AMERICAN ACADEMY OF OPHTHALMOLOGY®

Improving Artificial Intelligence–based Microbial Keratitis Screening Tools Constrained by Limited Data Using Synthetic Generation of Slit-Lamp Photos

Daniel Wang, BA,[a] *Bonnie Sklar, MD,*[b] *James Tian, MD,*[c] *Rami Gabriel, MD,*[d] *Matthew Engelhard, MD, PhD,*[e] *Ryan P. McNabb, PhD,*[d] *Anthony N. Kuo, MD*[d,f]

Objective: We developed a novel slit-lamp photography (SLP) generative adversarial network (GAN) model using limited data to supplement and improve the performance of an artificial intelligence (AI)–based microbial keratitis (MK) screening model.

Design: Cross-sectional study.

Subjects: Slit-lamp photographs of 67 healthy and 36 MK eyes were prospectively and retrospectively collected at a tertiary care ophthalmology clinic at a large academic institution.

Methods: We trained the GAN model StyleGAN2-ADA on healthy and MK SLPs to generate synthetic images. To assess synthetic image quality, we performed a visual Turing test. Three cornea fellows tested their ability to identify 20 images each of (1) real healthy, (2) real diseased, (3) synthetic healthy, and (4) synthetic diseased. We also used Kernel Inception Distance (KID) to quantitatively measure realism and variation of synthetic images. Using the same dataset used to train the GAN model, we trained 2 DenseNet121 AI models to grade SLP images as healthy or MK with (1) only real images and (2) real supplemented with GAN-generated images.

Main Outcome Measures: Classification performance of MK screening models trained with only real images compared to a model trained with both limited real and supplemented synthetic GAN images.

Results: For the visual Turing test, the fellows on average rated synthetic images as good quality (83.3% ± 12.0% of images), and synthetic and real images were found to depict pertinent anatomy and pathology for accurate classification (96.3% ± 2.19% of images). These experts could distinguish between real and synthetic images (accuracy: 92.5% ± 9.01%). Analysis of KID score for synthetic images indicated realism and variation. The MK screening model trained on both limited real and supplemented synthetic data (area under the receiver–operator characteristic curve: 0.93, bootstrapping 95% CI: 0.77–1.0) outperformed the model trained with only real data (area under the receiver–operator characteristic curve: 0.76, 95% CI: 0.50–1.0), with an improvement of 0.17 (95% CI: 0–0.4; 2-tailed *t* test $P = 0.076$).

Conclusions: Artificial intelligence–based MK classification may be improved by supplementation of limited real training data with synthetic data generated by GANs.

Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2025;5:100676 © 2024 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Supplemental material available at www.ophthalmologyscience.org.

Artificial intelligence (AI) has shown to be a promising tool for screening eye conditions like diabetic retinopathy in nonophthalmology settings.[1,2] Artificial intelligence models require large amounts of data to train. However, many conditions, like microbial keratitis (MK)—despite being a top 5 cause of blindness worldwide—have limited availability of public data for training.[3,4] Early diagnosis of MK is important because early treatment can limit the extent of corneal damage from infection. Diagnosis can be particularly difficult when specialized expertise may not be readily available. Slit-lamp microscopy is the most common ophthalmic examination used to diagnose and monitor the progression of MK. However, slit-lamp photography (SLP) comprises only 1% of globally available public datasets, which hampers training for AI-based MK screening tools.[5]

In settings of limited data availability, generative adversarial networks (GANs) have been shown to be effective in creating synthetic data for data augmentation. This technique has been used to create augmented datasets that were then used to train models for detecting skin lesions, chest x-rays, and inherited retinal diseases.[6–8] Prior literature on the

https://doi.org/10.1016/j.xops.2024.100676
ISSN 2666-9145/25

use of GANs in generating synthetic ophthalmic images has focused on retinal fundus and OCT images.[9,10] The resulting synthetic images have been found to be indiscernible from real images and potentially useful in dataset augmentation for deep learning models.[11,12]

Early GAN models required substantial amounts of data to train. Small datasets often lead to unstable training and mode collapse, a failure in training that results in the model being fixed on a limited set of data features, ignoring the full diversity of the dataset.[13] Attempts to augment the data frequently resulted in leaking of augmentations into the final generated images. However, more recent GANs—including StyleGAN2-ADA—use novel image augmentation strategies to train on limited data.[14] This allows GAN training with smaller datasets (hundreds instead of tens of thousands of images) to create synthetic data to supplement limited datasets.

