UC Irvine
UC Irvine Previously Published Works

Title
Real-world readiness: Evaluating AI model performance in degraded photographic imaging for cancer detection

Permalink
https://escholarship.org/uc/item/9qj8h8nv

Journal
Expert Systems with Applications, 303

ISSN
0957-4174

Authors
You, Ruilin
Chen, Jiabin
Wang, Yihan
et al.

Publication Date
2026-03-01

DOI
10.1016/j.eswa.2025.130569

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Expert Systems With Applications 303 (2026) 130569



Contents lists available at ScienceDirect

Expert Systems With Applications

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





Real-world readiness: Evaluating AI model performance in degraded photographic imaging for cancer detection

Ruilin You [a], Jiabin Chen [a], Yihan Wang [a], Zhihan Hong [a], Cherie Wink [b], Petra Wilder-Smith [b], Rongguang Liang [a], Bofan Song [a,*]

[a] Wyant College of Optical Sciences, University of Arizona, Tucson, AZ, USA
[b] Beckman Laser Institute & Medical Clinic, University of California Irvine, Irvine, CA, USA

ARTICLE INFO

Keywords:
Artificial intelligence robustness
Image degradation
Cancer detection

ABSTRACT

Photographic imaging is a vital tool for cancer detection and screening, enabling non-invasive visualization of abnormal tissues and early-stage malignancies. It is particularly effective for detecting and screening cancers that are visually accessible, such as skin, oral, cervical, and certain gastrointestinal cancers. Techniques such as dermoscopy, intraoral imaging, and endoscopy leverage high-resolution photographic methods to capture detailed views of body tissues. A key advantage of photographic imaging lies in its simplicity and the feasibility of constructing compact, low-cost systems, making it especially well-suited for large-scale cancer screening programs. However, this streamlined approach can lead to variability in image quality due to factors such as lighting conditions, focus, and operator handling. These inconsistencies pose challenges for AI-based diagnostic systems, where AI model performance can be highly sensitive to input variations. Consequently, ensuring the robustness of AI algorithms across diverse and degraded image conditions is essential for real-world deployment. In this study, we comprehensively evaluate the performance and robustness of state-of-the-art convolutional neural networks (CNNs), transformers, and general-purpose foundation models under common system-induced colorimetric/tonal and capture-induced degradations, and when degradation type combines. These degradations were carefully designed to simulate realistic image quality issues frequently encountered in clinical and screening settings of photographic imaging. Our findings provide a solid foundation and actionable insights for developing reliable deep learning classification systems in practical medical environments. We identified how training strategies, neural network architectures, and image degradation types individually and jointly influence AI-based diagnostic performance in photographic cancer detection. Furthermore, by accurately assessing how variations in photographic imaging devices and capture-induced image-quality changes affect AI-integrated diagnostic performance, this work highlights critical considerations for building future AI-integrated imaging systems that are resilient to real-world operational challenges.

1. Introduction

Cancer continues to be a major global health challenge, responsible for millions of deaths each year and placing a substantial burden on healthcare systems (Bray et al., 2024). Early and accurate detection plays a pivotal role in improving treatment outcomes and survival rates, particularly for cancers that are often asymptomatic in their early stages (Crosby et al., 2022; Fitzgerald et al., 2022). Traditional diagnostic methods, such as biopsies and laboratory tests, are often invasive, time-consuming, and resource intensive. Non-invasive diagnostic tools that can support timely screening and diagnosis are thus of significant interest, especially in resource-constrained or high-risk settings (Chacko & Ankri, 2024; Narayanamurthy et al., 2018).

Photographic imaging has gained prominence as a practical and non-invasive tool for cancer detection due to its ability to capture visual characteristics of lesions and tissue abnormalities (Cho et al., 2020; Warin et al., 2021). High-resolution intraoral photographs allow clinicians to screen for early signs of oral potentially malignant disorders (OPMDs) and carcinomas, particularly in high-risk populations (Lin et al., 2021; Mirfendereski et al., 2025; Song et al., 2021). Dermoscopic and standard photographic images are widely used to identify cutaneous lesions that may indicate melanoma or non-melanoma skin cancers

* Corresponding author.
E-mail addresses: ruilinyou@arizona.edu (R. You), jiabinchen@arizona.edu (J. Chen), wangyihan@arizona.edu (Y. Wang), zhihanhong@arizona.edu (Z. Hong), cwink@hs.uci.edu (C. Wink), pwsmith@hs.uci.edu (P. Wilder-Smith), rliang@optics.arizona.edu (R. Liang), songb@arizona.edu (B. Song).

https://doi.org/10.1016/j.eswa.2025.130569
Received 8 August 2025; Received in revised form 28 October 2025; Accepted 22 November 2025
Available online 1 December 2025
0957-4174/© 2025 Published by Elsevier Ltd.

(Marka et al., 2019; Thomas & Puig, 2017). These imaging modalities are not only non-invasive and relatively inexpensive, but also scalable and suitable for telemedicine applications, making them especially valuable in point-of-care, low-resource and remote settings. Furthermore, the image-based nature of these modalities lends itself well to computational analysis, enabling the development of automated diagnostic systems that can support or augment clinical expertise (Fu et al., 2020; Song et al., 2018).

Artificial Intelligence (AI), has revolutionized image-based medical diagnostics (Chan et al., 2020; Panayides et al., 2020). Convolutional Neural Networks (CNNs) have demonstrated strong performance in visual pattern recognition tasks, including cancer classification from photographic images (Anwar et al., 2018). More recently, transformer-based vision models, have introduced new architectures that can capture global and long-range dependencies in image data (Azad et al., 2024). These models have shown promising results across a range of computer vision benchmarks and are increasingly being explored for medical imaging tasks (Shamshad et al., 2023). However, a major limitation of these approaches lies in their sensitivity to image quality (Dodge & Karam, 2016). Real-world medical images are often affected by various types of degradations, such as motion blur, improper lighting, low contrast, and optical glare, arising from factors like patient movement, inconsistent imaging conditions, and limitations of portable imaging devices (Lepcha et al., 2023; Pei et al., 2019). These artifacts can significantly reduce model performance and undermine clinical trust in AI-based systems, especially when deployed outside controlled environments (Herath et al., 2025; Rajpurkar & Lungren, 2023).

In this study, we systematically investigate the robustness of representative deep learning models under various types of image degradation commonly encountered in both clinical and field settings. We present a comprehensive comparison of symbolic CNNs, transformers, and foundation models applied to two datasets: intraoral images acquired using our custom-built imaging device (Birur N et al., 2022; Uthoff et al., 2019, 2018b), and a publicly available skin cancer photographic image dataset (Wen et al., 2022). The image degradations we simulate closely mimic real-world variations in imaging quality, including device inconsistencies, capture-induced artifact, and optical distortions. We evaluated eight distinct types of image degradation across various levels of severity to comprehensively assess the robustness of different deep learning architectures. Capture induced and optical distortion degradations, including glare, resolution reduction, motion blur, and defocus, were evaluated across 11 severity levels ranging from 0 to 10. Additionally, colorimetric and tonal degradations, specifically contrast, brightness, saturation, and hue, were evaluated across 21 severity levels ranging from negative 10 to positive 10. Each type and level of degradation was evaluated independently using fivefold cross validation for both the intraoral and skin image datasets. Additionally, we introduced dual-mixed degradations to form a two-dimensional evaluation matrix, enabling the study of combined effects when two types of degradation occur simultaneously. This is crucial, as in real-world image acquisition, multiple degradations often co-occur in non-linear and non-additive ways. Exploring these compound scenarios provides valuable insight for designing more robust imaging systems and improving the reliability of AI-assisted diagnostics.We also investigated how different augmentation strategies used during the training phase affect model performance. Our evaluation spans multiple model families and degradation types, enabling a comprehensive comparison of how different architectures handle suboptimal input quality. By benchmarking model performance across varying levels of distortion, we aim to understand the reliability of these models in real-world scenarios, where ideal imaging conditions are not always guaranteed. Overall, this work contributes by uniquely analyzing (i) the effects of distinct and combined image quality degradations that mimic real-world clinical screening, (ii) the impact of state-of-art deep learning architectures including CNNs, attention based Transformers, and foundation models, and (iii) training strategies and novel evaluation protocols for assessing diagnostic performance and robustness on photographic oral and skin medical images.

The significance of this foundational study is multi-faceted and has profound implications for both the development of AI in medicine and the design of future diagnostic tools. This study provides a comprehensive empirical assessment of how different AI architectures, CNNs, Transformer-based models, and multimodal foundation models (Zhang & Metaxas, 2024) (CLIP and SigLIP vision models) (Radford et al., 2021; Zhai et al., 2023) respond to various forms of image quality degradation in the context of photographic-based cancer detection. This nuanced understanding is essential for the development and deployment of reliable and robust AI systems in clinical settings, ensuring their diagnostic performance remains resilient under real-world imaging conditions. Another practical significance, the findings offer actionable insights for the optimization and future design of AI-integrated medical imaging devices and their operational protocols. By systematically identifying which specific types of degradation individually and jointly impact model performance, stakeholders can strategically focus their efforts on mitigating the most critical image quality issues. This targeted approach will support the development of next-generation AI-integrated imaging systems that are better aligned with the requirements of AI tools, ultimately improving the accuracy, reliability, and clinical utility of cancer detection workflows.

2. Related work

Robustness evaluation has become a critical area of research in medical imaging, especially as deep learning models transition from controlled environments to real-world clinical applications. Prior studies have explored the impact of various noise sources, resolution loss, and artifact corruptions on model performance, particularly in radiology modalities such as CT, X-ray, and MRI (Litjens et al., 2017; Lundervold & Lundervold, 2019). As deep learning becomes an increasingly accessible and reliable tool for assisting clinical image classification, several benchmarking efforts have been conducted to evaluate model performance on corrupted medical images, particularly in MRI data (Boone et al., 2023). However, unlike standard MRI and CT data, photographic medical imaging involves a wider variety of acquisition procedures and more uncontrollable environments–such as lighting variations, differences in imaging devices, patient movement, uneven surfaces, image cropping, and other factors encountered during the workflow. These challenges are particularly evident in fields like dermatology, endoscopy, and intraoral imaging, each with its own unique imaging characteristics and difficulties. Studies have examined the robustness of convolutional neural networks (CNNs) under various types of image degradation, using datasets such as the public available ISIC skin lesion dataset. These evaluations often involve adding noise and other commonly encountered artificial artifacts to assess and benchmark the brittleness of neural networks under non-ideal imaging conditions (Maron et al., 2021a,b). They provided valuable insights into training strategies and neural network architectures for various binary classification tasks. Recent work in endoscopic medical imaging has also focused on evaluating and developing training strategies, as well as exploring network architectures through robustness analysis (Jaspers et al., 2024). In the work by Kusters et al. (2025), the authors presented a state-of-the-art comparison between transformer-based architectures and conventional CNN-based classifiers for gastrointestinal imaging. There are many intraoral devices for OCT probes or CT scanning used in dental assessments. Photographic imaging of oral lesions is extremely important for identifying oral cancers. Previous studies have evaluated the performance of conventional neural networks in classifying cancerous oral lesions (Tanriver et al., 2021; Welikala et al., 2020). There has also been work focused on distinguishing between benign and malignant oral cancers using fine-tuned transfer learning strategies (Islam et al., 2023). While such previous studies have provided valuable insights into how deep learning can assist in training classification models, they have not tackled how image quality impacts

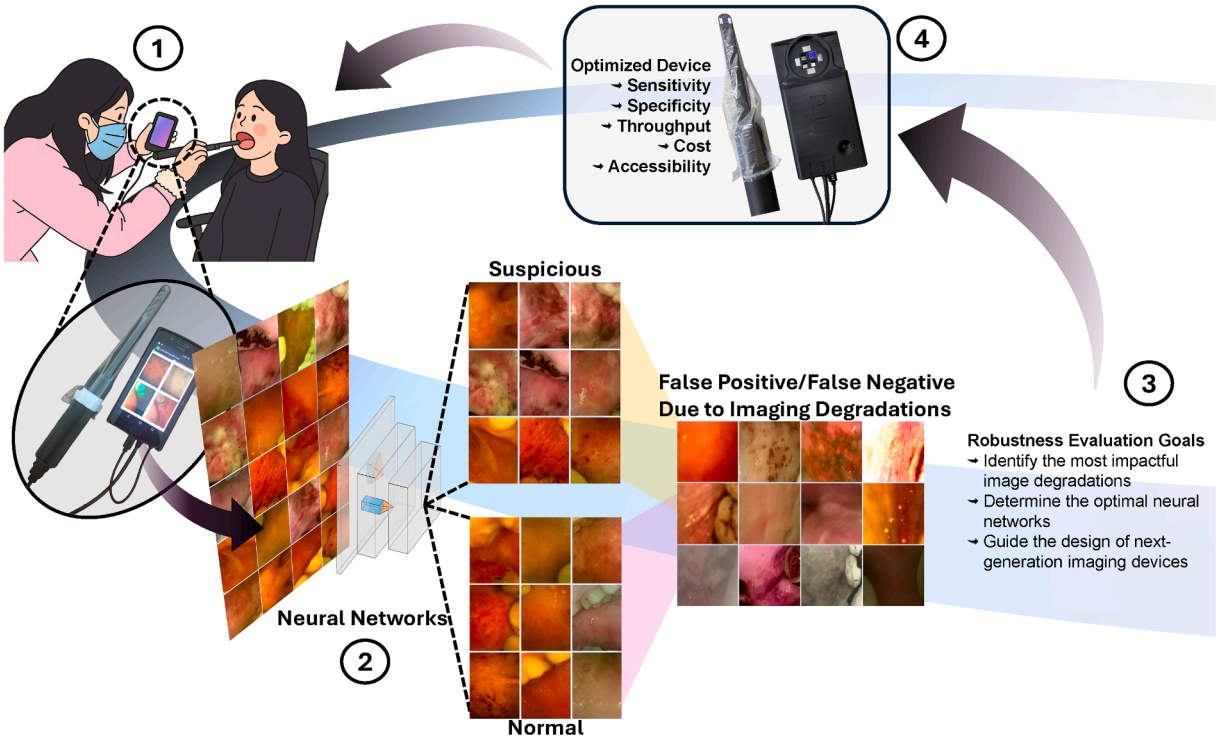


Fig. 1. Cycle of the next-generation, deep learning-assisted oral photographic device development. Step 1: Large-scale image data is acquired using our custom-designed and in-house-built mobile oral imaging device. Step 2: Clinical experts annotate, categorize, and verify the images, organizing them into normal and cancerous cases for classifier training. Step 3: Comprehensive robustness evaluations are performed using various neural network architectures and training strategies under a range of common imaging and device-induced uncertainties. Step 4: Insights from these evaluations guide iterative improvements to both the device design and image acquisition protocols–enhancing throughput, sensitivity, cost-effectiveness, and accessibility.

final accuracy and AUC performance. We believe it is crucial to evaluate the robustness of different neural network models and benchmark their performance under varying conditions.We also demonstrate that multiple image degradations have class dependent effects, differing between positive and negative cases. In particular, color and tone degradations interact with lesion appearance, creating dependencies between image quality and lesion characteristics. As a result, sensitivity and specificity can shift asymmetrically depending on the colorimetric and tonal degradations applied. This uncovers an important insight that previous work has not addressed, that inappropriate imaging device settings can significantly affect outcomes and bias models toward positive or negative predictions, with consequences for both clinical decisions and fair evaluation. We believe that this can ultimately help advance the field by enabling more accurate and reliable early detection of oral cancer using deep learning.

3. Experiments

In this section, we first introduce our oral datasets, which were acquired using an in-house built imaging device, as well as skin datasets obtained from the ISIC repository. We then describe the neural network architectures used for evaluation, along with the training procedures applied to each model. Next, we present various types of image degradation commonly encountered during image capture. These degradations are tested across different neural network architectures and are categorized into two main types: (1) colorimetric and tone degradations, and (2) capture-induced and optical distortions. Both categories can arise either from the imaging device itself or due to user-related factors during scene capture. Finally, we describe the evaluation metrics used to assess the performance of the trained classifiers for both oral and skin datasets.

3.1. Datasets

3.1.1. Oral dataset

The oral image dataset was acquired using a custom-built, portable, and cost-effective imaging device specifically designed for intraoral photographic documentation (Uthoff et al., 2018a; Yang et al., 2023). A total of 3991 clinically sourced images were collected from a diverse patient cohort and rigorously curated by experienced oral healthcare professionals for the purpose of training and validating diagnostic models. These images were then separated based on the presence of diagnostically relevant features indicative of oral potentially malignant disorders (OPMDs) and oral cancers (Birur N et al., 2022; Keerthi et al., 2023). Of the total dataset, 2281 images were classified as normal or non-cancerous, while 1710 were labeled as clinically suspicious for OPMDs and oral malignancies, as illustrated in Fig. 2. To evaluate model robustness under real-world conditions, we further examined the impact of eight distinct types of image degradation on intraoral photographs. These degradation modes–described in detail in Section 3.4–were designed to reflect common artifacts encountered during clinical image acquisition, including those arising from suboptimal lighting, motion blur, and optical defocus.

