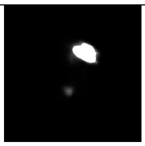
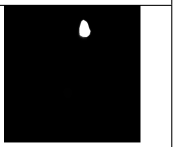
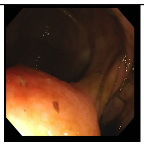
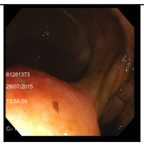
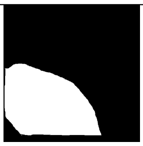
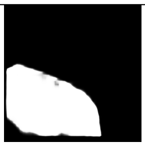
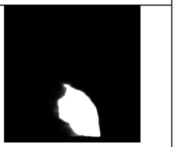
Artifact	Original Image	Image with Artifact	Ground Truth	Output (Original)	Output (Artifact)
Saturation					
Contrast					
White-Balance					
Specularity					
Blur					
Instrument					
Feces					
Blood					
Text					

Table 11: Comparison of SANet Model Outputs with Different Artifacts

Artifact	Original Image	Image with Artifact	Ground Truth	Output (Original)	Output (Artifact)
Saturation					
Contrast					
White-Balance					
Specularity					
Blur					
Instrument					
Feces					
Blood					
Text					

Answer to RQ3: We are still in the process of re-training the academic medical image diagnosis models to achieve a better performance.

4.5 RQ4: How would different factors affect the performance of MedTest?

This section delves into how three distinct external factors influence the efficacy of MedTest.

Image Structure and Overlay The heterogeneity in the source images' locations and orientations presents challenges in our automated object perturbation system. The lack of comprehensive image analysis during object addition precludes optimal object selection and placement, potentially leading to incongruous object positioning in the synthesized images. Efforts to mitigate this include excluding objects from atypically laid out images (such as instruments positioned at corners in partial views) and constraining the target positions for merging objects to more closely resemble their original context, albeit with slight positional variations. These measures aim to minimize the incongruities arising from layout and positioning discrepancies.

Polyp Characteristics The extensive diversity in the dataset, particularly regarding polyp size and shape, poses challenges for the automated synthesis of object-related perturbations. The presence of large polyps can complicate object addition, necessitating refined automation protocols for object selection and placement. This adjustment must accommodate the variation in polyp characteristics, striking a balance between overall performance and the generation of some suboptimal results.

Ambient Lighting Conditions Divergent lighting conditions in the seed images, especially those that are overly dark or bright, can lead to unnatural effects in object perturbations. To address this, we have implemented contrast, brightness, and color assessments for both the target and object-containing images. This enables the imposition of similarity constraints when selecting objects for synthesis, thereby mitigating unnatural outcomes. Lighting conditions also affect light-related perturbations such as saturation, contrast, and specularity. Excessively dark or bright seed images can render the application of saturation or contrast effects counterintuitive and unnatural. Specularity synthesis is similarly impacted in underexposed images.

5 Discussion

5.1 Threats to Validity

This section elucidates potential threats that could affect the validity of our study.

Variability in Diagnosis Ground Truth A primary concern is the potential alteration in the diagnostic accuracy of test cases generated by MedTest, especially after numerous perturbations, which could lead to false positives. To mitigate this risk, we engaged in expert annotation to affirm the diagnostic ground truth of these generated test cases. Additionally, annotators were instructed to assess whether the test cases authentically represent artifacts encountered in real-world clinical settings. The findings confirm that the artifacts generated by our methodology are diagnosis-neutral.

Scope of Application on Endoscope Image Analysis Another concern is the applicability of MedTest primarily to endoscope image analysis, which may not be universally extendable to other types of medical images. The selection of endoscope imagery was a deliberate decision, considering its representativeness in a specific medical imaging context. However, we posit that the MRs developed can be readily adapted to other medical imaging modalities. We provide a comprehensive framework encompassing the study of clinical artifacts, formulation and design of MRs, generation of test cases, and utilization of failure cases to enhance robustness.