We present a novel synthetic SLP generative model created with a dataset extremely limited by image availability. Additionally, we test the hypothesis that training deep learning models with augmentation from GAN-generated synthetic images improves classification performance.

Methods

Data Collection

Normal and MK SLPs were prospectively and retrospectively collected at a tertiary care ophthalmology clinic at a large academic institution. Prospective patients were recruited from February 2024 to March 2024. Retrospective patient data were collected from clinical encounters of patients seen from August 2023 to December 2023. This study was approved by the Institutional Review Board of Duke Health and followed the tenets of the Declaration of Helsinki and the Health Insurance Portability and Accountability Act. Informed consent was obtained through a signed written form for each prospectively recruited patient. The requirement of informed consent was waived for retrospective data review. Inclusion criteria for prospective patients included age of $\geq$18 years, ability to independently provide consent, and diagnosis of MK as confirmed by a clinician (for the MK population) or not diagnosed with MK by a clinician (for the normal population). Exclusion criteria for prospective patients included if the patient was unable to be imaged for any reason. For retrospective patients, inclusion criteria included age of $\geq$18 years, a clinical diagnosis of MK as documented in the patient's chart (for the MK population), or not diagnosed with MK (for the normal population) as documented in the patient's chart, and exclusion criteria included not having SLP. Patients were also excluded if they had other corneal or external examination findings besides MK (e.g., pterygium, LASIK flap, and corneal scarring). All external photography images used for this study were taken at 10$\times$ magnification with diffuse white light. No fluorescein was instilled prior to photography. Normal eyes collected for this study included patients who presented for routine ophthalmology visits, like diabetic retinopathy annual screening; patients with treated disease, such as squamous cell carcinoma of the conjunctiva who presented for follow-up surveillance without abnormal external examination findings; and patients with unilateral eye disease who had the opposite unaffected normal eye photographed as a baseline. Clinical notes and eye examination results were reviewed by one of the authors (D.W.) to ensure that the above criteria were met and that MK eyes had ongoing active MK at the time of photography. Past ocular history was reviewed to ensure patients did not have any additional diagnoses that would impact the cornea or surrounding structures. Patient charts were reviewed for demographic characteristics such as sex, age, race and ethnicity, and eye laterality.

From the total image dataset, we withheld 10 normal images and 9 diseased images as an independent test set for the classification model testing at the end. Neither the GAN nor the classification models were trained on this test set.

Generative Adversarial Network (Synthetic Image) Model

We trained StyleGAN2-ADA with a limited SLP dataset of normal images (n = 27) and diseased images (n = 57). Training images were first preprocessed by manually cropping a square region surrounding the cornea and resizing each image to 1024 $\times$ 1024 pixels. Model parameters were initialized to values learned from training (i.e., transfer learning) on the Flickr-Faces-HQ dataset (found at https://github.com/NVlabs/ffhq-dataset) with x-flip image augmentation.[15] The model was trained to generate 1024 $\times$ 1024 synthetic images. Details of image augmentation and GAN architecture are discussed in previous literature introducing StyleGAN2-ADA.[16]

Kernel Inception Distance (KID), a widely adopted metric of GAN performance, was used to assess training.[17] This metric was designed to be suited for smaller datasets. A smaller KID indicates that the latent features between real images and synthetic images are more similar. Kernel Inception Distance was measured using model checkpoints for every 200 epochs. The models were trained for 800 epochs, and the model with the lowest KID was used to generate 200 images each of normal and diseased eyes using a list of random seeds.