3.1.2. Skin dataset

The skin image dataset used in this study originates from the International Skin Imaging Collaboration (ISIC) archive (Cassidy et al., 2022), which aggregates data from three major sources: BCN20K, HAM10K, and the ISIC Challenge 2020 dataset (Hernández-Pérez et al., 2024; Tschandl et al., 2018). For this work, we selected only cases corresponding to melanocytic lesions, specifically nevus and melanoma classes, resulting in a total of 12,405 dermoscopic images–comprising 6705 nevus cases and 5700 melanoma cases. We present a subset of the selected

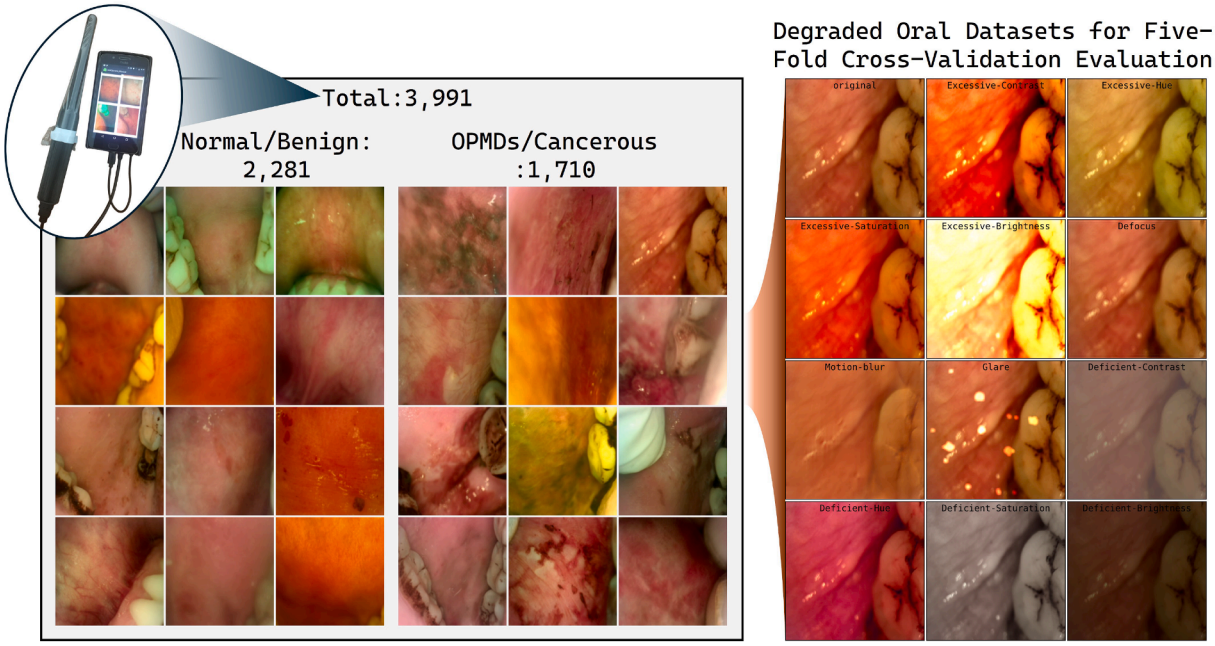


Fig. 2. Oral Dataset. On the left, we present a sample oral image acquired using our custom-built oral photographic device. From the collected dataset, we selected 3991 viable images for training and validation. The dataset was categorized into two groups: suspicious/positive cases representing oral potentially malignant disorders (OPMDs) and cancerous lesions, and negative cases representing normal and benign oral mucosa. On the right, we present an example reference image from the oral dataset. Various types of image degradation were applied to this image to qualitatively illustrate how these alterations can affect visual clarity and potentially impact clinical diagnostic accuracy.

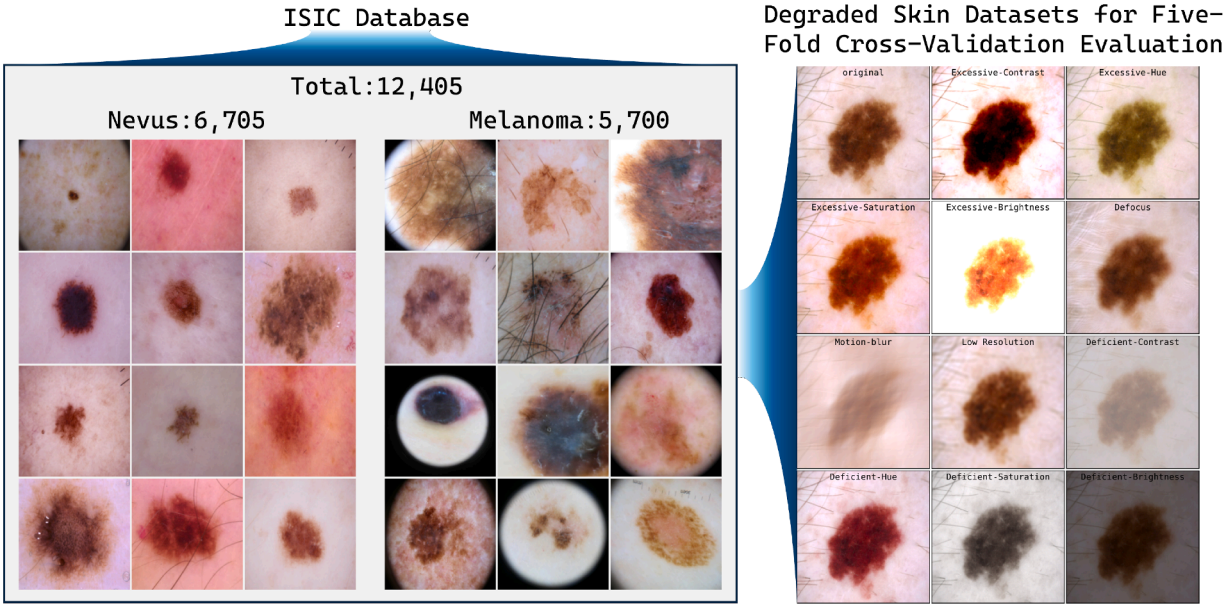


Fig. 3. Skin Dataset. On the left, we present representative examples of nevus (negative) and melanoma (positive) cases, selected from a combined dataset of 12,405 images sourced from three public repositories: BCN20K, HAM10K, and the ISIC Challenge 2020. On the right, we show a reference image from the skin dataset, to which we applied various types of image degradation. This visualization qualitatively demonstrates the visual impact of each degradation type and highlights their potential influence on skin clinical diagnostic outcomes.

dataset in Fig. 3. Given that photographic images are similarly susceptible to various forms of image degradation, we applied the same set of degradation conditions to this skin dataset (Fig. 3). This allows for a consistent benchmarking framework across both our intraoral dataset and an established public dataset, facilitating a more comprehensive evaluation of network robustness under clinically relevant image quality variations.

3.2. Neural network architecture and training strategies

To provide a comprehensive and state-of-the-art assessment of various network architectures and their corresponding training strategies, we conducted a comprehensive evaluation across symbolic models, including CNNs, transformers, and foundation models. For both CNNs and transformers, we used models pretrained on ImageNet-22k as the

initialization baseline. These models then were trained with different data augmentation techniques to evaluate robustness.

3.3. Architecture

Convolutional Neural Networks (CNNs) have long been regarded as foundational and highly effective for image classification tasks. In this study, we investigate several widely adopted CNN architectures in approximate chronological order of their introduction: VGG-19 (Simonyan & Zisserman, 2014), ResNet-50 (He et al., 2016), DenseNet-121 (Huang et al., 2017), and EfficientNet-B2 (Tan & Le, 2019). These models represent key milestones in CNN design and continue to serve as strong baselines for performance benchmarking.

Transformer-based architectures have recently demonstrated substantial improvements in image classification tasks, owing to their attention-based mechanisms and ability to model long-range dependencies. In our evaluation, we focus on vision transformers that use only the encoder component as the backbone for feature extraction. The models selected include Vision Transformer (ViT, Dosovitskiy et al., 2020), DeiT (Touvron et al., 2021), Swin Transformer (Liu et al., 2021), and EfficientFormer-L3 (Li et al., 2022). These models represent recent advancements in transformer-based image classifiers and are compared directly against CNNs to assess performance differences.

We also compared the two previous groups with current deep learning foundation models–Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) and Sigmoid Loss for Language-Image Pre-training (SigLIP) (Zhai et al., 2023)–to evaluate whether these models offer any advantages over conventional approaches in standard binary image classification tasks.

3.3.1. Training

To ensure consistency and minimize variability in the training process, all models were initialized with fixed random seeds, and pretrained weights were directly loaded. We used the Adam optimizer with a learning rate of 1e-4 to perform gradient descent during training. Early stopping was applied when both validation loss and validation accuracy did not improve for 10 consecutive epochs. All images from the oral and skin datasets were resized to 224×224 pixels, with a patch size of 32, and training was performed using the cross-entropy loss function. All training and validation were conducted using a 5-fold cross-validation setup, applied uniformly across all augmentation strategies. These strategies included: (1) no augmentation (only normalization), (2) spatial geometric augmentations such as cropping, flipping, rotation, and scaling, and (3) combined spatial and color augmentations including adjustments to contrast, brightness, saturation, and hue. Training was performed in a Python 3.10.9 environment using PyTorch, with all models trained on an NVIDIA V100 GPU using CUDA Toolkit 11.8.

3.4. Image degradation

We applied four common types of colorimetric degradations: brightness, contrast, saturation, and hue to simulate basic image distortions. These distortions may result from device limitations or user errors, such as unintentionally using incorrect settings during image acquisition. In addition, we introduced four common capture-induced and optical distortions frequently encountered in clinical trials: motion blur, defocus, resolution degradation, and glare. These degradations can arise from patient movement due to the use of handheld imaging devices, the limited depth of field of the imaging system, as well as environmental factors and sample-related issues such as uneven and specular surfaces. The simulation details for each type of degradation are described below.

3.4.1. Colorimetric and tone degradation

For colorimetric and tonal degradations, we tested brightness, contrast, saturation, and hue, each across 21 levels ranging from −10 to +10. For example, a brightness level of −10 results in a very dark image, while +10 produces a highly brightened image. A level of 0 corresponds to the original, unaltered image.

Each type of degradation was individually applied to a five-fold cross-validation dataset. We also tested combinations of two degradation types, resulting in a two-dimensional matrix with $N \times N$ data points, where $N = 21$ represents the number of levels tested for each degradation type. The chosen levels were qualitatively validated and confirmed based on clinical relevance and practical scenarios observed in photographic based diagnostics.These four colorimetric and tonal degradations capture major issues that arise during clinical image acquisition due to device limitations in camera or sensor control, dynamic range, and variable illumination common in photographic clinical settings. These degradations often occur together; For example, saturated LED lighting or improper exposure, whether too long or too short, can simultaneously alter brightness and saturation, leading to the loss of subtle, diagnostically important features. When contrast is set incorrectly, it can also exaggerate irrelevant details, such as saliva in intraoral images and hair in skin imaging diagnostics. Because clinical photographs are typically captured with color cameras, the visual manifestations of cancerous or infected tissue in skin and oral imaging are closely tied to color appearance, which makes hue control and calibration particularly important, yet some devices lack hue adjustment altogether or provide poorly tuned defaults for clinical use. These effects can strongly affect the final clinical diagnosis, increasing the risk of missed findings or false positives.

3.4.2. Capture-induced and optical distortion

The four common types of frequency-related image degradation we aim to simulate for oral photographic imaging devices are: defocus, motion blur, resolution loss, and glare. These degradations commonly occur in clinical practice and can significantly affect image quality and diagnostic performance. Defocus is one of the most prevalent issues across both conventional photography and optical imaging systems. It typically results from limitations in depth-of-field relative to the system's resolution. To simulate defocus, we explicitly isolate its effect using the theoretical form of optical aberration: the second-order Zernike polynomial (Von F, 1934) corresponding to defocus. This is implemented in the Fourier domain (Goodman, 2005) by modulating the pupil function, allowing us to simulate pure defocus artifacts without the influence of higher-order aberrations. The applied defocus model is expressed as follows:

$$I_{defocus}(x,y) = \left| \mathcal{F}^{-1}[\mathcal{F}\{I_{original}(x,y)\} \exp(i\alpha Z_2^0)] \right|^2, \tag{1}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the 2D Fourier transform and its inverse, respectively. The defocus is modeled using the second-order Zernike polynomial $Z_2^0(\rho,\theta) = \sqrt{3}(2\rho^2 - 1)$, which represents the defocus component of the wavefront in the pupil function. The parameter α is varied to simulate 10 severity levels of defocus in the image. Motion blur can also be simulated in Fourier space as phase shifts across the entire frequency domain, using a motion blur kernel (Gonzalez, 2009) expressed as:

$$H_{motion-blur}(k_x,k_y) = \frac{1}{k_\theta(k_x,k_y)} \sin(k_\theta(k_x,k_y)) \exp(-j(k_\theta(k_x,k_y)), \tag{2}$$

$$k_\theta = k_x + k_y = k_0(\omega) \cdot \sin(\theta) + k_0(\omega) \cdot \cos(\theta). \tag{3}$$

Here, H denotes the motion blur kernel in the Fourier domain. The severity of the blur can be adjusted using ω, while the direction is controlled by θ. We pre-compute this motion blur kernel for $\theta \in \{0°, 18°, 36°, 54°, 72°, 90°\}$ across 10 levels of ω. For each severity level, a random direction θ is applied to the cross-validation dataset. The simulated motion blur image can be readily expressed by the following

equation,

$$I_{motion-blur} = \left| \mathcal{F}^{-1}[\mathcal{F}\{I_{original}(x,y)\}H(k_\theta)] \right|^2. \quad (4)$$

To simulate resolution degradation, we applied a circular low-pass filter in the frequency domain by progressively shrinking its radius, thereby removing high-frequency components from the original captured images. Glare artifacts were simulated by controlling the number and intensity of saturated regions to mimic the specular reflections commonly observed in oral images. These reflections, primarily caused by saliva, often lead to significant local overexposure, which can adversely affect image quality and reduce clinical diagnostic accuracy. In our robustness evaluations, we apply image degradations primarily in the Fourier domain using 2D Fast Fourier Transforms (FFT) followed by element-wise multiplication. This method is particularly advantageous for large kernel sizes or global operations, where it offers significant computational efficiency compared to direct spatial-domain convolutions. While FFT-based methods may not outperform spatial convolutions for small kernel sizes due to transform overhead, they become increasingly efficient for large images and batch processing. Furthermore, this approach is highly compatible with CUDA acceleration, enabling scalable and fast evaluation across large datasets.

All of the above capture induced and optical distortion related degradations are common in photographic clinical diagnostics, especially when the imaging system is constrained by optical design and the need for portability. Limited depth of field increases the risk of defocus, particularly with handheld devices. This can significantly reduce resolution and cause loss of fine details that are critical for diagnosis. There are practical trade offs between frame rate and image resolution: although higher resolution is desirable, lower frame rates make motion blur more likely, which degrades information, introduces artifacts, and elongates features that may mislead diagnosis. Unexpected glare from saliva can occlude relevant oral structures or be mistaken for important findings. The simulated severity levels for these degradations were qualitatively validated and thoroughly compared with our collected data, confirming that they closely reflect conditions encountered in clinical imaging and diagnosis.

3.5. Evaluation metrics

We evaluate our model's robustness using four main metrics: sensitivity, specificity, accuracy, and the area under the receiver operating characteristic curve (AUC). These metrics offer critical insights into how various forms of image degradation influence diagnostic performance. In terms of sensitivity and specificity. Our findings indicate that certain types of degradation disproportionately affect the detection of abnormal cases, likely due to alterations in color and feature representation that are more susceptible to quality loss. In the oral imaging datasets, high-contrast degradation appears to have a greater impact on specificity. This may be attributed to the fact that normal oral mucosa typically exhibits smoother, more uniform textures with less prominent features, making it more susceptible to misclassified as positive case under excessive contrast-enhanced conditions. In contrast, the sensitivity remains relatively unaffected, likely because abnormal cases retain distinct morphological characteristics even when image contrast is altered. In addition to these conditional metrics for positive and negative cases, we also utilized top-1 accuracy and AUC as aggregate indicators to provide a comprehensive assessment of each model's ability to distinguish between normal and abnormal cases. This approach establishes a valuable benchmark for understanding the differential impact of image degradation on diagnostic outcomes. The results have important implications for the development of next-generation AI-assisted clinical imaging devices, particularly in guiding design priorities–whether to enhance sensitivity or specificity–based on the degradation profiles most likely to be encountered in real-world settings.

4. Results

We present the evaluation results of the trained neural networks, which demonstrate state-of-the-art performance. The results are organized into two main parts for both the oral and skin datasets. First, we evaluate the models under isolated image degradation conditions, where each type of degradation is applied individually at varying levels. This is followed by evaluations under mixed degradation conditions, where multiple degradations are applied simultaneously. In real-world clinical settings, it is common for several types of image degradation to occur concurrently. These combined effects are often non-linear and non-additive, making them more challenging to address. To better reflect these realistic conditions, we simulated various combinations of degradations and tested the performance of the classification models under these scenarios.Importantly, we did not simply apply all possible degradations indiscriminately. Instead, our goal was to identify which specific types or combinations of degradations have the most significant impact on model performance. This insight is essential for designing next-generation imaging systems that are optimized for practical use. By focusing on the most critical degradation scenarios, we can reduce development cost and effort while enhancing classification robustness in real-world applications.

4.1. Oral evaluation results

We evaluated the performance of our oral disease classifier under various types of imaging degradation, as illustrated in Fig. 2, with severity levels defined in Section 3.4. A five-fold cross-validation was conducted across all neural network architectures. Additionally, we tested three different augmentation strategies for both VGG-19 and Swin Transformer models. To further investigate the combined effects of common degradations, we applied mixed dual-image degradation scenarios for advanced analysis of their impact on classification performance in 4.1.3.