Evaluation on a Limited Set of Medical Image Analysis Systems Our evaluation encompassed six medical image analysis systems, which may not comprehensively represent MedTest's efficacy across diverse systems. To address this, our evaluation targeted both commercial software employing Large Language Models (LLMs) and SOTA academic models pertinent to our focused task. Future endeavors will involve extending our testing to a broader array of commercial and research models to further validate and enhance the generalizability of MedTest's performance.

6 Related Work

6.1 Enhanced Testing Approaches for AI Software

AI software has revolutionized various fields with a wide array of applications, ranging from autonomous vehicle technology to sophisticated face recognition systems. However, a critical concern surrounding these AI-based models is their inherent lack of robustness. This vulnerability potentially leads to undesirable outputs, which can culminate in serious mishaps or accidents, as highlighted in several studies [79, 36]. In response to these challenges, researchers have diligently worked on developing a plethora of methods aimed at creating adversarial examples or test cases. These are specifically designed to deceive or 'fool' AI software, thus exposing their weaknesses [8, 47, 67, 77, 76, 65, 40, 46, 75, 51, 26, 25]. Concurrently, there has been a significant effort in proposing various algorithms and strategies to bolster the robustness of AI software. Notable among these are robust training mechanisms and advanced network debugging techniques [42, 4, 18, 70, 41, 61]. Our research is particularly focused on examining the robustness of a key AI application—medical image diagnosis software—which, until now, has not been systematically scrutinized in existing literature.

6.2 Comprehensive Analysis of Robustness in Medical Image Analysis Software

In our extensive survey of the literature, we have delved into the methodologies employed for testing and attacking medical image analysis systems, drawing insights from related domains such as natural language processing (NLP) and computer vision (CV). Over the years, a diverse array of metamorphic testing techniques has been proposed for NLP software, exploring novel approaches and methodologies [9, 10, 21, 23, 24, 50, 58]. Alongside metamorphic testing, the field has also seen significant advancements in identifying errors in NLP software, inspired by adversarial attack methodologies prevalent in the CV domain [20, 30, 32, 37, 74]. The realm of AI-driven CV software is a double-edged sword, offering both unprecedented convenience and potential risks in daily life. For instance, it has been observed that criminals can manipulate photos to deceive face recognition systems, and autopilot systems sometimes fail to detect imminent hazards. To address these concerns, several automated testing frameworks like DeepTest have been developed, aimed at rigorously testing the robustness of CV algorithms [62].

However, our research provides a unique and substantial contribution compared to the aforementioned studies. Firstly, MedTest, our proposed method, is specifically tailored for medical imaging, encompassing a comprehensive array of MRs suited for various perturbations encountered in real clinical settings. To the best of our knowledge, the MRs proposed in MedTest are novel and unexplored in existing literature across these research areas. Furthermore, unlike previous studies that focus on single tasks, MedTest is implemented for multiple tasks including segmentation, classification, and visual-question answering. Importantly, all the MRs in our study are grounded in real-world clinical scenarios, as evidenced by our preliminary studies, marking a departure from previous research which often conceptualized perturbations without empirical validation. Moreover, while most existing studies evaluate their methodologies on research models, MedTest extends its evaluation to include two leading commercial software products. This comprehensive approach positions MedTest as a pioneering and holistic testing framework for medical image analysis systems.

7 Future Work

7.1 Testing on Multimodal Models on Medical Visual Question Answering

Amid the fast development of large language models and the derived multimodal models, it has been seen that multimodal models possess sufficient ability to take in specific image input and generate text output based on its knowledge on the given image, which can also be utilized in medical diagnosis. Visual question answering on medical context is therefore a crucial application on such large-scale multimodal models. Based on this, we plan to further investigate the most popular and advent

Large Vision-Language Models (LVLM), including the newly released GPT-4 with Vision (GPT-4V) and Bard, which demostrated exceptional performance in tackling multimodal tasks. This particular ability also gives rise to the potential expansion of LVLMs on the realm of VQA.