Visual Turing Test

We conducted a visual Turing test with cornea fellows to quantify the quality and inclusion of normal anatomy and pathology in synthetic images, as well as to understand how well experts distinguished between real and synthetic images. These experiments were modeled on those previously conducted GAN-generated synthetic images to evaluate image quality.[12] The synthetic images used in these experiments were generated by a list of random seeds and were not curated prior to rating. The fellows have completed an ophthalmology residency and were completing specialized subspecialty training in cornea and external disease. In these visual Turing tests, the cornea fellows performed manual annotation of 80 images, which included equal numbers of real and synthetic images as well as normal and diseased images. Fellows were initially not told that the dataset included synthetic images to prevent any evaluation of whether the image was real or synthetic from biasing their decision about the quality of the image and whether the image included pertinent anatomy and pathology. During the first round of the test, the fellows were given 2 prompts and asked to review all 80 images:

- Prompt 1: "Is the image quality sufficient for grading?" This question was used to assess the quality of the synthetic images and determine if they contained artifacts that rendered them ungradable.
- Prompt 2: "What is the diagnosis? Is the image depicting an eye that is normal or an eye with MK?" This question was used to establish the diagnostic capability of the cornea fellows and to assess whether the images contained the

necessary features of normal anatomy or MK pathology to allow human graders to correctly diagnose the image.

After reviewing the entire dataset once for the prompts above, the fellows were then asked the following prompt and reviewed all 80 images a second time:

- Prompt 3: "For each image, is the image real or synthetic?" The fellows were told that the dataset contained both real and synthetic images but were not told what percentage were real or synthetic to avoid bias. We utilized this question to assess an expert's ability to differentiate between real and synthetic images of the cornea.

Microbial Keratitis Classification/Screening Model

Previous studies of AI-based MK classification models have shown that DenseNet121 is superior to other classic deep learning algorithms such as Inception-v3 and ResNet50 in differentiating between keratitis, normal, and other corneal pathology in slit-lamp images of eyes. Li et al[16] showed that DenseNet121 had a higher area under the receiver–operator characteristic (AUROC) curve, and the t-distributed stochastic neighbor embedding indicated that the features of each category learned by the DenseNet121 algorithm were more separable than those of other models. Therefore, we utilized DenseNet121 in training our classification model. DenseNet121 was trained with PyTorch (version 1.6.0). We used the following weights and values for our model: adaptive moment estimation optimizer was set to a 0.001 initial learning rate, β1 of 0.9, β2 of 0.999, and weight decay of 1e-4. Each model was trained for 80 epochs. Standard image augmentation techniques such as random cropping, horizontal and vertical flipping, and rotations were applied to the images in preprocessing. After precedents established in previous literature, validation loss during training was calculated on the validation dataset after each epoch.[16] A checkpoint was saved each time validation loss decreased. The model with the lowest validation loss during the entirety of training was saved as the final model for use on the test set.

For our first model (only real), we trained DenseNet121 with only real data (n = 84 images). The real dataset of preprocessed images was split into training (70%) and validation (30%) sets. Our second model (real + synthetic) was trained with the same real data used to train the first model with the addition of synthetic data; this augmented training set comprised 400 synthetic images with a 1:1 normal:diseased ratio. The validation set for this second model was composed of all the real data (except for the images withheld for the independent test set).

Area under the receiver–operator characteristic curve was calculated using the scikit-learn Python package by comparing DenseNet121's predicted probability that the image depicted an eye with MK to the true labels (normal vs. MK). Accuracy, sensitivity, and specificity were calculated after assigning a predicted label to each image based on whether the probability of normal or MK was higher. Our entire pipeline of StyleGAN2-ADA and DenseNet121 training is depicted in the Supplement (Fig S1, available at www.ophthalmologyscience.org).