4.1.1. Independent image degradation evaluation

We evaluated the robustness of our oral disease classifier using five-fold cross-validation under various image degradation conditions. For each fold, the validation dataset was first subjected to image degradation, followed by normalization, before being input into the trained model. The results are presented as line plots showing the mean values of four evaluation metrics, with standard deviation represented as shaded areas around each curve, across all types of image degradation. We began by analyzing the performance of convolutional neural networks (CNNs), which are widely used as benchmark models for image classification tasks. The performance of these representative CNNs is shown in Fig. 4. Among the four degradation types–contrast, brightness, saturation, and hue–hue had the least impact on classification performance, likely because subtle color shifts do not significantly alter the structural features of the sample images. In contrast, specificity was particularly sensitive to changes in contrast, as normal (negative) oral mucosa typically exhibits smooth, uniform pink surfaces with minimal distinct features. When contrast is excessively enhanced, these subtle features may be artificially sharpened, leading to false positives and reduced specificity. However, this same enhancement does not significantly affect sensitivity, as the features of abnormal cases remain prominent. Interestingly, when contrast is significantly reduced, the opposite effect is observed: sensitivity drops sharply while specificity remains relatively stable. This suggests that low contrast obscures critical features necessary for detecting abnormal cases. Brightness also had a substantial impact on specificity when either excessively increased or decreased. This is expected, as overexposure results in uniformly bright images that wash out important features, while underexposure leads to dark images where features are obscured–both of which hinder accurate classification. Among the four degradation types, brightness had the most pronounced effect on performance. Saturation degradation led

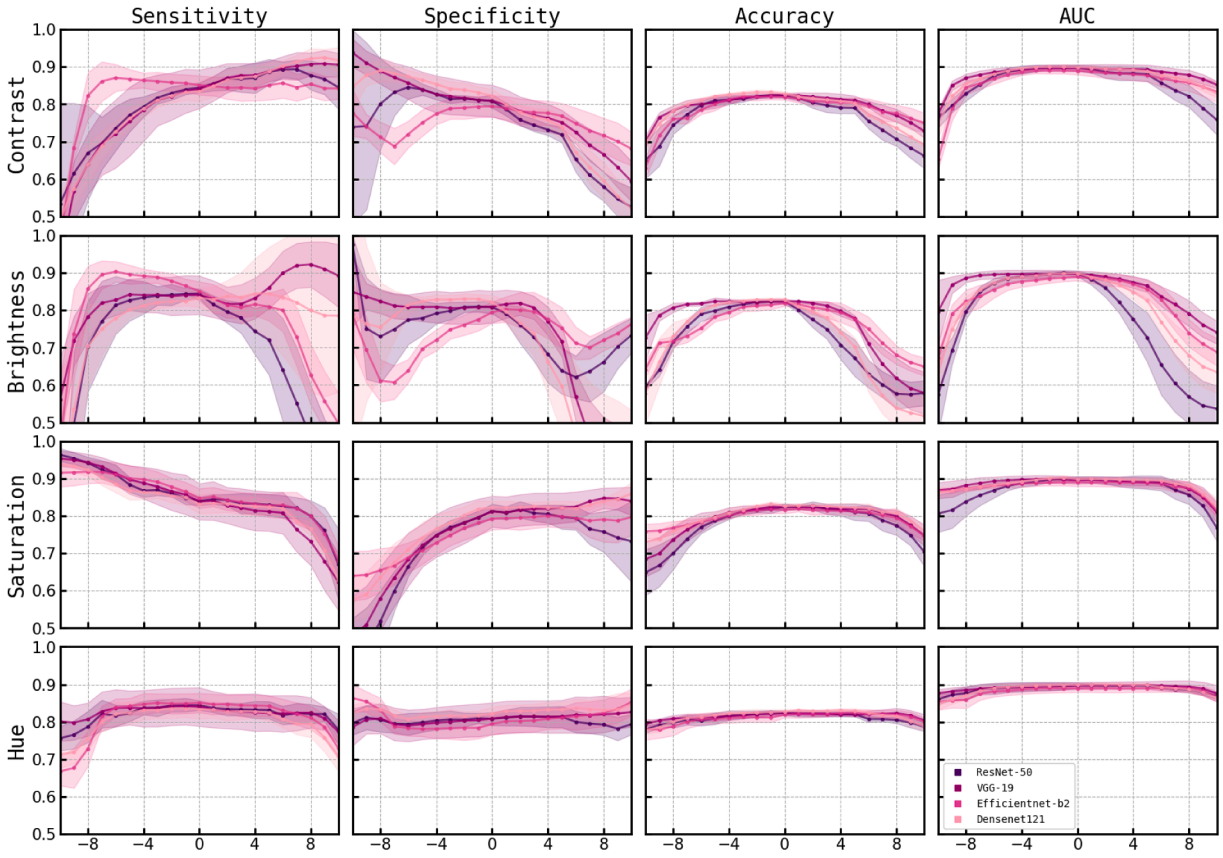


Fig. 4. CNN-based oral image robustness evaluation under colorimetric and tonal degradations. The 4×4 grid of plots illustrates how CNN performance varies when different types of image degradation are independently applied. Each plot presents one of four evaluation metrics across 21 severity levels (ranging from −10 to +10). Solid lines represent the mean values from five-fold cross-validation, while the shaded areas indicate the standard deviation at each severity level. Among the models evaluated, VGG-19 demonstrated the most robust performance for oral image classification across all degradation types.

to an increase in false negatives. As color saturation decreased, images appeared smoother and less distinct, causing the classifier to misidentify abnormal cases with sharp features as normal. This effect was inversely reflected in specificity, which improved under desaturated conditions for the same reason. Among the CNNs we tested on oral dataset, VGG-19 maintained better performance under degraded image quality, particularly in deficient brightness (−10 severity level) conditions, where it achieved 15 % to 22 % higher AUC than the other CNNs.

In the CNN robustness evaluation, we are able to extract several valuable insights regarding the effects of the four types of image degradation. Among the CNN-based architectures tested, VGG-19 consistently outperformed the others, as shown in Fig. 4. Based on this performance, we selected VGG-19 as the representative CNN model for comparison with recently introduced attention-based transformers, as illustrated in Fig. 5. The overall trends in sensitivity and specificity observed in the transformer models were consistent with our previous analysis. However, Vision Transformers (ViT) did not perform as well as other transformer-based models or VGG-19 in the oral clinical images classification task. ViT demonstrated much stronger performance in the skin dataset classification task compared to oral, suggesting its potential as a promising candidate in a different domain. We will discuss these findings in more detail in Section 4.2.1. Overall, transformer-based models demonstrated greater robustness compared to VGG-19 and other CNN-based classifiers. Among them, the Swin Transformer achieved the best performance with an AUC of 0.92 under original (non-degraded) image conditions. It also demonstrated the strongest robustness across all types of image degradation. The maximum AUC decreases under contrast, brightness, saturation, and hue degradations were −0.07, −0.13, −0.06, and −0.02, respectively. The second-best model, DeiT, achieved an AUC of 0.91 without degradation, with corresponding performance drops of −0.12, −0.15, −0.09, and −0.05 under the same conditions. Both worst-case performances occurred at the +10 brightness severity level.

We further compared transformer- and CNN-based models with two widely used foundation models: CLIP and SigLIP, under the same robustness evaluation framework. However, these foundation models did not outperform the task-specific models across varying levels of image degradation or severity in the binary classification of oral images, as shown in Fig. 6. We attribute this performance gap to several factors. First, both CLIP and SigLIP are primarily trained for image-text alignment tasks, whereas in our experiments, we utilized only their image encoder components as feature extractors for classification. This limited adaptation may hinder performance on specialized tasks. Second, there is a significant domain mismatch between the training data of these foundation models and our intraoral imaging dataset. CLIP and SigLIP are trained on large-scale, general-purpose datasets sourced from the internet, which differ substantially in content, resolution, and visual patterns from clinical oral photographs. These findings highlight the need for open-source, domain-specific foundation models tailored to photographic medical imaging applications–where subtle visual cues and high diagnostic fidelity are essential for accurate disease detection.

We also evaluated the impact of capture-induced image degradations on oral image classification performance. This is particularly important for handheld devices, where motion blur frequently occurs in clinical settings due to instability from either the patient or the imaging device. Additionally, intraoral imaging often suffers from defocus blur caused by uneven surface geometry, making it more challenging to acquire diagnostically useful images. A third common issue is glare, typically caused by saliva or other fluids in the mouth, which can result

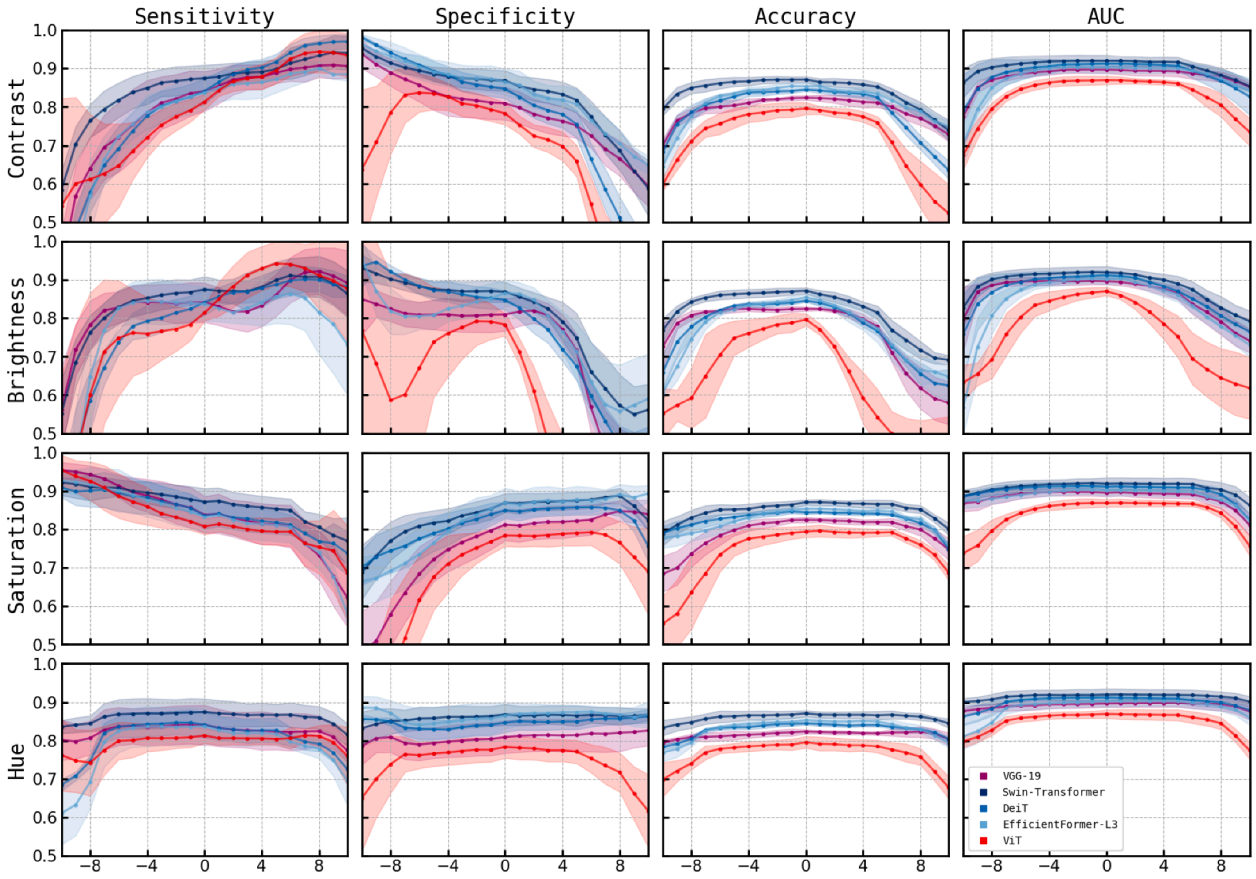


Fig. 5. Transformer-based oral image robustness evaluation (4 × 4 grid of plots). Here, we compares the performance of transformer models with VGG-19, which serves as a reference CNN-based classifier for oral image classification. Overall, transformer models demonstrated superior robustness across various types of image degradation. The ViT underperformed relative to the other transformer models, which may be attributed to the nature of the oral imaging task and the limited size of the dataset. ViT performed better on the skin dataset, and this improved performance points to a possible dependency on data type or domain. Among all models, the Swin Transformer achieved the highest overall performance across all evaluation metrics and degradation types.

in saturated reflections that obscure critical tissue features. This is a domain-specific degradation frequently encountered with photographic imaging for cancer detection. We conducted the same robustness evaluation on these three types of image degradations, motion blur, defocus, and glare, as shown in Fig. 7. The Swin-Transformer remained robust across most degradation conditions. The Swin Transformer and DeiT demonstrated strong robustness across most image degradation conditions, with accuracy decreasing by only about 1 % to 2 % until a significant drop occurred at severity level +7 and beyond. Although CLIP and SigLIP generally underperformed compared to CNN- and transformer-based models, they exhibited considerably smaller performance declines under severe defocus conditions. Specifically, SigLIP's accuracy decreased by only 0.02 to 0.72, and its AUC dropped by just 0.03 to 0.78, whereas the Swin Transformer experienced a larger decrease, with accuracy dropping by 0.20 to 0.64 and AUC dropping by 0.22 to 0.68.

4.1.2. Augmentation impact in training phase

For all previous assessments and evaluations, models were trained using both geometric and color-space augmentations. In this section, we investigate how different augmentation strategies used during the training phase affect model performance (Liu et al., 2024; Martínez-Martínez et al., 2025). Geometric augmentations included random rotations, flips, cropping, and scaling, applied consistently across the training data using a fixed random seed for all five-fold cross-validation splits. Color augmentations involved random changes in contrast, brightness, saturation, and hue, also applied with a fixed seed to ensure consistency across folds. To illustrate the impact of these augmentation strategies, we compare the performance of two representative models: VGG-19 and Swin Transformer, under three conditions: (1) no augmentation (baseline with only normalization), (2) geometric augmentation only, and (3) geometric plus color augmentation. The results are summarized in Fig. 7. We observed that adding more augmentation during training does not necessarily lead to improved robustness when evaluated under different types of image degradation. In fact, training with only geometric augmentations sometimes confused the model, reducing its ability to handle colorimetric degradations and resulting in overall worse performance compared to training without any augmentation, particularly on the oral image dataset. However, incorporating colorimetric augmentations during training made the Swin Transformer model more robust to saturation and hue degradations compared to the no-augmentation baseline. On the other hand, the model trained without any augmentation still performed better under brightness and contrast degradations. This demonstrates the need to carefully tailor augmentation strategies to the specific degradation types likely to be encountered in the target domain. To conclude, implementing both geometric and colorimetric augmentations improves overall performance. In a few cases, the mean accuracy was slightly lower than with non augmented training. However, the standard deviation across five fold cross validation was consistently smaller with augmentation. We also observed more uniform improvements in AUC across all models and across all degradation categories. For our in house oral dataset, which is relatively small, performance can be inconsistent, but augmentation still reduces variance.

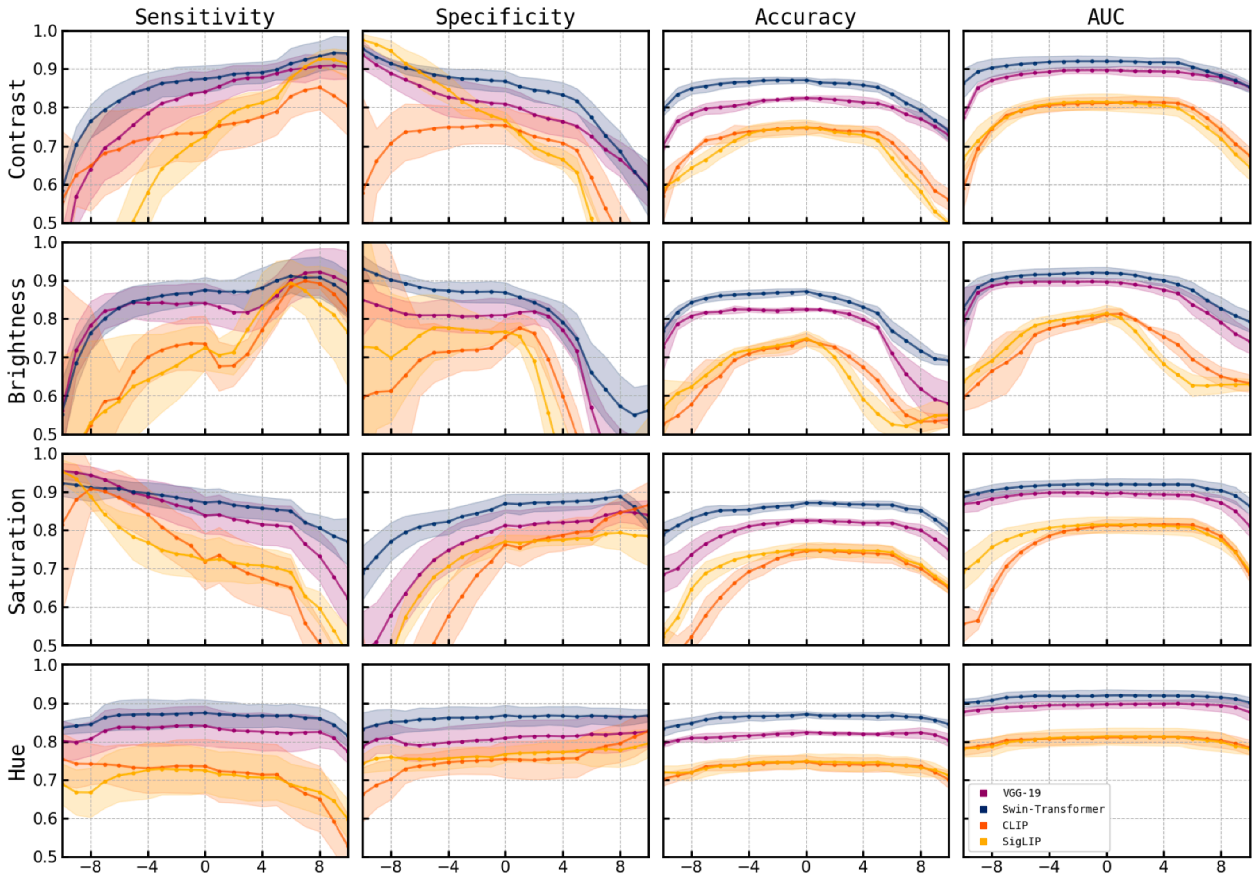


Fig. 6. Robustness evaluation of foundation models (CLIP and SigLIP). We present a 4×4 grid of classification performance plots for CLIP and SigLIP under various image degradation types and severity levels, compared against VGG-19 and Swin Transformer. The foundation models consistently underperformed relative to both the transformer- and CNN-based task-specific models across all conditions. We attribute this to two main factors: (1) only the image encoder components of CLIP and SigLIP were used for classification, and (2) the training datasets for these models primarily consist of natural, everyday scenes, which differ substantially from photographic medical images. These results highlight the importance and growing need for foundation models specifically designed for medical imaging tasks.