In the desired experiment setting, we plan to leverage the proposed MRs and generate the test cases based on seed dataset specialized for VQA tasks. Then, we may utilized these generated images as input to direct the LVLMs to answer the given questions regarding the input image.

Questions to be answered by our tested multimodal models should be ones that only require text output, i.e., can be illustrated by language. Sample questions may includes, ”How many polyps are in the image?”, or, ”Are there any instruments in the image?”. In this case, the multimodal models can provide straightforward answers, such as ”Yes/No” or the number, which also benefits our decision on evaluation criteria.

7.2 Re-training Models for Performance Improvement

A question yet to answer is whether test cases generated by MedTest can applied to improve the performance of academic SOTA medical image diagnosis models. As illustrated in our RQ2, evident errors can be triggered by our generated dataset, which may imply that the models lack sufficient knowledge on these corner cases, leading to inappropriate representations of vital features within the images for diagnosis. Therefore, one intuitive approach is to construct a more comprehensive dataset for training, especially including those synthesized images with specific artifact types associated with relatively unsatisfactory performance. In this way, we hope to cover more corner cases in the model outputs and expand the activated neurons within the models.

7.3 Image Synthesis with Generative Adversarial Networks

Since our MedTest in producing MRs mainly involves mathematical-representation-based image transformation and processing techniques, limitations exist on producing large-scale dataset with more variations within each artifact class, especially object perturbations. Due to the scarce data to serve as artifact candidates, our generated samples have specific artifact patterns, which restrict the images from being more natural. Besides, we cannot simulate the possible large-area presence of artifacts and potential existence on vital areas, such as on polyps, without affecting the original ground truth label. For instance, blood may appear in a contiguous and pervasive manner, but our simulation method only extract small parts from it and cannot produce the same effect as original.

Therefore, generative adversarial networks (GAN) may exhibit its potential in creating a more realistic blending different elements into medical images as we desired. After we surveyed the related work, we have asserted that GANs have exceptional power in generating natural fusion of image contents and styles according to the given images and segmentation label of different instance categories. [14, 19] Similar application in medical images, even in polyp related tasks, have been witnessed with promising performance. Because of this, we plan to explore deeper into this topic and try to generate more realistic images regarding the object perturbations for our customized dataset, so that we can further improve the overall evaluation on our target models.

8 Conclusion

In this report, we embarked on an in-depth analysis of AI-driven diagnostic tools in medical imaging, with a particular emphasis on endoscopic image diagnosis. The choice to focus initially on this area stems from its critical importance in healthcare. Accurate and reliable medical diagnostics are fundamental to patient care, and the increasing integration of AI tools in this domain necessitates a rigorous evaluation of their performance. Our development of MedTest, a specialized metamorphic testing framework, marks a significant step in this direction, enabling a detailed assessment of these tools under various clinically relevant scenarios.

Through our comprehensive pilot study, we identified and categorized common artifacts that pose challenges to the diagnostic accuracy of these tools. We generated the 9 different types of artifacts on 4 datasets, involving more than $2,000$ images and generated over $18,000$ images with artifacts. Our findings reveal that even SOTA algorithms exhibit varying degrees of performance degradation when faced with these realistic test cases, underscoring the need for continual improvement and rigorous testing.

While this study provides valuable insights into the robustness of medical image diagnosis software, it also sets the stage for our next ambitious endeavor: evaluating the performance of multimodal models. Multimodal models, which integrate and interpret data from various modalities, are poised to revolutionize medical diagnostics by offering a more comprehensive analysis than single-modality models. However, the complexity of these models necessitates a nuanced approach to testing and validation.

To this end, our future work will focus on extending the methodologies and lessons learned from our current research to the realm of multimodal models. This includes developing testing frameworks that can effectively assess the performance of these models in integrating and analyzing data from diverse sources. Our ultimate goal is to ensure that as these advanced AI tools become integral to medical diagnostics, they do so with the highest standards of accuracy and reliability, thus enhancing patient outcomes and advancing healthcare services.