Testing the 2 Models

Each model (first model: only real; second model: real + synthetic) was tested by applying it to a test set that was withheld from both the GAN model training and the classification model training. For each image, the model's prediction that the image was normal and the prediction that the image was diseased were recorded. The AUROC was calculated for the entire test set using the predicted probability that the image was diseased.

We then used bootstrapping to generate CIs for our MK DenseNet121 AUROC measurements.[18] Each test set was resampled with replacement 1000 times to construct 95% CIs for each AUROC based on empirical quantiles (2.5th to 97.5th percentile) among all 1000 resamples. A CI for the difference in AUROCs between the first model trained with only real data and the second model trained with real and GAN-generated data was similarly calculated based on empirical quantiles of this difference across all resamples. A 2-sided *P* value to determine whether the difference in AUROC between models was statistically significant was calculated as the proportion of resamples in which this difference was ≥97.5th percentile and ≤2.5th percentile.

Results

Data Collected from Clinic

We collected SLPs from 67 normal eyes of 51 patients and 36 diseased eyes from 35 patients to be included in the total dataset. Of these, 4 normal eyes from 2 patients and 1 diseased eye from 1 patient were collected prospectively. The remaining photographs were collected retrospectively. Demographic details of the dataset used in training and validation of the models are described in Table 1, and details of the dataset reserved for testing of the models are shown in Table 2. The characteristics of the test set were similar to the corresponding training and validation set. Microbial keratitis eyes with laboratory-confirmed etiology (n = 15) included patients with viral (n = 3), fungal (n = 5), bacterial (n = 5), and parasitic (n = 2) etiologies of keratitis. The remaining MK eyes (n = 14) were clinically documented and treated as MK.

Normal eye images were more common in our dataset than MK eye images. The average age of patients with normal eyes was slightly higher than the average age of patients with MK eyes. The dataset of normal eyes had a

Table 1. Training and Validation Set Demographic Characteristics of Eyes

Variables	Normal	MK
No. of eyes (patients)	57 (42)	27 (26)
Age, mean ± SD	63.7 ± 16.2	51.5 ± 15.3
Female, n (%)	20 (47.6)	16 (61.5)
Race/ethnicity, n (%)		
White	35 (83.3)	17 (65.4)
Black	4 (9.52)	5 (19.2)
Other	3 (7.14)	4 (15.4)
Laterality, n (%)		
OD	28 (49.1)	17 (63.0)
OS	29 (50.9)	10 (37.0)

MK = microbial keratitis; OD = right eye; OS = left eye; SD = standard deviation.
Demographic characteristics of the eyes included in the training of StyleGAN2-ADA as well as DenseNet121 Microbial Keratitis model. Data were collected prospectively and retrospectively under institutional review board–approved protocol from a large, academic institution. Patients included those who presented for ocular surface disease examination, ophthalmology consultation, and routine ophthalmology evaluations.

Table 2. Test Set Demographic Characteristics of Eyes

Variables	Normal	MK
No. of eyes (patients)	10 (9)	9 (9)
Age, mean ± SD	63.2 ± 14.8	55.9 ± 22.0
Female, n (%)	5 (55.6)	4 (44.4)
Race/ethnicity, n (%)		
White	8 (88.9)	5 (55.6)
Black	1 (11.1)	4 (44.4)
Other	0 (0)	0 (0)
Laterality, n (%)		
OD	8 (80.0)	5 (55.6)
OS	2 (20.0)	4 (44.4)

MK = microbial keratitis; OD = right eye; OS = left eye; SD = standard deviation.
Demographic characteristics of the eyes included in the test set for final DenseNet121 Microbial Keratitis model testing. These data were not included in the training or validation set of either StyleGAN2-ADA or DenseNet121 model development. Data were collected prospectively and retrospectively under institutional review board–approved protocol from a large academic institution. Patients included those who presented for ocular surface disease examination, ophthalmology consultation, and routine ophthalmology evaluations.

higher percentage of patients who were White and a lower percentage of patients who were Black compared to the MK dataset. Examples of collected images that depict normal and MK eyes are shown in Figure 2 as the Real Healthy and Real Diseased examples.