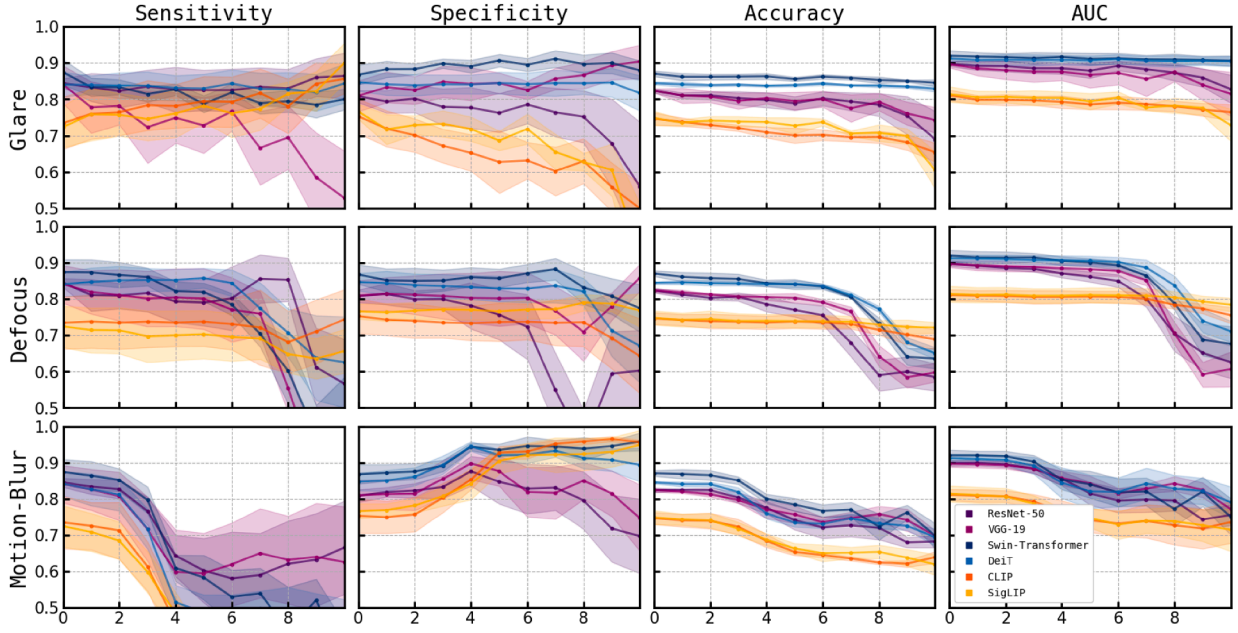


Fig. 7. Oral robustness evaluation on capture-induced and optical distortion image degradation. The 4×4 grid plots depict the performance of different neural network architectures under four evaluation metrics when 10 levels of glare, motion blur, and defocus are applied to the images.

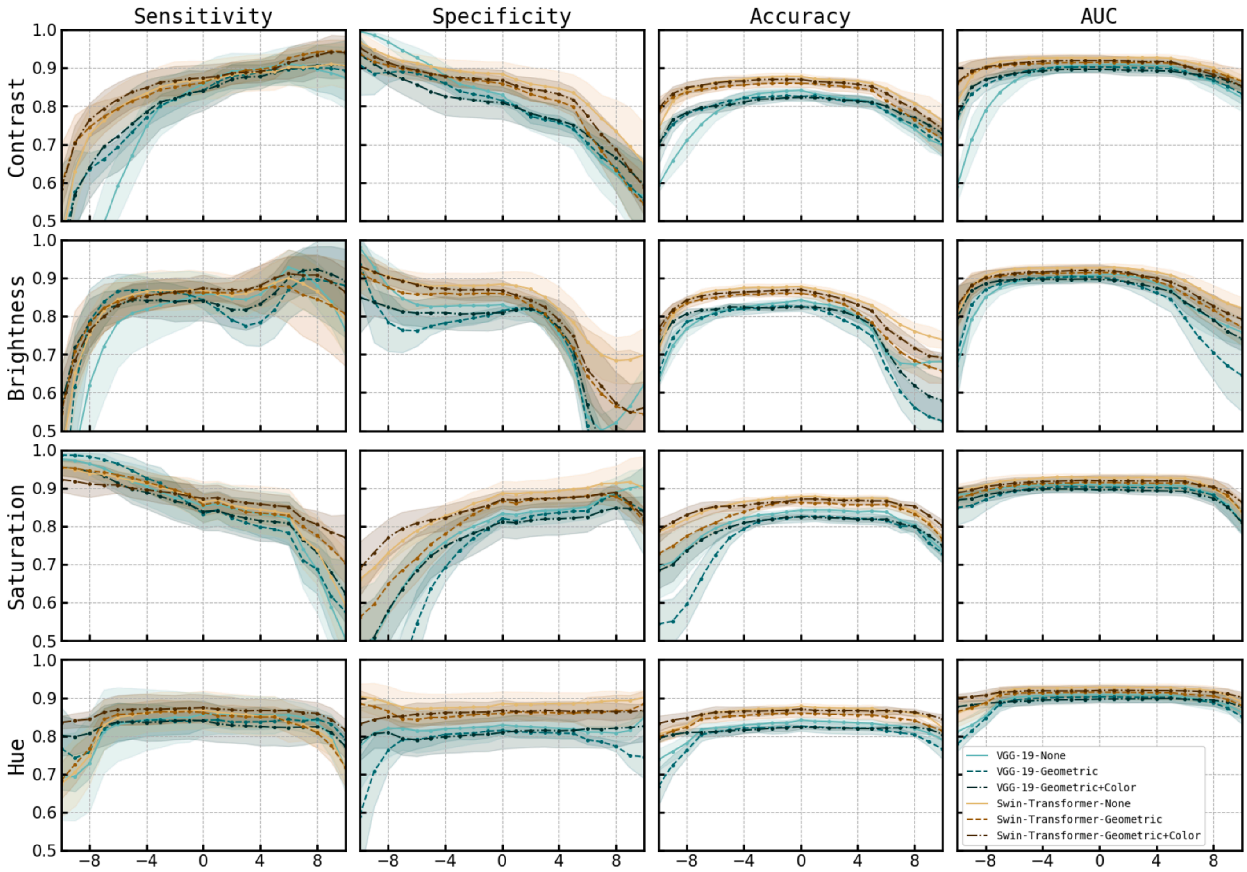


Fig. 8. Comparison of VGG-19 and Swin transformer performance under different training augmentation strategies for oral image classification. The plots depict the impact of different data augmentation strategies during training: no augmentation, geometric augmentation only, and combined geometric and color augmentation. Darker lines represent models trained with more extensive augmentation. The comparison highlights how augmentation influences robustness and classification performance on oral images.

4.1.3. Mixed image degradation evaluation

In most clinical settings, image degradation is often complex and typically involves multiple types of degradation occurring simultaneously. To analyze this, we conducted an advanced evaluation of mixed colorimetric degradations by testing combinations of two degradation types across a 121-point (11×11) grid. This evaluation was performed across all five cross-validation folds, as shown in Fig. 9. The corresponding standard deviation plot is presented in Fig. A.16. This analysis provides valuable insights into how classifier performance changes when two degradations occur together at varying levels. In our experiments, the order of applying image degradations was fixed: brightness, followed by contrast, saturation, and then hue. In Fig. 9, the degradation labeled on the x-axis is applied first, followed by the one on the y-axis. We report the mean AUC (in percentage) across five-fold cross-validation using the Swin Transformer model, which consistently outperformed other architectures in previous evaluations. From earlier single-degradation experiments, we observed that excessively increasing brightness reduces classification accuracy and AUC more than decreasing it. However, in the brightness-contrast matrix, we found that when contrast was enhanced, the negative effects of low brightness were significantly amplified, especially in the lower left corner of the matrix. This suggests a nonlinear interaction between brightness and contrast, where one degradation can exacerbate the effect of the other.A similar behavior was observed in the contrast-saturation matrix: performance was notably worse when contrast was at its maximum positive level compared to when it was reduced. These results indicate that under conditions of excessive contrast enhancement, the model becomes highly sensitive to additional degradations, leading to a significant drop in classification performance. We also observed an interesting effect when saturation and hue were applied simultaneously: the classification performance dropped significantly when saturation was enhanced, and the performance variation across different hue levels also increased markedly across all severity levels.

4.2. Skin evaluation results

We evaluated the performance of our skin classifier in distinguishing nevus (negative) from melanoma (positive) using a robustness evaluation approach similar to that used for the oral dataset. We explored different model architectures–including CNNs, transformers, and foundation models–and evaluated their performance under individual image degradations. The skin dataset belongs to a different domain and contains a larger number of training samples compared to our in-house collected oral dataset. We believe this not only serves as a valuable reference for comparing model performance across domains, but also broadens the relevance of our findings to a wider audience interested in skin cancer diagnostics. For the skin images, we additionally simulated resolution degradation, as high spatial resolution is often critical for accurately analyzing skin lesions. Examples showing how various image degradations visually affect skin images compared to a clean reference image are presented in Fig. 2. These results offer a meaningful point of comparison for evaluating data acquired using our in-house oral imaging device.

4.2.1. Independent image degradation evaluation

In this section, we conducted a similar analysis to that presented in Section 4.1.1. We compared different model architectures under

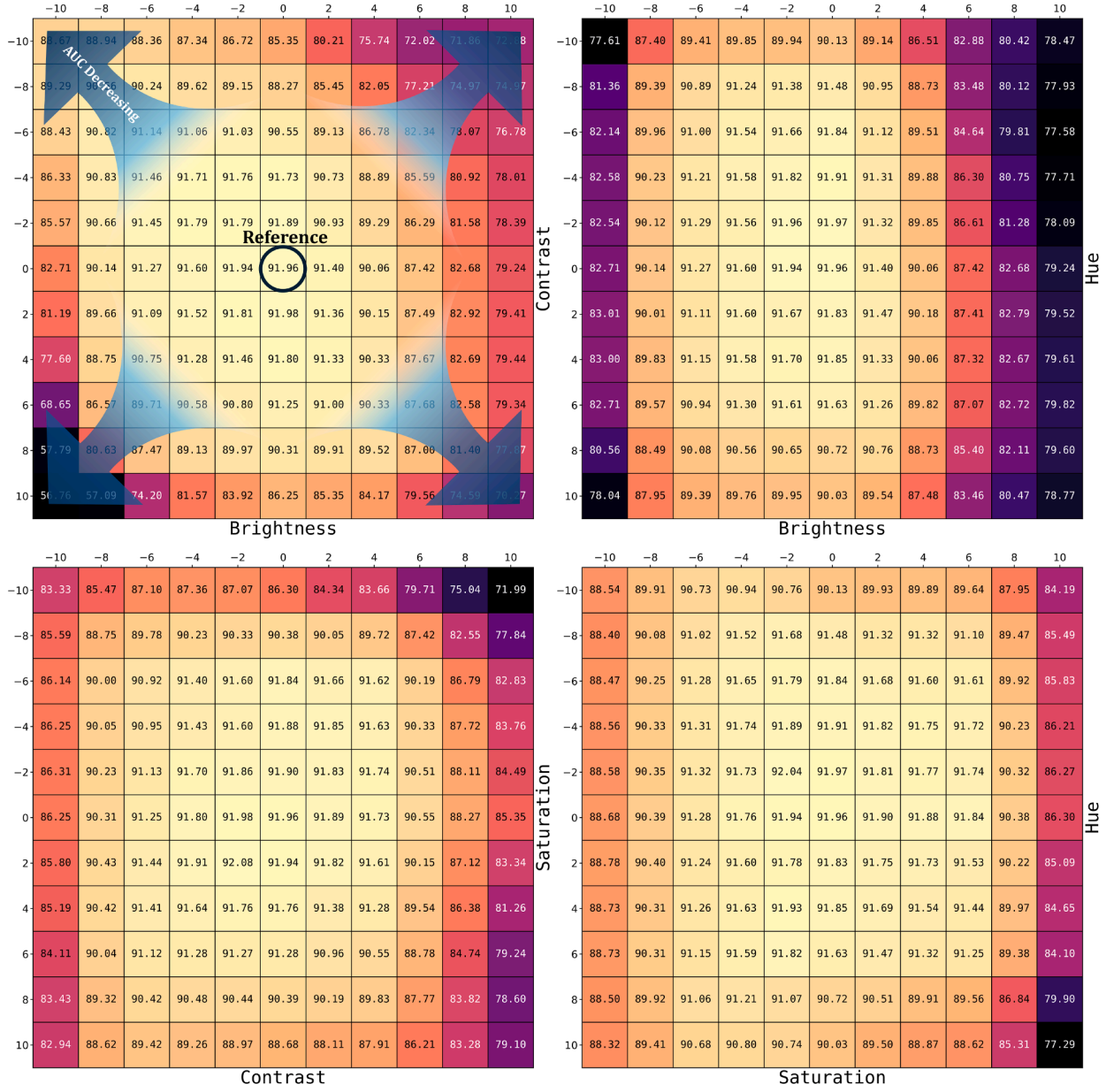


Fig. 9. AUC performance metrics for swin transformer under dual image degradations. The values represent the mean AUC (in percentage) across five-fold cross-validation. The black circled value in the first matrix indicates the reference AUC for the 2D matrix (no image degradation). In each subsequent matrix, the center value also denotes the reference (no-degradation) AUC for that matrix. The blue arrows highlight the decreasing AUC trend as the two degradation types are combined at higher severities. Each matrix shows results for 11 severity levels, ranging from −10 to +10, applied across different types of image degradations. The degradation on the x-axis is always applied first, followed by the one on the y-axis. Color maps are individually scaled for each matrix, with bright yellow indicating the maximum AUC value and black indicating the minimum. The corresponding standard deviation for each matrix is shown in Fig. A.16.

individual image degradations and evaluated performance using four standard metrics, forming a 4 × 4 grid of plots. Fig. 10 shows the results for CNN-based models. Overall, the relative performance across CNN architectures is consistent with our observations on the oral dataset. However, VGG-19 did not stand out as strongly as it did in the oral imaging task. The standard deviations across all evaluation metrics were slightly narrower compared to those from the oral dataset, and the models' performances were more closely aligned. Notably, sensitivity and specificity appeared more stable and more consistent with accuracy, in contrast to the greater variability observed in the oral dataset. We attribute this improved consistency to two main factors: (1) the skin dataset contains a larger number of training samples, which improves generalization across models, and (2) skin images exhibit less structural and background variation compared to oral images. Oral photographs are typically captured within the entire mouth cavity, often including complex geometries, occlusions, and varying viewpoints due to 3D tilt and uneven surfaces. In contrast, skin images are predominantly 2D, with more uniform backgrounds and simpler visual structures, contributing to greater consistency in model performance across degradations.

Given the relatively similar performance among CNN models, we selected VGG-19 as the representative CNN baseline for comparison with transformer-based models, as shown in Fig. 11. Overall, most transformer models outperformed VGG-19 across the various image degradation types. Notably, the Vision Transformer (ViT) performed significantly better on the skin dataset than it did on the oral dataset, where its performance had been closer to that of VGG-19. For example, under the highest brightness degradation level (+10), ViT achieved an AUC of approximately 0.85, compared to 0.76 for VGG-19 and

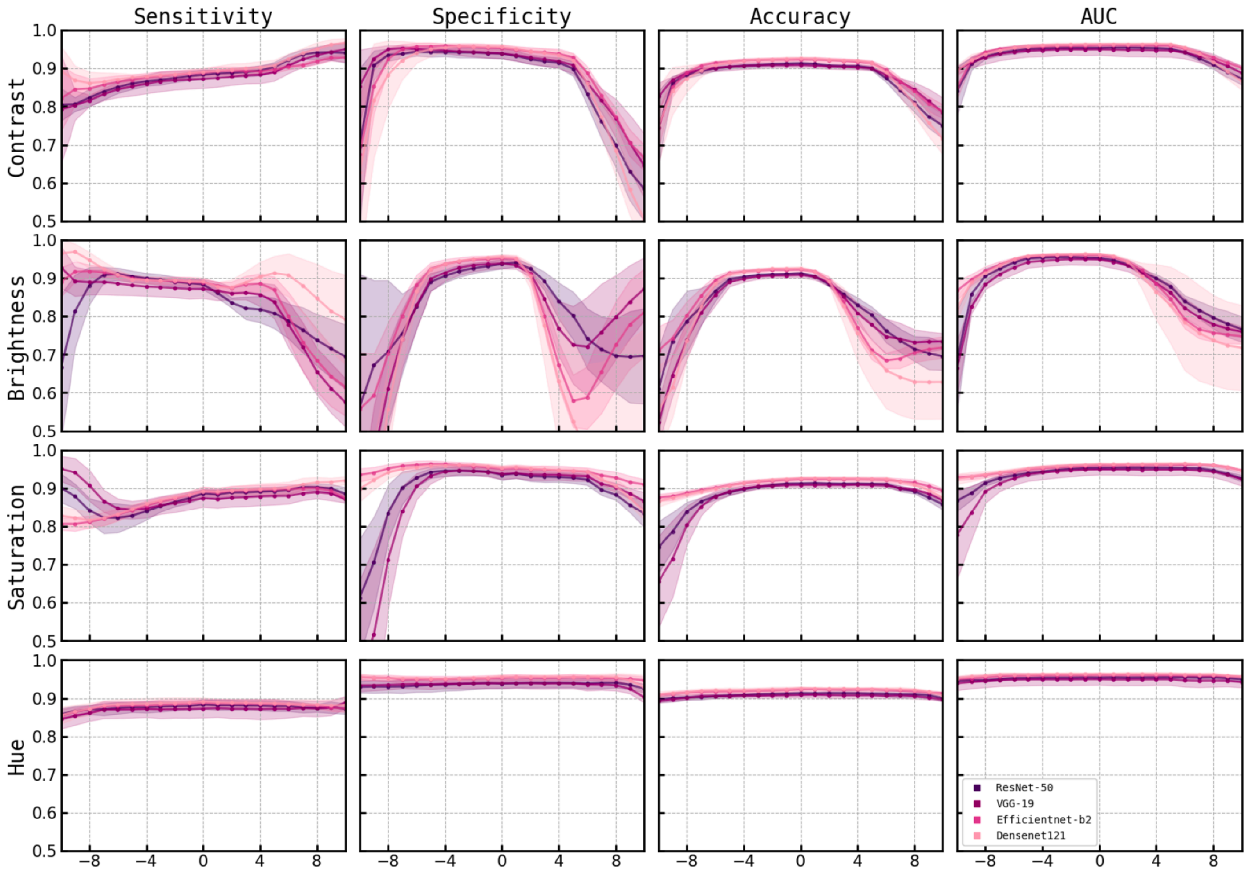


Fig. 10. CNN-based robustness evaluation on skin dataset under colorimetric and tonal degradations. The 4 × 4 grid of plots illustrates CNN performance when different types of image degradation are independently applied. Each plot shows one of four evaluation metrics across 21 severity levels (ranging from −10 to +10). Solid lines represent mean values obtained from five-fold cross-validation, and shaded regions indicate the corresponding standard deviations at each severity level.