In conclusion, this report not only sheds light on the vulnerabilities of current medical image diagnosis software but also lays the groundwork for future explorations into the broader domain of AI-driven diagnostic tools, including multimodal models. As we continue to push the boundaries of AI in healthcare, rigorous testing and continual improvement of these tools will be paramount to fully realizing their potential in improving patient care.

References

[1] Ali, S., Zhou, F., Bailey, A., Braden, B., East, J., Lu, X., and Rittscher, J. A deep learning framework for quality assessment and restoration in video endoscopy, 2019.

[2] Ali, S., Zhou, F., Daul, C., Braden, B., Bailey, A., Realdon, S., East, J., Wagnières, G., Loschenov, V., Grisan, E., Blondel, W., and Rittscher, J. Endoscopy artifact detection (ead 2019) challenge dataset, 2019.

[3] Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., and Asari, V. K. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *ArXiv abs/1802.06955* (2018).

[4] Asyrofi, M. H., Yang, Z., Shi, J., Quan, C. W., and Lo, D. Can differential testing improve automatic speech recognition systems? *2021 IEEE International Conference on Software Maintenance and Evolution (ICSME)* (2021), 674–678.

[5] Bengio, Y., et al. Learning deep architectures for ai. *Foundations and trends® in Machine Learning 2*, 1 (2009), 1–127.

[6] Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., and Vilariño, F. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics 43* (2015), 99–111.

[7] Brody, H. Medical imaging. *Nature 502*, 7473 (2013), S81–S81.

[8] Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M. E., Shields, C., Wagner, D. A., and Zhou, W. Hidden voice commands. In *USENIX Security Symposium* (2016).

[9] Chen, S., Jin, S., and Xie, X. Testing your question answering software via asking recursively. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)* (2021), IEEE, pp. 104–116.

[10] Chen, S., Jin, S., and Xie, X. Validation on machine reading comprehension software without annotated labels: A property-based method. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (2021), pp. 590–602.

[11] Chen, T. Y., Cheung, S. C., and Yiu, S.-M. Metamorphic testing: A new approach for generating next test cases. *ArXiv abs/2002.12543* (2020).

[12] Chen, T. Y., Ho, J. W. K., Liu, H., and Xie, X. An innovative approach for testing bioinformatics programs using metamorphic testing. *BMC Bioinformatics 10* (2008), 24 – 24.

[13] Chen, Y., Juttukonda, M., Su, Y., Benzinger, T., Rubin, B. G., Lee, Y. Z., Lin, W., Shen, D., Lalush, D., and An, H. Probabilistic air segmentation and sparse regression estimated pseudo ct for pet/mr attenuation correction. *Radiology 275*, 2 (2015), 562–569.

[14] Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 8789–8797.

[15] Chornyy, R. Artificial intelligence in healthcare: Market size, growth, and trends. `https://binariks.com/blog/artificial-intelligence-ai-healthcare-market/`, 2023. Accessed: 2023-11-01.

[16] Dwarakanath, A., Ahuja, M., Sikand, S., Rao, R. M., Bose, R. P. J. C., Dubash, N., and Podder, S. Identifying implementation bugs in machine learning based image classifiers using metamorphic testing. *Proceedings of the 27th ACM SIGSOFT International Symposium on Software Testing and Analysis* (2018).

[17] Fan, D.-P., Ji, G.-P., Zhou, T., Chen, G., Fu, H., Shen, J., and Shao, L. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention* (2020), Springer, pp. 263–273.

[18] Gao, X., Saha, R. K., Prasad, M. R., and Roychoudhury, A. Fuzz testing based data augmentation to improve robustness of deep neural networks. *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)* (2020), 1147–1158.

[19] Günther, E., Gong, R., and Van Gool, L. Style adaptive semantic image editing with transformers. In *European Conference on Computer Vision* (2022), Springer, pp. 187–203.

[20] Guo, J., Zhang, Z., Zhang, L., Xu, L., Chen, B., Chen, E., and Luo, W. Towards variable-length textual adversarial attacks. *arXiv preprint arXiv:2104.08139* (2021).