GAN Model Evaluation

Kernel Inception Distance curves during training of the normal and diseased models are shown in Figure 3. Kernel Inception Distance decreased during training on images of normal and MK eyes and remained low. This indicates that the synthetic images generated by the GAN during training contained similar latent features as the dataset the model was training on, validating realism and variety of synthetic images generated by the GAN. Examples of synthetic images generated by the trained GANs that depict normal and MK eyes are shown in Figure 2 as the Synthetic Healthy and Synthetic Diseased examples.

Visual Turing Test Results

Table 3 displays the results from the visual Turing test. On average (± standard deviation), experts rated 94.2% (± 2.36%) of real images and 83.3% (± 12.0%) of synthetic images as high enough quality to rate for anatomy and pathology. The average accuracy of classifying images as healthy and diseased did not vary significantly by real (95.8% ± 2.89%) or synthetic (96.7% ± 3.82%). Experts performed well on rating images as real or synthetic (accuracy: 92.5% ± 9.01%, sensitivity: 98.3% ± 1.44%, and specificity: 86.7% ± 17.0%). The Fleiss free-marginal kappa for real and synthetic ratings was 0.72 (95% CI: 0.60–0.84), indicating intermediate to good interrater agreement.

Testing the 2 Models

First model: Training DenseNet121 on only the real dataset alone yielded an AUROC of 0.76 (Fig 4A) with bootstrapping 95% CI: 0.50 to 1.0 (Fig 5). The accuracy of this MK DenseNet121 model trained only on real data was 84.2%, sensitivity was 88.9%, and specificity was 80%.

Second model: Augmenting the MK DenseNet121 model with synthetic data training yielded an AUROC of 0.93 with bootstrapping 95% CI: 0.77 to 1.0 (Fig 4B and 5). The accuracy of this MK DenseNet121 model trained on real and synthetic data was 94.7%, sensitivity was 88.9%, and specificity was 100%.

There was a 0.17 improvement in AUROC of the model trained with real and synthetic data over the model trained with only real data. The 95% CI of improvement in AUROC was 0 to 0.4. The 2-tailed t test yielded P-value of 0.076.

Discussion

StyleGAN2-ADA successfully generated synthetic images of normal and MK eyes from a small, limited dataset. Our study demonstrated that these images were high quality for grading and correctly depicted anatomy and pathology that reflected normal and MK eyes. Cornea fellows were able to distinguish between real and synthetic images. Importantly, we demonstrated that augmenting the training of DenseNet121 with synthetic images yielded an improved AUROC for our test set. Bootstrapping suggested that the model trained with real and synthetic data statistically performed better than random, but the difference in AUROC between the model trained with only real data and the model trained with real and synthetic data does include 0 at the boundary of the 95% CI.

StyleGAN2-ADA has been shown to generate realistic-looking images using training datasets with hundreds of images. Our normal eye dataset was comprised of 57 images, and the MK eye dataset was comprised of 27 images. Given these small dataset sizes, the synthetic image quality could be affected. When reviewed in the visual Turing test, the fellows noted specific features outside of the cornea that led them to identify an image as synthetic. Fellows noted repeating textures in the iris, conjunctiva, or eyelid in the synthetic images, as can be noted in the generated images displayed in Figure 2 (e.g., where the blue of the iris extends into the upper eyelid in the bottom right Synthetic Healthy image). Although the synthetic images did not pass the third stage of the visual Turing test in their ability to fool corneal experts, the synthetic images did pass the first and second stages of the visual Turing test because the fellows rated the images as overall high quality and capable of depicting normal corneal anatomy and MK in the cornea.