0.84 for Swin Transformer. This suggests that ViT is more robust in the skin dataset. We believe this highlights the importance of domain characteristics–beyond dataset size–in influencing model performance. While ViT showed notable improvement, Swin Transformer still achieved the best overall performance, consistently outperforming other transformers by a margin of 1 % to 3 % across all types of image degradations.

We also compared the previously discussed models with foundation models on the skin dataset, following the same evaluation approach used in Section 4.1.1. Similar to our observations with ViT, both CLIP and SigLIP narrowed the performance gap with VGG-19 compared to what was observed in the oral dataset. This resulted in a more comparable performance between foundation models and VGG-19 on the skin classification task. However, Swin Transformer still showed a clear performance advantage over both foundation models across most degradation scenarios. These results are consistent with our findings from the oral dataset and support the expectation that general-purpose foundation models may not offer superior robustness for domain-specific tasks such as photographic clinical imaging. This highlights the need for tailored models that are better aligned with the characteristics of medical image data. We also observed that for capture-induced image degradations–specifically resolution, defocus, and motion blur–CLIP and SigLIP slightly outperformed other models under extreme defocus conditions (e.g., at the +10 severity level), achieving 1–2 % higher performance, as shown in Fig. 13. In contrast, ResNet-50 was highly sensitive to resolution degradation, exhibiting a significant drop in specificity–indicating an increase in false positive detections on the skin dataset–starting around severity level 7 of the resolution degradation.

4.2.2. *Augmentation impact in training phase*

We compared the performance of VGG-19 and Swin Transformer under different training augmentation strategies. Overall, the results are consistent with those observed in the oral dataset. However, excluding colorimetric augmentation during training led to a substantial performance drop–up to 30 % (under a +10 severity level in hue) for both VGG-19 and Swin Transformer–compared to models trained with full augmentation. Interestingly, the Swin Transformer trained without any augmentation outperformed the version trained with geometric augmentation alone. This suggests a potential interaction between model architecture and the type of augmentation used, highlighting that augmentation strategies may have degradation-specific effects. Given that subtle differences in skin color are important for accurate diagnosis, hue-based augmentation during training may be particularly beneficial for improving robustness in skin imaging tasks. Overall, when both geometric and color augmentations are applied, performance improves in both accuracy and AUC. Although the mean performance across the validation dataset is marginal in some cases, the standard deviation is significantly narrower than in models trained without augmentation for the skin datasets.

4.2.3. *Mixed image degradation evaluation*

We also used the Swin Transformer–identified as the best-performing model for the skin dataset–to conduct an advanced dual-image degradation evaluation, following the same approach used for the oral dataset (Section 4.1.3). The overall AUC performance of the Swin Transformer across the degradation matrix remained stable. Among the combinations tested, brightness and contrast were the most impactful, as shown in Fig. 15, while saturation and hue had a

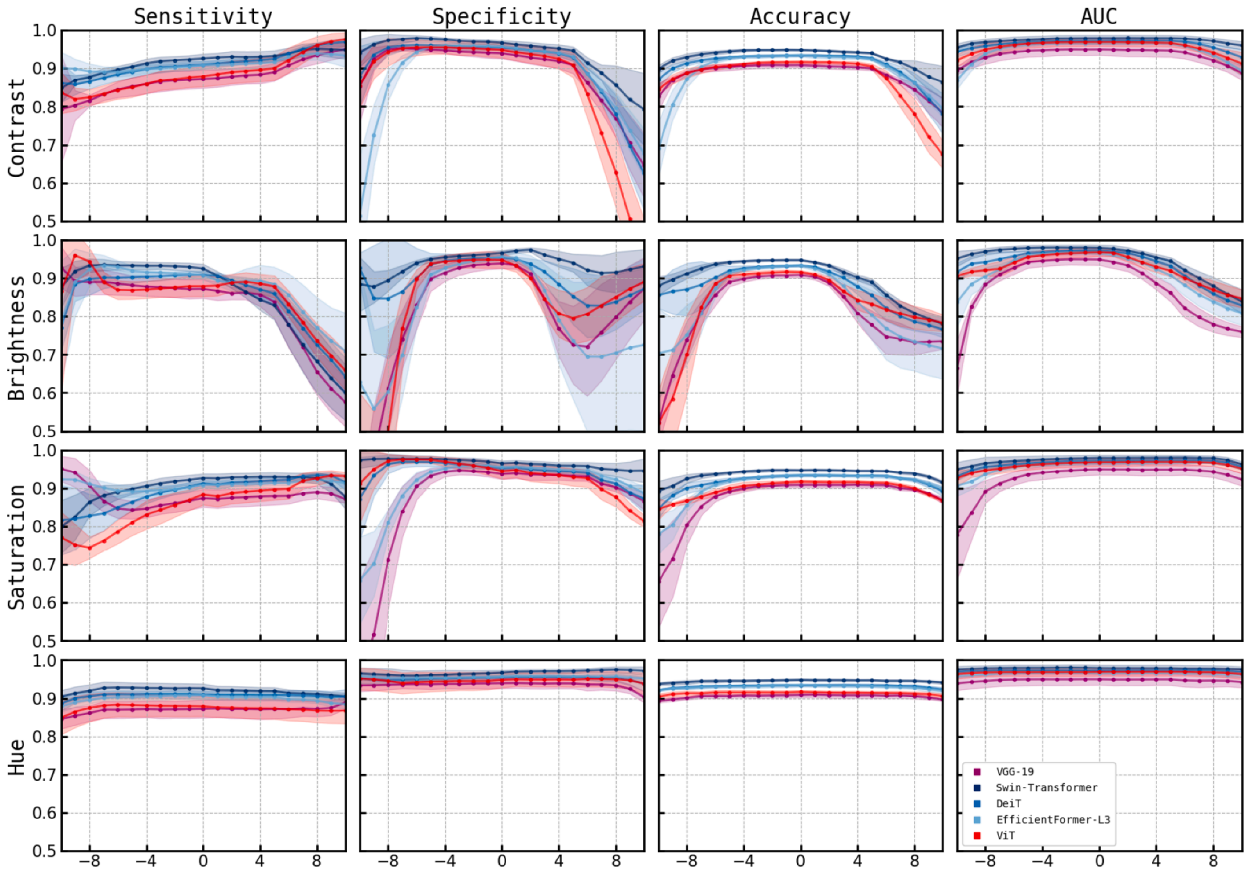


Fig. 11. Transformer-based model performance evaluation on skin dataset. Similarly, we compared the CNN-based VGG-19 model with several recent transformer architectures to evaluate their robustness under colorimetric image degradations.

comparatively minor effect on classification performance. The corresponding standard deviation map is presented in Fig. A.17, highlighting the consistency of the model under different degradation conditions.

5. Discussion

The rapid development of deep learning architectures has shown significant promise in image classification and pattern recognition tasks. Photographic imaging is emerging as an important tool in current and future clinical diagnostics, particularly in cancer detection (Bi et al., 2019; Song & Liang, 2025), as advances in low-cost sensors and 3D printing technologies (Hong et al., 2025; You et al., 2025a,b) make imaging devices more affordable and accessible. Accurate interpretation of these images has traditionally relied on the clinician's expertise and experience. However, assessing imaging outputs can be challenging, especially in the presence of variability across devices and imaging conditions. Deep learning models offer the potential to support and standardize this diagnostic process, reducing reliance on individual experience. Systematic distortions or artifacts introduced by the device–often due to limitations in its optical components or sensor design–can influence interpretation. Such device-specific variations may lead to misinterpretation if not properly accounted for. Therefore, it is equally critical for clinicians to be aware of these limitations and how they may vary between devices. These factors motivated our comprehensive robustness evaluation across a variety of image degradation types and model architectures. We believe this work provides valuable insights into how different neural network models respond to real-world imaging imperfections. Beyond model benchmarking, our findings also offer practical guidance for the development of next-generation AI-assisted photographic devices, particularly in clinical applications.

One of the key findings in our study is that recent attention-based transformer architectures consistently demonstrate greater robustness than conventional CNNs when handling photographic medical images. On average, transformers outperform CNNs by 1–5 % in terms of AUC and Top-1 accuracy under original image quality. When image degradation is introduced, this performance gap can widen significantly–up to 25 % in some cases. This improved robustness can be attributed to the self-attention mechanisms in transformer architectures, which enable the model to capture long-range dependencies and contextual features more effectively than the local receptive fields used in CNNs (Bai et al., 2021; Zhou et al., 2022). Our results in the photographic clinical imaging setting are well-aligned with previous findings in general computer vision tasks. Another important observation is the dependency of sensitivity and specificity on the type of image degradation, as well as the underlying medical image morphology and colorimetric properties. This highlights the critical importance of considering the visual and physical characteristics of the targeted tissue structures when designing an imaging device, especially when it is intended to use AI-assist diagnosis. as these factors can significantly influence final AUC and accuracy outcomes.

The results from the general-purpose image-text foundation models, CLIP and SigLIP, highlight the need for open-source foundation models specifically designed for photographic medical imaging tasks. While several foundation models have been developed for radiology (Azad et al., 2023; Ma et al., 2025) and pathology imaging (Wang et al., 2024), there is currently no unified foundation model tailored for photographic medical images. In particular, there is a lack of publicly available models that address site-specific needs in skin, oral, and endoscopic imaging.

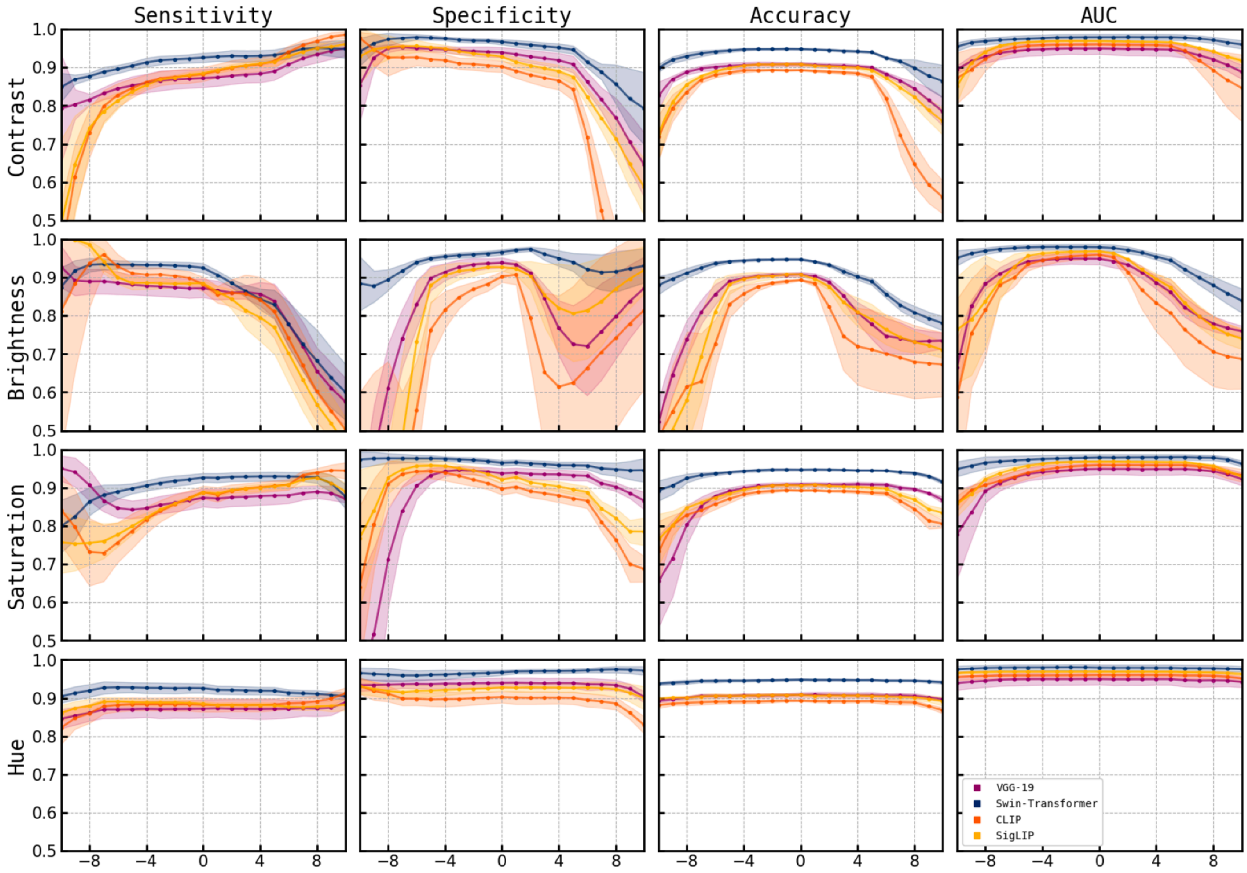


Fig. 12. Robustness evaluation of foundation models compared with VGG-19 and Swin Transformer. The same 4×4 grid of plots illustrates model performance across various degradations. Compared to results on the oral dataset, the foundation models narrowed the performance gap with VGG-19 but still underperformed relative to the Swin Transformer.

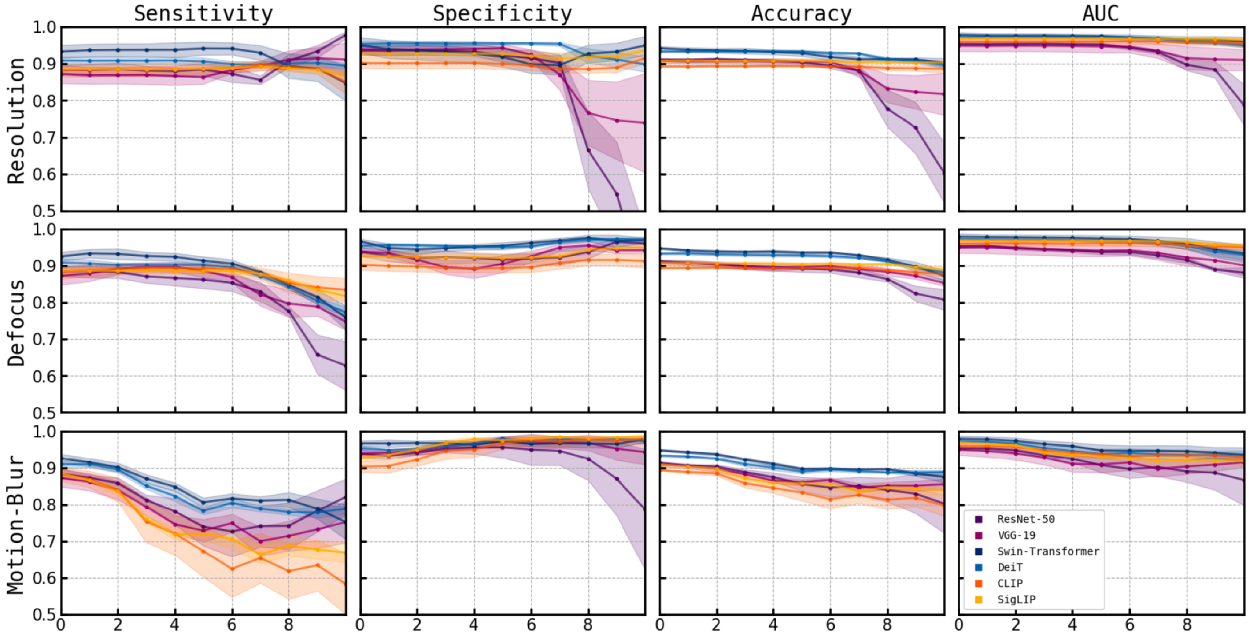


Fig. 13. Skin dataset robustness evaluation under capture-induced image degradations: resolution, defocus, and motion blur. In contrast to the oral imaging dataset, glare was not included here, as skin imaging typically does not suffer significantly from glare-related degradation due to the absence of highly reflective liquid environments. However, resolution remains critically important, as dermatologists typically require higher magnification and greater detail to accurately assess subtle skin features during diagnosis.