[21] Gupta, S., He, P., Meister, C., and Su, Z. Machine translation testing via pathological invariance. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (2020), pp. 863–875.

[22] Hall, K. K., Shoemaker-Hunt, S. J., Hoffman, L., Richard, S., Gall, E. M., Schoyer, E., Costar, D., Gale, B., Schiff, G. D., Miller, K., Earl, T. R., Katapodis, N. D., Sheedy, C. K., Wyant, B. E., Bacon, O., Hassol, A., Schneiderman, S. R., Woo, M., Leroy, L., Fitall, E., Long, A.-M., Holmes, A., Riggs, J. S., and Lim, A. Making healthcare safer iii: A critical analysis of existing and emerging patient safety practices [internet].

[23] He, P., Meister, C., and Su, Z. Structure-invariant testing for machine translation. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering* (2020), pp. 961–973.

[24] He, P., Meister, C., and Su, Z. Testing machine translation via referential transparency. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)* (2021), IEEE, pp. 410–422.

[25] Huang, W., Sun, Y., Zhao, X.-E., Sharp, J., Ruan, W., Meng, J., and Huang, X. Coverage-guided testing for recurrent neural networks. *IEEE Transactions on Reliability* (2021).

[26] Humbatova, N., Jahangirova, G., and Tonella, P. Deepcrime: mutation testing of deep learning systems based on real faults. *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis* (2021).

[27] Jha, D., Ali, S., Emanuelsen, K., Hicks, S. A., VajiraThambawita, Garcia-Ceja, E., Riegler, M. A., de Lange, T., Schmidt, P. T., Johansen, H. D., Johansen, D., and Halvorsen, P. Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy, 2020.

[28] Jha, D., Riegler, M., Johansen, D., Halvorsen, P., and Johansen, H. D. Doubleu-net: A deep convolutional neural network for medical image segmentation. *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)* (2020), 558–564.

[29] Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., de Lange, T., Johansen, D., and Johansen, H. D. Kvasir-seg: A segmented polyp dataset, 2019.

[30] Jia, R., Raghunathan, A., Göksel, K., and Liang, P. Certified robustness to adversarial word substitutions. *arXiv preprint arXiv:1909.00986* (2019).

[31] Jie, Z., Zhiying, Z., and Li, L. A meta-analysis of watson for oncology in clinical application. *Scientific Reports 11* (2021).

[32] Jin, D., Jin, Z., Zhou, J. T., and Szolovits, P. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence* (2020), vol. 34, pp. 8018–8025.

[33] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. Segment anything, 2023.

[34] Kirk, H. R., Vidgen, B., Röttger, P., Thrush, T., and Hale, S. A. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. *ACL abs/2108.05921* (2021).

[35] LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature 521*, 7553 (2015), 436–444.

[36] Levin, S. Tesla fatal crash: 'autopilot' mode sped up car before driver killed, report finds [online]. `https://www.theguardian.com/technology/2018/jun/07/tesla-fatal-crash-silicon-valley-autopilot-mode-report`, 2018. Accessed: 2018-06.

[37] Li, D., Zhang, Y., Peng, H., Chen, L., Brockett, C., Sun, M.-T., and Dolan, B. Contextualized perturbation for textual adversarial attack. *arXiv preprint arXiv:2009.07502* (2020).

[38] Lin, Y., Wu, J., Xiao, G., Guo, J., Chen, G., and Ma, J. Bsca-net: Bit slicing context attention network for polyp segmentation. *Pattern Recognition 132* (2022), 108917.

[39] Litjens, G. J. S., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J., van Ginneken, B., and Sánchez, C. I. A survey on deep learning in medical image analysis. *Medical image analysis 42* (2017).

[40] Luo, Y., Meghjani, M., Ho, Q. H., Hsu, D., and Rus, D. Interactive planning for autonomous urban driving in adversarial scenarios. *2021 IEEE International Conference on Robotics and Automation (ICRA)* (2021), 5261–5267.