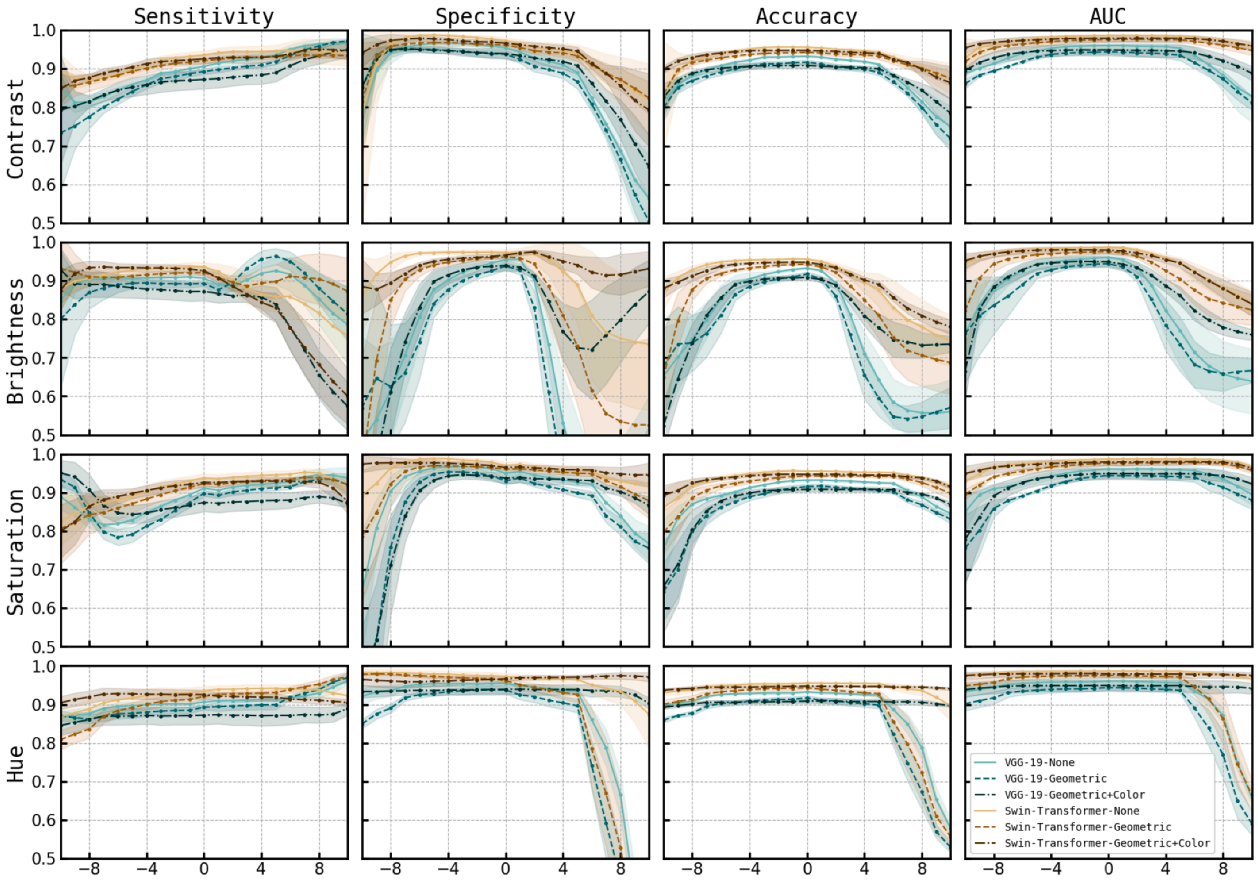


Fig. 14. Comparison of VGG-19 and Swin Transformer performance under different augmentation strategies on the skin dataset. Plots show results with no augmentation, geometric augmentation, and combined geometric and color augmentation. Darker lines indicate more extensive augmentation during the training phase.

Our findings underscore the importance of developing domain-specific photographic approaches to support diagnostic accuracy and robustness in these clinical applications.

In addition, we acknowledge that our findings may be limited by the datasets used in the experiments. Although intraoral and skin imaging make up a large portion of photographic medical imaging, other clinical environments exist, such as gastrointestinal and endoscopic imaging. For our own in house oral imaging data, the patient pool is further limited. Even so, our study provides a useful perspective on the robustness of models and device design when only small datasets are available, compared with the much larger skin dataset. The generalizability of these robustness conclusions can be strengthened through further analysis and by incorporating larger, more diverse photographic datasets, extending applicability to a broader range of clinical environments, including underrepresented settings. With the current data, we believe the study offers a comprehensive and reasonably general view for skin cancer screening and a useful baseline for the oral dataset we are building.

The primary goal of this paper is not to identify the best-performing neural network model for classifying oral or skin images. Instead, the motivation lies in systematically characterizing how image classification tasks, imaging devices, and potential degradations affect the final accuracy of the results. To provide a more comprehensive understanding, we investigate how AI-assisted photographic model choices and variations in image quality influence performance at different levels. All the image degradations that we introduced, as noted in the previous section, are common in standard clinical practice. By analyzing the robustness of different deep learning models, we are able to identify how external factors such as illumination conditions, camera settings, and user operation affect results. This supports practical guidelines for clinicians to adjust image quality to a level that maximizes AI assisted accuracy, including ensuring adequate illumination, choosing appropriate exposure times, and applying hue or white balance correction. These recommendations can be applied to our in house oral imaging device, to general skin imaging environments, and potentially to other photographic imaging systems with similar imaging acquisition workflow. These findings also serve as a reference, setting expectations for how each imaging factor can affect the final results during operations. For capture induced artifacts and optical aberrations, such as defocus, motion blur, and limited resolution, this robustness analysis can greatly benefit research on developing AI assisted medical devices by helping to make informed trade offs in optical system design, balancing between portability, affordability, and effectiveness. Our aim is to use these findings to inform the selection and training of effective classifiers, promote best practices in image acquisition, and guide the design of photographic imaging devices.

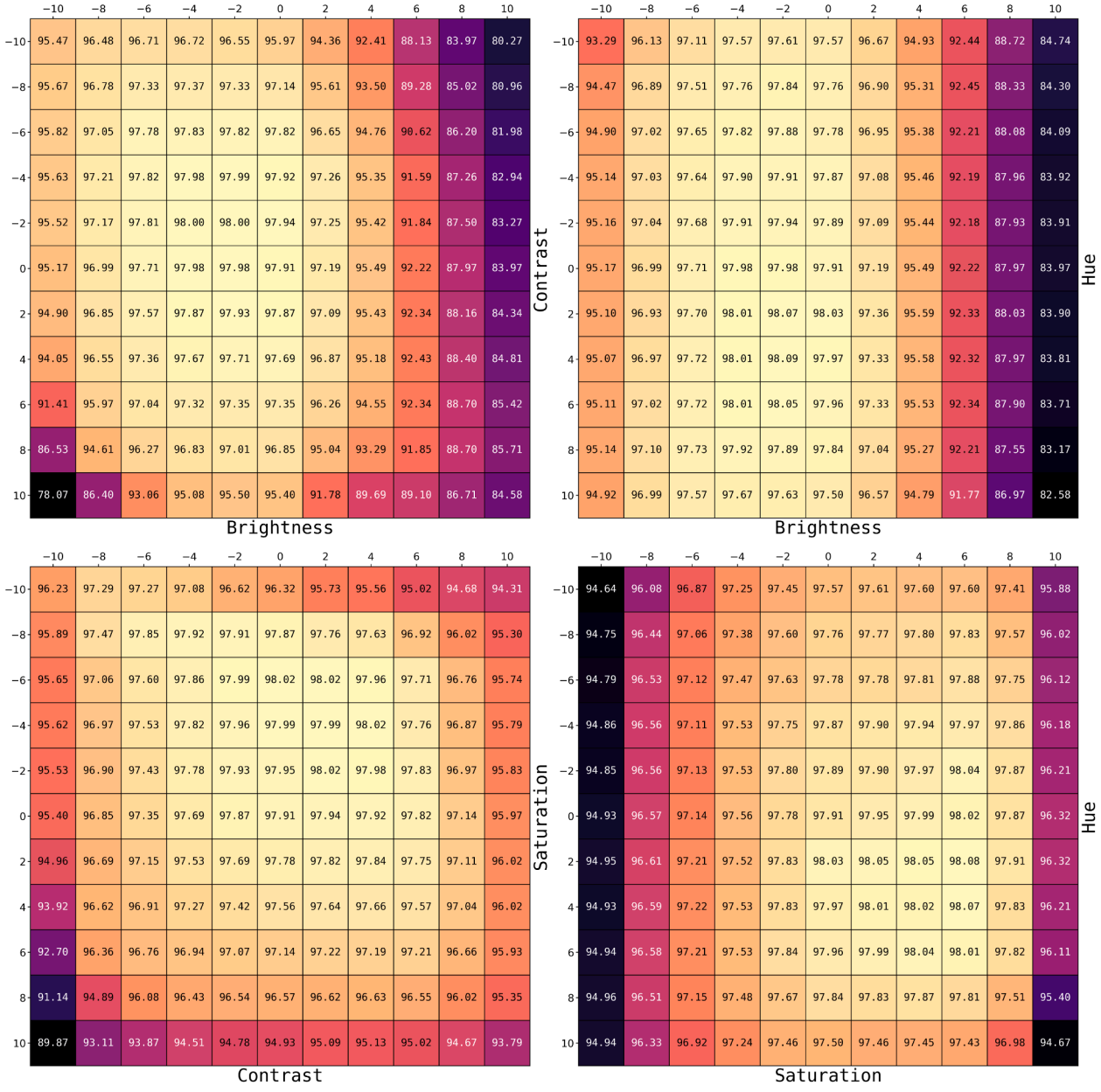


Fig. 15. Mean AUC of Swin Transformer under dual image degradations on the skin dataset. Each matrix shows the average AUC (%) over five-fold cross-validation for 11 severity levels (−10 to +10). The x-axis degradation is applied first, followed by the y-axis. Color maps are individually scaled, with bright yellow indicating the highest AUC and black the lowest. Corresponding standard deviations are shown in Fig. A.17.

6. Conclusion

In this study, we present a comprehensive and systematic robustness evaluation of AI-assisted photographic image classification using supervised deep learning. We investigated how various types of image degradation impact model performance, revealing several critical insights. Beyond overall correlations with accuracy and AUC, we found that sensitivity and specificity are also strongly affected by image degradations–particularly due to how different degradation types alter the visibility and color properties of diagnostic features. From an architectural perspective, recently developed transformer models consistently outperformed CNN-based models in most scenarios. However, foundation models such as CLIP and SigLIP, when used solely as image encoders, did not demonstrate the same level of robustness as CNNs or transformers in either the oral or skin imaging tasks. We believe this highlights both a current limitation and an important opportunity: there is a growing need for foundation models specifically designed and trained for photographic medical imaging tasks. We also show that, with well-designed augmentation strategies, robustness improves and variance narrows for both convolutional networks and attention-based networks. Most importantly, our systematic evaluation of how individual image degradations influence model performance provides a valuable benchmark and reference standard for future development. To conclude, Our robustness analysis shows that recently developed attention-based architectures achieve promising results on photographic medical imaging classification tasks. We also observe asymmetric effects on sensitivity and specificity across degradation types and severities, providing informative guidance for imaging-hardware design and data-augmentation strategies. Together, these findings inform the development of next-generation AI-assisted clinical imaging systems for photographic modalities, particularly in oral and dermatological diagnostics.

CRediT authorship contribution statement

Ruilin You: Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization; **Jiabin Chen:** Investigation, Visualization; **Yihan Wang:** Investigation, Formal analysis; **Zhihan Hong:** Methodology; **Cherie Wink:** Conceptualization, Data curation, Resources; **Petra Wilder-Smith:** Conceptualization, Methodology, Resources, Funding acquisition; **Rongguang Liang:** Supervision, Project administration, Funding acquisition, Writing – review & editing; **Bofan Song:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

Data availability

Besides the publicly available ISIC skin lesion image datasets, the oral image data is this research will not be made public. The code can be obtained upon request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by NIH (R01DE030682 National Institute of Dental and Craniofacial Research, R21CA274717 National Cancer Institute, U01CA279862 National Cancer Institute).

Appendix A. Supporting information

A.1. Oral and skin mixed-image degradation matrix standard deviation

Here, we present the standard deviation matrices for mixed image degradations on the oral and skin datasets, as discussed in Sections 4.1.3 and 4.2.3.

A.2. Tabular results of robustness performance across all models on oral and skin datasets

We organized the results into a numerical comparison between models and image degradations. The tables provide an accessible and accurate reference for assessing the robustness of different models under various image degradation conditions (Tables A.1, A.2, A.3, A.4, A.5, A.6, A.7 and A.8).

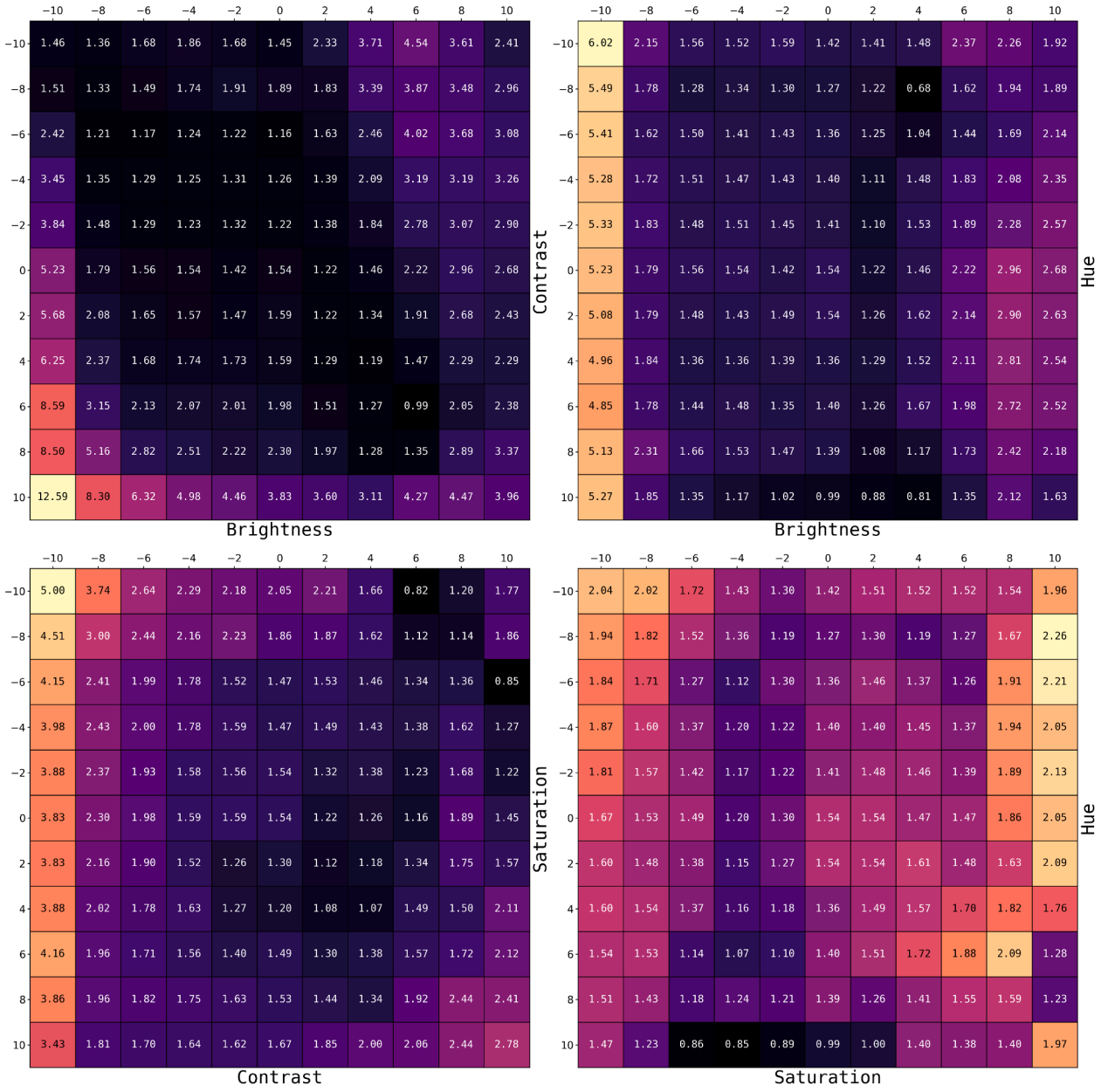


Fig. A.16. AUC standard deviation for swin transformer. Standard deviations from Fig. 9 show performance variability across five-fold cross-validation. Higher values indicate less consistency under specific degradation combinations.

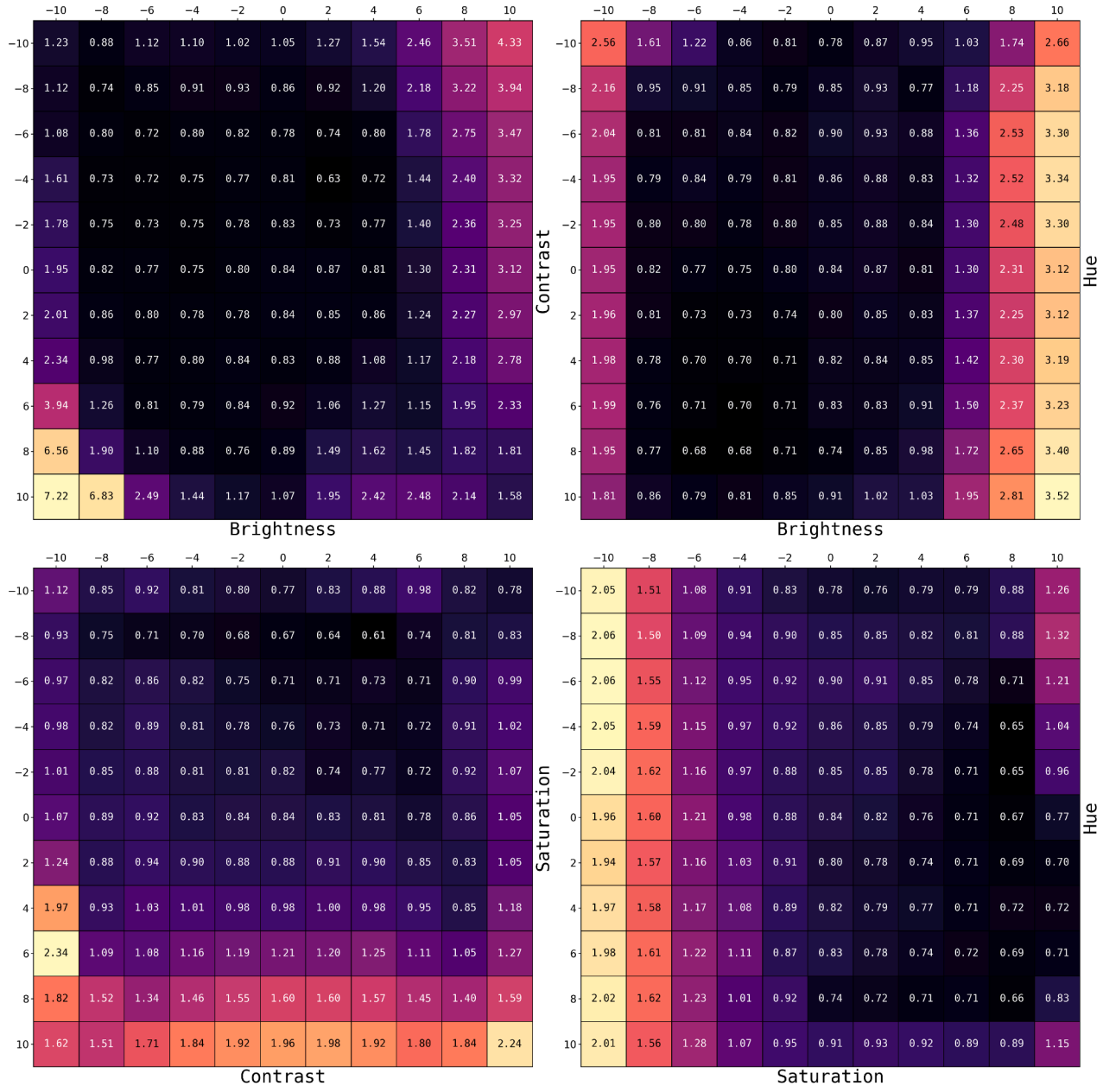


Fig. A.17. AUC standard deviation for Swin Transformer on the skin dataset. Standard deviations from Fig. 15 indicate performance variability across five-fold cross-validation.