[41] Ma, S., Liu, Y., Lee, W.-C., Zhang, X., and Grama, A. Y. Mode: automated neural network model debugging via state differential analysis and input selection. *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (2018).

[42] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *ICLR abs/1706.06083* (2018).

[43] Makary, M. A. Medical error—the third leading cause of death in the us. `https://www.bmj.com/content/353/bmj.i2139`, 2016. Accessed: 2023-11-01.

[44] Merative. Ibm watson health's product has been equipped in nine of the top 10 us hospitals and decreased 32% in ed visits by high utilizers. `https://www.merative.com/`, 2023. Accessed: 2023-11-01.

[45] Newman-Toker, D. E., Nassery, N., Schaffer, A., Yu-Moe, C. W., Clemens, G. D., Wang, Z., Zhu, Y., Tehrani, A. S. S., Fanai, M., Hassoon, A., and Siegal, D. Burden of serious harms from diagnostic error in the usa. *BMJ Quality & Safety* (2023).

[46] Pei, K., Cao, Y., Yang, J., and Jana, S. S. Deepxplore: Automated whitebox testing of deep learning systems. *Proceedings of the 26th Symposium on Operating Systems Principles* (2017).

[47] Pham, H. V., Kim, M., Tan, L., Yu, Y., and Nagappan, N. Deviate: A deep learning variance testing framework. *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)* (2021), 1286–1290.

[48] Rajan, R. E., and Koonambaikulathamma, V. Spatial and hierarchical feature extraction based on sift for medical images.

[49] Reid, D. Google's deepmind a.i. beats doctors in breast cancer screening trial. `https://www.cnbc.com/2020/01/02/googles-deepmind-ai-beats-doctors-in-breast-cancer-screening-trial.html`, 2020. Accessed: 2023-11-01.

[50] Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), pp. 4902–4912.

[51] Riccio, V., Jahangirova, G., Stocco, A., Humbatova, N., Weiss, M., and Tonella, P. Testing machine learning based systems: a systematic mapping. *Empir. Softw. Eng. 25* (2020), 5193–5254.

[52] S., A., F., Z., and et al., B. B. An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. *Scientific Reports 10* (2020), 2748.

[53] Schmidhuber, J. Deep learning in neural networks: An overview. *Neural networks 61* (2015), 85–117.

[54] Segura, S., Fraser, G., Sánchez, A. B., and Cortés, A. R. A survey on metamorphic testing. *IEEE Transactions on Software Engineering 42* (2016), 805–824.

[55] Shao, Y., Gao, Y., Guo, Y., Shi, Y., Yang, X., and Shen, D. Hierarchical lung field segmentation with joint shape and appearance sparse learning. *IEEE transactions on medical imaging 33*, 9 (2014), 1761–1780.

[56] Shen, Z., Fu, H., Shen, J., and Shao, L. Modeling and enhancing low-quality retinal fundus images. *IEEE transactions on medical imaging 40*, 3 (2020), 996–1006.

[57] Suk, H.-I., Lee, S.-W., Shen, D., and Initiative, A. D. N. Deep sparse multi-task learning for feature selection in alzheimer's disease diagnosis. *Brain Structure and Function 221* (2016), 2569–2587.

[58] Sun, Z., Zhang, J. M., Harman, M., Papadakis, M., and Zhang, L. Automatic testing and improvement of machine translation. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering* (2020), pp. 974–985.

[59] Tajbakhsh, N., Gurudu, S. R., and Liang, J. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging 35*, 2 (2015), 630–644.

[60] Tang, C., Dong, Y., and Su, X. Automatic registration based on improved sift for medical microscopic sequence images. *2008 Second International Symposium on Intelligent Information Technology Application 1* (2008), 580–583.

[61] Tao, G., Ma, S., Liu, Y., Xu, Q., and Zhang, X. Trader: Trace divergence analysis and embedding regulation for debugging recurrent neural networks. *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)* (2020), 986–998.

[62] Tian, Y., Pei, K., Jana, S., and Ray, B. Deeptest: Automated testing of deep-neural-network-driven autonomous cars. In *Proceedings of the 40th international conference on software engineering* (2018), pp. 303–314.

[63] Tian, Y., Pei, K., Jana, S. S., and Ray, B. Deeptest: Automated testing of deep-neural-network-driven autonomous cars. *2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)* (2017), 303–314.

[64] Tomar, N. K., Jha, D., Bagci, U., and Ali, S. Tganet: Text-guided attention for improved polyp segmentation, 2022.

[65] Tu, J., Li, H., Yan, X., Ren, M., Chen, Y., Liang, M., Bitar, E., Yumer, E., and Urtasun, R. Exploring adversarial robustness of multi-sensor perception systems in self driving. *ArXiv abs/2101.06784* (2021).

[66] Vázquez, D., Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., López, A. M., Romero, A., Drozdzal, M., and Courville, A. A benchmark for endoluminal scene segmentation of colonoscopy images, 2016.

[67] Wang, J., Chen, J., Sun, Y., Ma, X., Wang, D., Sun, J., and Cheng, P. Robot: Robustness-oriented testing for deep learning systems. *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)* (2021), 300–311.

[68] Wang, J., Huang, Q., Tang, F., Meng, J., Su, J., and Song, S. Stepwise feature fusion: Local guides global, 2022.

[69] Wang, L., Chen, K. C., Gao, Y., Shi, F., Liao, S., Li, G., Shen, S. G., Yan, J., Lee, P. K., Chow, B., et al. Automated bone segmentation from dental cbct images using patch-based sparse representation and convex optimization. *Medical physics 41*, 4 (2014), 043503.

[70] Wang, W., Huang, J., Chen, C., Gu, J., Zhang, J., Wu, W., He, P., and Lyu, M. R. Validating multimedia content moderation software via semantic fusion. *ArXiv abs/2305.13623* (2023).

[71] Wang, W., tse Huang, J., Wu, W., Zhang, J., Huang, Y., Li, S., He, P., and Lyu, M. R. Mttm: Metamorphic testing for textual content moderation software. *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)* (2023), 2387–2399.

[72] Wei, J., Hu, Y., Zhang, R., Li, Z., Zhou, S. K., and Cui, S. Shallow attention network for polyp segmentation, 2021.

[73] Xie, X., Ho, J. W. K., Murphy, C., Kaiser, G. E., Xu, B., and Chen, T. Y. Testing and validating machine learning classifiers by metamorphic testing. *The Journal of systems and software* (2011).

[74] Zang, Y., Qi, F., Yang, C., Liu, Z., Zhang, M., Liu, Q., and Sun, M. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), pp. 6066–6080.

[75] Zhang, J., Harman, M., Ma, L., and Liu, Y. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering 48* (2022), 1–36.

[76] Zhang, J., tse Huang, J., Wang, W., Li, Y., Wu, W., Wang, X., Su, Y., and Lyu, M. R. Improving the transferability of adversarial samples by path-augmented method. *ArXiv abs/2303.15735* (2023).

[77] Zhang, J., Wu, W., tse Huang, J., Huang, Y., Wang, W., Su, Y., and Lyu, M. R. Improving adversarial transferability via neuron attribution-based attacks. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 14973–14982.

[78] Zhang, M., Zhang, Y., Zhang, L., Liu, C., and Khurshid, S. Deeproad: Gan-based metamorphic testing and input validation framework for autonomous driving systems. *2018 33rd IEEE/ACM International Conference on Automated Software Engineering (ASE)* (2018).

[79] Ziegler, C. A google self-driving car caused a crash for the first time. [online]. `https://www.theverge.com/2016/2/29/11134344/google-self-driving-car-crash-report`, 2016. Accessed: 2016-09.

[80] Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M., Kaus, M. R., Haker, S. J., Wells, W. M., Jolesz, F. A., and Kikinis, R. Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. *Academic Radiology 11*, 2 (2004), 178–189.