Table A.1
Oral dataset robustness evaluation: Top-1 accuracy (mean ± standard deviation) under reference and degraded conditions. Bold text indicates the highest mean accuracy at each degradation level within each model category.

Model	Reference 0	Brightness −10	Brightness +10	Contrast −10	Contrast +10	Saturation −10	Saturation +10	Hue −10	Hue +10
CNNs									
ResNet50	0.82 ± 0.01	0.59 ± 0.01	0.58 ± 0.03	0.65 ± 0.05	0.66 ± 0.03	0.65 ± 0.06	0.70 ± 0.04	0.78 ± 0.01	0.79 ± 0.01
VGG19	0.82 ± 0.01	**0.73 ± 0.06**	0.58 ± 0.06	**0.70 ± 0.02**	0.73 ± 0.02	0.68 ± 0.05	0.75 ± 0.03	**0.79 ± 0.01**	**0.80 ± 0.02**
EfficientNet-B2	0.82 ± 0.01	0.64 ± 0.02	**0.65 ± 0.03**	**0.63 ± 0.03**	**0.75 ± 0.03**	**0.76 ± 0.02**	**0.75 ± 0.01**	**0.78 ± 0.03**	**0.80 ± 0.01**
DenseNet121	**0.83 ± 0.01**	0.58 ± 0.08	0.52 ± 0.05	0.69 ± 0.02	0.69 ± 0.03	0.73 ± 0.02	**0.76 ± 0.02**	0.78 ± 0.01	0.79 ± 0.01
Transformers									
Swin Transformer	**0.87 ± 0.01**	**0.77 ± 0.03**	**0.69 ± 0.01**	**0.79 ± 0.02**	0.74 ± 0.03	**0.79 ± 0.03**	**0.80 ± 0.02**	**0.83 ± 0.02**	**0.85 ± 0.01**
DeiT	0.84 ± 0.00	0.66 ± 0.05	0.63 ± 0.04	0.68 ± 0.03	0.64 ± 0.02	**0.79 ± 0.02**	0.75 ± 0.01	0.78 ± 0.01	0.80 ± 0.01
EfficientFormer-L3	0.85 ± 0.01	0.61 ± 0.02	0.65 ± 0.03	0.65 ± 0.03	**0.75 ± 0.01**	0.78 ± 0.03	0.76 ± 0.01	0.77 ± 0.02	0.79 ± 0.01
ViT	0.80 ± 0.02	0.55 ± 0.07	0.49 ± 0.05	0.60 ± 0.02	0.52 ± 0.07	0.55 ± 0.07	0.69 ± 0.02	0.70 ± 0.04	0.68 ± 0.03
Foundation Models									
CLIP	0.75 ± 0.02	0.53 ± 0.07	0.54 ± 0.02	0.57 ± 0.07	0.56 ± 0.03	0.49 ± 0.06	0.65 ± 0.02	0.70 ± 0.02	0.70 ± 0.02
SigLIP	**0.75 ± 0.02**	**0.57 ± 0.07**	**0.55 ± 0.03**	**0.59 ± 0.01**	**0.50 ± 0.01**	**0.52 ± 0.03**	**0.65 ± 0.02**	**0.72 ± 0.02**	**0.71 ± 0.01**

Table A.2
Oral dataset robustness evaluation: AUC (mean ± standard deviation) under reference and degraded conditions. Bold text indicates the highest mean accuracy at each degradation level within each model category.

Model	Reference 0	Brightness −10	Brightness +10	Contrast −10	Contrast +10	Saturation −10	Saturation +10	Hue −10	Hue +10
CNNs									
ResNet50	0.90 ± 0.01	0.58 ± 0.11	0.54 ± 0.06	0.76 ± 0.03	0.76 ± 0.04	0.81 ± 0.05	0.77 ± 0.04	0.86 ± 0.01	0.87 ± 0.01
VGG19	0.90 ± 0.01	**0.80 ± 0.08**	**0.74 ± 0.03**	**0.77 ± 0.01**	**0.85 ± 0.01**	**0.87 ± 0.02**	0.81 ± 0.03	**0.88 ± 0.01**	**0.88 ± 0.02**
EfficientNet-B2	0.89 ± 0.01	0.67 ± 0.04	0.69 ± 0.04	0.65 ± 0.03	0.83 ± 0.03	**0.87 ± 0.01**	**0.82 ± 0.02**	0.85 ± 0.02	0.87 ± 0.01
DenseNet121	**0.90 ± 0.01**	0.64 ± 0.06	0.63 ± 0.06	0.75 ± 0.04	0.83 ± 0.01	0.86 ± 0.01	**0.82 ± 0.02**	0.85 ± 0.02	0.86 ± 0.01
Transformers									
Swin Transformer	**0.92 ± 0.02**	**0.83 ± 0.05**	**0.79 ± 0.03**	**0.86 ± 0.04**	**0.85 ± 0.01**	**0.89 ± 0.02**	**0.86 ± 0.02**	**0.90 ± 0.01**	**0.90 ± 0.01**
DeiT	0.91 ± 0.01	0.76 ± 0.10	0.76 ± 0.03	0.79 ± 0.09	0.83 ± 0.05	0.89 ± 0.01	0.82 ± 0.01	0.86 ± 0.02	0.88 ± 0.02
EfficientFormer-L3	0.90 ± 0.01	0.57 ± 0.07	0.73 ± 0.03	0.70 ± 0.05	0.84 ± 0.01	0.87 ± 0.02	0.82 ± 0.01	0.81 ± 0.03	0.86 ± 0.02
ViT	0.87 ± 0.01	0.63 ± 0.05	0.62 ± 0.08	0.67 ± 0.04	0.73 ± 0.04	0.74 ± 0.04	0.76 ± 0.03	0.80 ± 0.02	0.77 ± 0.02
Foundation Models									
CLIP	0.81 ± 0.02	0.60 ± 0.04	0.63 ± 0.02	0.59 ± 0.05	**0.67 ± 0.02**	0.56 ± 0.05	0.68 ± 0.02	0.78 ± 0.02	0.78 ± 0.02
SigLIP	**0.81 ± 0.02**	**0.64 ± 0.03**	**0.63 ± 0.01**	**0.67 ± 0.03**	0.65 ± 0.03	**0.69 ± 0.05**	**0.70 ± 0.01**	**0.78 ± 0.02**	**0.78 ± 0.01**

Table A.3
Top-1 Accuracy (mean ± std) under reference and severity levels (+5, +10) for spatial degradations. Bold indicates the highest mean per degradation level within each model category.

Model	Reference	Glare		Defocus		Motion	
	0	+5	+10	+5	+10	+5	+10
CNNs							
ResNet50	0.82 ± 0.01	0.80 ± 0.02	0.69 ± 0.08	0.77 ± 0.02	0.59 ± 0.04	0.74 ± 0.03	0.68 ± 0.02
VGG19	**0.82 ± 0.01**	**0.81 ± 0.02**	**0.73 ± 0.02**	**0.80 ± 0.01**	**0.60 ± 0.03**	**0.76 ± 0.02**	**0.70 ± 0.02**
Transformers							
Swin Transformer	**0.87 ± 0.01**	**0.86 ± 0.01**	**0.84 ± 0.01**	**0.84 ± 0.01**	0.64 ± 0.03	**0.76 ± 0.02**	0.70 ± 0.03
DeiT	0.84 ± 0.00	0.84 ± 0.01	0.82 ± 0.00	**0.84 ± 0.01**	**0.65 ± 0.02**	0.75 ± 0.01	**0.71 ± 0.02**
ViT	0.80 ± 0.02	0.78 ± 0.01	0.74 ± 0.02	0.79 ± 0.02	0.64 ± 0.02	0.72 ± 0.01	0.68 ± 0.00
Foundation Models							
CLIP	0.75 ± 0.02	0.72 ± 0.02	0.63 ± 0.03	**0.74 ± 0.02**	0.69 ± 0.03	0.66 ± 0.01	**0.62 ± 0.01**
SigLIP	0.75 ± 0.02	**0.72 ± 0.01**	**0.66 ± 0.03**	**0.74 ± 0.02**	**0.72 ± 0.02**	**0.67 ± 0.01**	0.61 ± 0.03

Table A.4
AUC (mean ± std) under reference and severity levels (+5, +10) for spatial degradations. Bold indicates the highest mean per degradation level within each model category.

Model	Reference	Glare		Defocus		Motion	
	0	+5	+10	+5	+10	+5	+10
CNNs							
ResNet50	0.90 ± 0.01	**0.89 ± 0.01**	0.83 ± 0.03	0.86 ± 0.01	**0.63 ± 0.05**	0.81 ± 0.03	0.76 ± 0.03
VGG19	**0.90 ± 0.01**	0.87 ± 0.03	**0.80 ± 0.05**	**0.88 ± 0.02**	0.61 ± 0.05	**0.84 ± 0.02**	**0.77 ± 0.02**
Transformers							
Swin Transformer	**0.92 ± 0.02**	**0.91 ± 0.01**	**0.91 ± 0.02**	0.90 ± 0.01	0.68 ± 0.05	0.81 ± 0.05	0.75 ± 0.05
DeiT	0.91 ± 0.01	0.91 ± 0.01	0.90 ± 0.01	**0.91 ± 0.01**	0.71 ± 0.03	**0.84 ± 0.04**	**0.80 ± 0.05**
ViT	0.87 ± 0.01	0.86 ± 0.01	0.85 ± 0.01	0.87 ± 0.01	**0.73 ± 0.05**	0.81 ± 0.01	0.75 ± 0.01
Foundation Models							
CLIP	**0.81 ± 0.02**	**0.80 ± 0.02**	**0.76 ± 0.01**	**0.81 ± 0.02**	0.76 ± 0.02	**0.75 ± 0.03**	**0.72 ± 0.04**
SigLIP	**0.81 ± 0.02**	0.79 ± 0.02	**0.76 ± 0.03**	**0.81 ± 0.02**	**0.78 ± 0.02**	**0.75 ± 0.03**	**0.72 ± 0.06**

Table A.5
Skin dataset robustness evaluation: Top-1 accuracy (mean ± standard deviation) under reference and degraded conditions. Bold text indicates the highest mean accuracy at each degradation level within each model category.

Model	Reference	Brightness −10	Brightness +10	Contrast −10	Contrast +10	Saturation −10	Saturation +10	Hue −10	Hue +10
CNNs									
ResNet50	0.91 ± 0.00	0.61 ± 0.11	0.69 ± 0.04	0.74 ± 0.10	0.75 ± 0.03	0.75 ± 0.07	0.86 ± 0.01	0.90 ± 0.01	0.90 ± 0.01
VGG19	0.91 ± 0.01	0.52 ± 0.06	**0.74 ± 0.02**	**0.83 ± 0.04**	0.78 ± 0.04	0.66 ± 0.12	**0.87 ± 0.01**	0.89 ± 0.00	0.90 ± 0.00
EfficientNet-B2	0.92 ± 0.01	**0.71 ± 0.06**	0.72 ± 0.03	0.74 ± 0.05	**0.79 ± 0.02**	**0.88 ± 0.01**	0.89 ± 0.01	**0.91 ± 0.00**	**0.91 ± 0.00**
DenseNet121	**0.93 ± 0.01**	0.55 ± 0.12	0.63 ± 0.10	0.75 ± 0.10	0.72 ± 0.04	0.87 ± 0.01	0.88 ± 0.01	**0.91 ± 0.01**	**0.91 ± 0.01**
Transformers									
Swin Transformer	**0.95 ± 0.00**	**0.88 ± 0.02**	**0.78 ± 0.02**	**0.90 ± 0.01**	0.86 ± 0.04	**0.89 ± 0.03**	**0.91 ± 0.01**	**0.94 ± 0.00**	**0.94 ± 0.00**
DeiT	0.93 ± 0.00	0.86 ± 0.03	0.77 ± 0.02	0.87 ± 0.02	0.78 ± 0.05	0.85 ± 0.03	0.89 ± 0.01	0.92 ± 0.00	0.92 ± 0.01
EfficientFormer-L3	0.93 ± 0.00	0.70 ± 0.12	0.72 ± 0.08	0.69 ± 0.07	0.80 ± 0.02	0.78 ± 0.05	0.90 ± 0.01	0.92 ± 0.00	0.92 ± 0.01
ViT	0.92 ± 0.00	0.52 ± 0.09	**0.78 ± 0.02**	0.85 ± 0.01	0.67 ± 0.04	0.85 ± 0.03	0.87 ± 0.01	0.90 ± 0.01	0.91 ± 0.01
Foundation Models									
CLIP	0.89 ± 0.00	**0.48 ± 0.05**	0.67 ± 0.08	0.72 ± 0.06	0.56 ± 0.04	0.73 ± 0.08	0.81 ± 0.01	0.88 ± 0.01	0.87 ± 0.01
SigLIP	**0.91 ± 0.00**	0.46 ± 0.01	**0.71 ± 0.02**	**0.73 ± 0.07**	**0.76 ± 0.04**	**0.77 ± 0.07**	**0.83 ± 0.01**	**0.90 ± 0.00**	**0.89 ± 0.01**

Table A.6
Skin dataset robustness evaluation: AUC (mean ± standard deviation) under reference and degraded conditions. Bold text indicates the highest mean accuracy at each degradation level within each model category.

Model	Reference	Brightness −10	Brightness +10	Contrast −10	Contrast +10	Saturation −10	Saturation +10	Hue −10	Hue +10
CNNs									
ResNet50	0.95 ± 0.00	0.68 ± 0.13	**0.77 ± 0.03**	0.84 ± 0.06	0.87 ± 0.02	0.87 ± 0.03	0.93 ± 0.01	**0.95 ± 0.01**	0.95 ± 0.01
VGG19	0.95 ± 0.01	0.66 ± 0.07	0.76 ± 0.01	**0.90 ± 0.01**	0.89 ± 0.02	0.78 ± 0.12	0.92 ± 0.02	0.94 ± 0.02	0.94 ± 0.01
EfficientNet-B2	**0.96 ± 0.01**	**0.87 ± 0.03**	0.75 ± 0.02	0.85 ± 0.03	**0.90 ± 0.01**	**0.93 ± 0.01**	**0.95 ± 0.00**	**0.95 ± 0.01**	**0.96 ± 0.00**
DenseNet121	**0.96 ± 0.00**	0.82 ± 0.03	0.72 ± 0.11	0.89 ± 0.01	0.87 ± 0.02	0.92 ± 0.01	**0.95 ± 0.00**	**0.95 ± 0.01**	**0.96 ± 0.00**
Transformers									
Swin Transformer	**0.98 ± 0.01**	**0.95 ± 0.02**	0.84 ± 0.03	**0.95 ± 0.01**	**0.96 ± 0.01**	**0.95 ± 0.02**	**0.96 ± 0.01**	**0.97 ± 0.01**	**0.98 ± 0.01**
DeiT	0.97 ± 0.00	0.92 ± 0.02	0.83 ± 0.02	0.94 ± 0.01	0.93 ± 0.02	0.93 ± 0.01	0.96 ± 0.01	**0.97 ± 0.00**	0.97 ± 0.01
EfficientFormer-L3	0.97 ± 0.00	0.84 ± 0.07	0.81 ± 0.03	0.87 ± 0.02	0.89 ± 0.01	0.91 ± 0.03	0.95 ± 0.00	0.96 ± 0.01	0.96 ± 0.01
ViT	0.97 ± 0.01	0.91 ± 0.01	**0.85 ± 0.02**	0.92 ± 0.01	0.91 ± 0.03	0.93 ± 0.02	0.95 ± 0.01	0.96 ± 0.01	0.96 ± 0.01
Foundation Models									
CLIP	0.96 ± 0.01	0.59 ± 0.26	0.69 ± 0.08	**0.87 ± 0.07**	0.85 ± 0.09	**0.86 ± 0.02**	**0.93 ± 0.02**	0.95 ± 0.01	0.95 ± 0.01
SigLIP	**0.97 ± 0.00**	**0.76 ± 0.14**	**0.74 ± 0.03**	0.85 ± 0.04	**0.92 ± 0.01**	**0.86 ± 0.03**	**0.93 ± 0.00**	**0.97 ± 0.00**	**0.96 ± 0.00**

Table A.7
AUC (mean ± std) under reference and severity levels (+5, +10) for spatial degradations (Resolution, Defocus, and Motion). Bold indicates the highest mean per degradation level within each model category.

Model	Reference	Resolution		Defocus		Motion	
	0	+5	+10	+5	+10	+5	+10
CNNs							
ResNet50	0.95 ± 0.00	**0.95 ± 0.00**	0.79 ± 0.06	**0.94 ± 0.01**	0.88 ± 0.01	**0.91 ± 0.02**	0.87 ± 0.07
VGG19	**0.95 ± 0.02**	**0.95 ± 0.02**	**0.91 ± 0.03**	0.94 ± 0.02	**0.90 ± 0.03**	**0.91 ± 0.02**	**0.91 ± 0.01**
Transformers							
Swin Transformer	**0.98 ± 0.01**	**0.97 ± 0.01**	**0.96 ± 0.01**	**0.97 ± 0.01**	**0.93 ± 0.02**	**0.95 ± 0.02**	**0.94 ± 0.02**
DeiT	0.97 ± 0.00	0.97 ± 0.01	0.95 ± 0.01	**0.97 ± 0.01**	0.93 ± 0.01	0.94 ± 0.01	0.93 ± 0.01
Foundation Models							
CLIP	0.96 ± 0.01	0.96 ± 0.01	0.96 ± 0.01	0.96 ± 0.01	**0.95 ± 0.01**	0.94 ± 0.01	**0.93 ± 0.01**
SigLIP	**0.97 ± 0.00**	**0.97 ± 0.00**	**0.97 ± 0.00**	**0.97 ± 0.00**	**0.95 ± 0.00**	0.93 ± 0.01	0.92 ± 0.01

Table A.8
Accuracy (mean ± std) under reference and severity levels (+5, +10) for spatial degradations (Resolution, Defocus, and Motion). Bold indicates the highest mean per degradation level within each model category.

Model	Reference	Resolution		Defocus		Motion	
	0	+5	+10	+5	+10	+5	+10
CNNs							
ResNet50	**0.91 ± 0.00**	0.90 ± 0.01	0.60 ± 0.08	0.89 ± 0.01	0.81 ± 0.03	**0.86 ± 0.02**	0.80 ± 0.08
VGG19	**0.91 ± 0.01**	**0.91 ± 0.01**	**0.82 ± 0.06**	**0.90 ± 0.01**	**0.85 ± 0.01**	**0.86 ± 0.01**	**0.86 ± 0.01**
Transformers							
Swin Transformer	**0.94 ± 0.01**	**0.93 ± 0.01**	**0.90 ± 0.01**	**0.94 ± 0.00**	0.87 ± 0.01	**0.90 ± 0.00**	0.87 ± 0.02
DeiT	0.93 ± 0.00	**0.93 ± 0.00**	**0.90 ± 0.01**	0.93 ± 0.00	**0.88 ± 0.01**	0.89 ± 0.01	**0.89 ± 0.01**
Foundation Models							
CLIP	0.89 ± 0.00	0.89 ± 0.00	0.89 ± 0.01	0.89 ± 0.00	0.88 ± 0.01	0.83 ± 0.02	0.80 ± 0.03
SigLIP	**0.91 ± 0.00**	**0.91 ± 0.00**	**0.90 ± 0.00**	**0.90 ± 0.00**	**0.89 ± 0.01**	**0.86 ± 0.01**	0.84 ± 0.01

References

Anwar, S. M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., & Khan, M. K. (2018). Medical image analysis using convolutional neural networks: A review. *Journal of Medical Systems*, *42*, 1–13.

Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., & Merhof, D. (2023). Foundational models in medical imaging: A comprehensive survey and future vision. arXiv preprint arXiv:2310.18689

Azad, R., Kazerouni, A., Heidari, M., Aghdam, E. K., Molaei, A., Jia, Y., Jose, A., Roy, R., & Merhof, D. (2024). Advances in medical image analysis with vision transformers: A comprehensive review. *Medical Image Analysis*, *91*, 103000.

Bai, Y., Mei, J., Yuille, A. L., & Xie, C. (2021). Are transformers more robust than cnns? *Advances in Neural Information Processing Systems*, *34*, 26831–26843.

Bi, W. L., Hosny, A., Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A., Allison, T., Arnaout, O., Abbosh, C., Dunn, I. F. et al. (2019). Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA: A Cancer Journal for Clinicians*, *69*(2), 127–157.

Birur N, P., Song, B., Sunny, S. P., Mendonca, P., Mukhia, N., Li, S., Patrick, S., AR, S., Imchen, T., Leivon, S. T. et al. (2022). Field validation of deep learning based point-of-care device for early detection of oral malignant and potentially malignant disorders. *Scientific Reports*, *12*(1), 14283.

Boone, L., Biparva, M., Forooshani, P. M., Ramirez, J., Masellis, M., Bartha, R., Symons, S., Strother, S., Black, S. E., Heyn, C. et al. (2023). Rood-mri: Benchmarking the robustness of deep learning segmentation models to out-of-distribution and corrupted data in mri. *NeuroImage*, *278*, 120289.

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *74*(3), 229–263.

Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., & Yap, M. H. (2022). Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. *Medical Image Analysis*, *75*, 102305.

Chacko, N., & Ankri, R. (2024). Non-invasive early-stage cancer detection: Current methods and future perspectives. *Clinical and Experimental Medicine*, *25*(1), 17.

Chan, H.-P., Samala, R. K., Hadjiiski, L. M., & Zhou, C. (2020). Deep learning in medical image analysis. *Deep Learning in Medical Image Analysis: Challenges and Applications*, (pp. 3–21).

Cho, B.-J., Choi, Y. J., Lee, M.-J., Kim, J. H., Son, G.-H., Park, S.-H., Kim, H.-B., Joo, Y.-J., Cho, H.-Y., Kyung, M. S. et al. (2020). Classification of cervical neoplasms on colposcopic photography using deep learning. *Scientific Reports*, *10*(1), 13652.

Crosby, D., Bhatia, S., Brindle, K. M., Coussens, L. M., Dive, C., Emberton, M., Esener, S., Fitzgerald, R. C., Gambhir, S. S., Kuhn, P. et al. (2022). Early detection of cancer. *Science*, *375*(6586), eaay9040.

Dodge, S., & Karam, L. (2016). Understanding how image quality affects deep neural networks. In *2016 eighth international conference on quality of multimedia experience (qoMEX)* (pp. 1–6). IEEE.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2020). An image is worth 16 × 16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929

Von F, Z. (1934). Beugungstheorie des schneidenver-fahrens und seiner verbesserten form, der phasenkontrastmethode. *Physica*, *1*(7–12), 689–704.

Fitzgerald, R. C., Antoniou, A. C., Fruk, L., & Rosenfeld, N. (2022). The future of early cancer detection. *Nature Medicine*, *28*(4), 666–677.

Fu, Q., Chen, Y., Li, Z., Jing, Q., Hu, C., Liu, H., Bao, J., Hong, Y., Shi, T., Li, K. et al. (2020). A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: A retrospective study. *EClinicalMedicine*, *27* 100558.

Gonzalez, R. C. (2009). Digital image processing. Pearson Education India.

Goodman, J. W. (2005). Introduction to Fourier optics. Roberts and Company Publishers.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Herath, H., Herath, H., Madusanka, N., & Lee, B.-I. (2025). A systematic review of medical image quality assessment. *Journal of Imaging*, *11*(4), 100.

Hernández-Pérez, C., Combalia, M., Podlipnik, S., Codella, N. C. F., Rotemberg, V., Halpern, A. C., Reiter, O., Carrera, C., Barreiro, A., Helba, B. et al. (2024). Bcn20000: Dermoscopic lesions in the wild. *Scientific Data*, *11*(1), 641.

Hong, Z., Zhang, Z., You, R., Chen, J., Li, S., Wang, Y., Sun, Y., Song, B., Ji, Z., Loy, D. A. et al. (2025). Dual-head multi-photon polymerization 3D printing for parallel additive manufacturing organic/inorganic materials in optics. *Additive Manufacturing*, *103*, 104772.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

Islam, M. M., Alam, K. M. R., Uddin, J., Ashraf, I., & Samad, M. A. (2023). Benign and malignant oral lesion image classification using fine-tuned transfer learning techniques. *Diagnostics*, *13*(21), 3360.

Jaspers, T. J. M., Boers, T. G. W., Kusters, C. H. J., Jong, M. R., Jukema, J. B., de Groof, A. J., Bergman, J. J., de With, P. H. N., & van der Sommen, F. (2024). Robustness evaluation of deep neural networks for endoscopic image analysis: Insights and strategies. *Medical Image Analysis*, *94*, 103157.

Keerthi, G., Mukhia, N., Sunny, S. P., Song, B., Raghavan, S. A., Gurudath, S., Mendonca, P., Li, S., Patrick, S., Imchen, T. et al. (2023). Inter-observer agreement among specialists in the diagnosis of oral potentially malignant disorders and oral cancer using store-and-forward technology. *Clinical Oral Investigations*, *27*(12), 7575–7581.

Kusters, C. H. J., Jaspers, T. J. M., Boers, T. G. W., Jong, M. R., Jukema, J. B., Fockens, K. N., de Groof, A. J., Bergman, J. J., van der Sommen, F., & De With, P. H. N. (2025). Will transformers change gastrointestinal endoscopic image analysis? A comparative analysis between CNNs and transformers, in terms of performance, robustness and generalization. *Medical Image Analysis*, *99*, 103348.

Lepcha, D. C., Goyal, B., Dogra, A., Sharma, K. P., & Gupta, D. N. (2023). A deep journey into image enhancement: A survey of current and emerging trends. *Information Fusion*, *93*, 36–76.

Li, Y., Yuan, G., Wen, Y., Hu, J., Evangelidis, G., Tulyakov, S., Wang, Y., & Ren, J. (2022). Efficientformer: Vision transformers at mobilenet speed. *Advances in Neural Information Processing Systems*, *35*, 12934–12949.

Lin, I., Datta, M., Laronde, D. M., Rosin, M. P., & Chan, B. (2021). Intraoral photography recommendations for remote risk assessment and monitoring of oral mucosal lesions.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022).

Liu, Z., Zhang, J., He, Q., & Wang, C. (2024). Understanding data augmentation from a robustness perspective. In *ICASSP 2024–2024 ieee international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6760–6764). IEEE.

Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift fuer Medizinische Physik*, *29*(2), 102–127.

Ma, D., Pang, J., Gotway, M. B., & Liang, J. (2025). A fully open AI foundation model applied to chest radiography. *Nature*, *643*(pp. 1–11).

Marka, A., Carter, J. B., Toto, E., & Hassanpour, S. (2019). Automated detection of nonmelanoma skin cancer using digital images: A systematic review. *BMC Medical Imaging*, *19*, 1–12.

Maron, R. C., Haggenmüller, S., von Kalle, C., Utikal, J. S., Meier, F., Gellrich, F. F., Hauschild, A., French, L. E., Schlaak, M., Ghoreschi, K. et al. (2021a). Robustness of convolutional neural networks in recognition of pigmented skin lesions. *European Journal of Cancer*, *145*, 81–91.

Maron, R. C., Schlager, J. G., Haggenmüller, S., von Kalle, C., Utikal, J. S., Meier, F., Gellrich, F. F., Hobelsberger, S., Hauschild, A., French, L. et al. (2021b). A benchmark for neural network robustness in skin cancer classification. *European Journal of Cancer*, *155*, 191–199.

Martínez-Martínez, J., Brown, O., Karami, M., & Nabavi, S. (2025). Robust training with data augmentation for medical imaging classification. arXiv preprint arXiv:2506.17133

Mirfendereski, P., Li, G. Y., Pearson, A. T., & Kerr, A. R. (2025). Artificial intelligence and the diagnosis of oral cavity cancer and oral potentially malignant disorders from clinical photographs: A narrative review. *Frontiers in Oral Health*, *6*, 1569567.

Narayanamurthy, V., Padmapriya, P., Noorasafrin, A., Pooja, B., Hema, K., Nithyakalyani, K., Samsuri, F. et al. (2018). Skin cancer detection using non-invasive techniques. *RSC Advances*, *8*(49), 28095–28130.

Panayides, A. S., Amini, A., Filipovic, N. D., Sharma, A., Tsaftaris, S. A., Young, A., Foran, D., Do, N., Golemati, S., Kurc, T. et al. (2020). Ai in medical imaging informatics: Current challenges and future directions. *IEEE Journal of Biomedical and Health Informatics*, *24*(7), 1837–1857.

Pei, Y., Huang, Y., Zou, Q., Zhang, X., & Wang, S. (2019). Effects of image degradation and degradation removal to CNN-based image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(4), 1239–1253.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PmLR.

Rajpurkar, P., & Lungren, M. P. (2023). The current and future state of AI interpretation of medical images. *New England Journal of Medicine*, *388*(21), 1981–1990.

Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2023). Transformers in medical imaging: A survey. *Medical Image Analysis*, *88*, 102802.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

Song, B., & Liang, R. (2025). Integrating artificial intelligence with smartphone-based imaging for cancer detection in vivo. *Biosensors and Bioelectronics*, *271*, 116982.

Song, B., Sunny, S., Li, S., Gurushanth, K., Mendonca, P., Mukhia, N., Patrick, S., Gurudath, S., Raghavan, S., Imchen, T. et al. (2021). Mobile-based oral cancer classification for point-of-care screening. *Journal of Biomedical Optics*, *26*(6), 065003.

Song, B., Sunny, S., Uthoff, R. D., Patrick, S., Suresh, A., Kolur, T., Keerthi, G., Anbarani, A., Wilder-Smith, P., Kuriakose, M. A. et al. (2018). Automatic classification of dual-modalilty, smartphone-based oral dysplasia and malignancy images using deep learning. *Biomedical Optics Express*, *9*(11), 5318–5329.

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114). PMLR.

Tanriver, G., Soluk Tekkesin, M., & Ergen, O. (2021). Automated detection and classification of oral lesions using deep learning to detect oral potentially malignant disorders. *Cancers*, *13*(11), 2766.

Thomas, L., & Puig, S. (2017). Dermoscopy, digital dermoscopy and other diagnostic tools in the early detection of melanoma and follow-up of high-risk skin cancer patients. *Acta Dermato-Venereologica*, *97*, 14–21.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning* (pp. 10347–10357). PMLR.

Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, *5*(1), 1–9.

Uthoff, R. D., Song, B., Birur, P., Kuriakose, M. A., Sunny, S., Suresh, A., Patrick, S., Anbarani, A., Spires, O., Wilder-Smith, P. et al. (2018a). Development of a dual-modality, dual-view smartphone-based imaging system for oral cancer detection. In *Design and quality for biomedical technologies XI* (pp. 106–112). SPIE (*vol. 10486*).

Uthoff, R. D., Song, B., Sunny, S., Patrick, S., Suresh, A., Kolur, T., Gurushanth, K., Wooten, K., Gupta, V., Platek, M. E. et al. (2019). Small form factor, flexible, dual-modality handheld probe for smartphone-based, point-of-care oral and oropharyngeal cancer screening. *Journal of Biomedical Optics*, *24*(10), 106003.

Uthoff, R. D., Song, B., Sunny, S., Patrick, S., Suresh, A., Kolur, T., Keerthi, G., Spires, O., Anbarani, A., Wilder-Smith, P. et al. (2018b). Point-of-care, smartphone-based, dual-modality, dual-view, oral cancer screening device with neural network classification for low-resource communities. *PLoS One*, *13*(12), e0207493.

Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., Li, R., Tang, H., Wang, K., Li, Y. et al. (2024). A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, *634*(8035), 970–978.

Warin, K., Limprasert, W., Suebnukarn, S., Jinaporntham, S., & Jantana, P. (2021). Automatic classification and detection of oral cancer in photographic images using deep learning algorithms. *Journal of Oral Pathology & Medicine*, *50*(9), 911–918.

Welikala, R. A., Remagnino, P., Lim, J. H., Chan, C. S., Rajendran, S., Kallarakkal, T. G., Zain, R. B., Jayasinghe, R. D., Rimal, J., Kerr, A. R. et al. (2020). Automated detection and classification of oral lesions using deep learning for early detection of oral cancer. *IEEE Access*, *8*, 132677–132693.

Wen, D., Khan, S. M., Xu, A. J., Ibrahim, H., Smith, L., Caballero, J., Zepeda, L., de Blas Perez, C., Denniston, A. K., Liu, X. et al. (2022). Characteristics of publicly available skin cancer image datasets: A systematic review. *The Lancet Digital Health*, *4*(1), e64–e74.

Yang, S. M., Song, B., Wink, C., Abouakl, M., Takesh, T., Hurlbutt, M., Dinica, D., Davis, A., Liang, R., & Wilder-Smith, P. (2023). Performance of automated oral cancer screening algorithm in tobacco users vs. non-tobacco users. *Applied Sciences*, *13*(5), 3370.

You, R., Chen, J., Wang, Y., Zhang, Z., Sun, Y., Song, B., Loy, D. A., Ji, Z., Hong, Z., & Liang, R. (2025a). Cost-effective motionless color coded ptychography for high-throughput biomedical imaging. *ACS Photonics*, *12*(6), 3236–3244.

You, R., Hong, Z., Chen, J., Zhang, Z., Wang, Y., Sun, Y., Song, B., Loy, D. A., Ji, Z., & Liang, R. (2025b). Extremely compact 3d printed glass ternary diffractive optical element for holographic images. *Advanced Optical Materials*, *13* (p. 2501074).

Zhai, X., Mustafa, B., Kolesnikov, A., & Beyer, L. (2023). Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 11975–11986).

Zhang, S., & Metaxas, D. (2024). On the challenges and perspectives of foundation models for medical image analysis. *Medical Image Analysis*, *91*, 102996.

Zhou, D., Yu, Z., Xie, E., Xiao, C., Anandkumar, A., Feng, J., & Alvarez, J. M. (2022). Understanding the robustness in vision transformers. In *International conference on machine learning* (pp. 27378–27394). PMLR.