In terms of DenseNet121's performance on the test set, training the model with synthetic data led to a decrease in the model's false-positive rate. After synthetic images were added to the training, the model was able to classify all normal images in the test set as normal. This improvement in the performance of the model gives a signal that synthetic images may contain information that is helpful to model

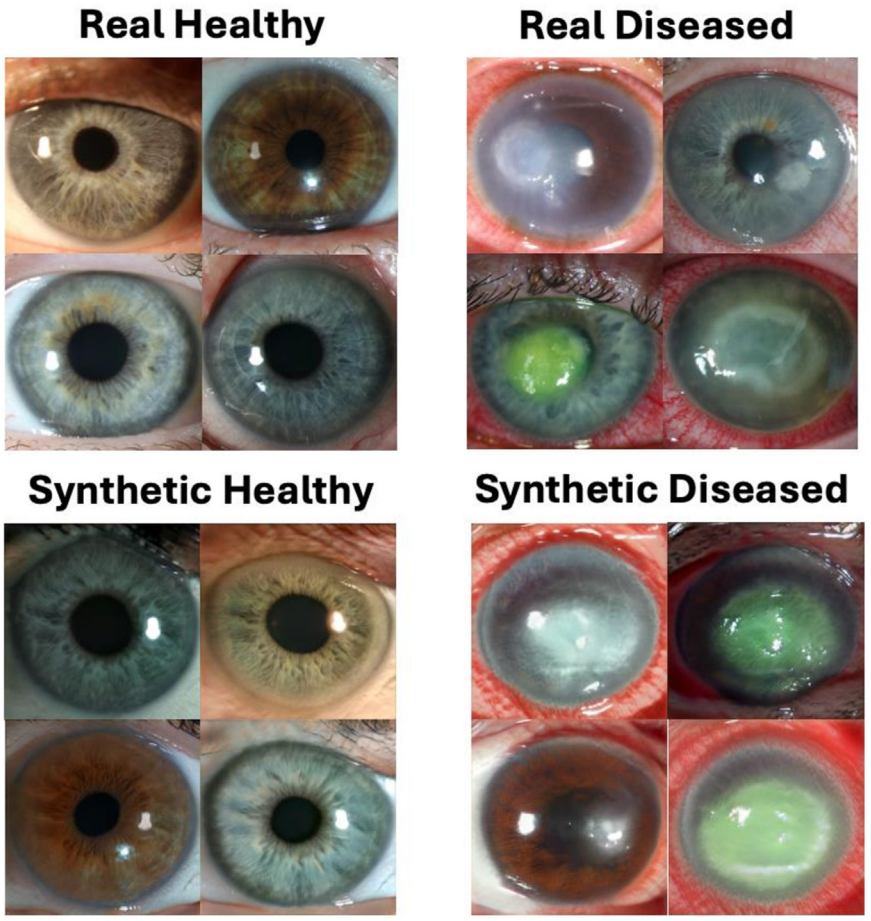

Figure 2. Synthetic images were generated by StyleGAN2-ADA using random seeds.

learning. Additionally, by training the model with synthetic images only in the training set and then testing the model with only real images in the validation set, we attempt to mitigate learning of any potential artifacts from the synthetic images that were seen by human expert reviewers in the visual Turing test.

In the setting of a limited test set size, we utilized bootstrapping to assess the potential range of performance in a generalized setting. Checking the robustness of our AUROC findings with bootstrapping, we found that the CI for the model trained only on real data contained 0.5, whereas the model trained on real and synthetic data did not. This suggests that the model trained on real and synthetic data performed better than random, whereas the model trained on only real data may not be performing better than random. In terms of the difference between the models, we find a P value of 0.076 with our 2-tailed t test, and the upper bound of the 95% CI is 0. Although bootstrapping assists in assessing potential imbalances in the test set that would limit generalizability, the performance of this model on a larger test set is a critical next step when more datasets become available.

In previous studies, DenseNet121 has been shown to perform well in separating categories of normal cornea and

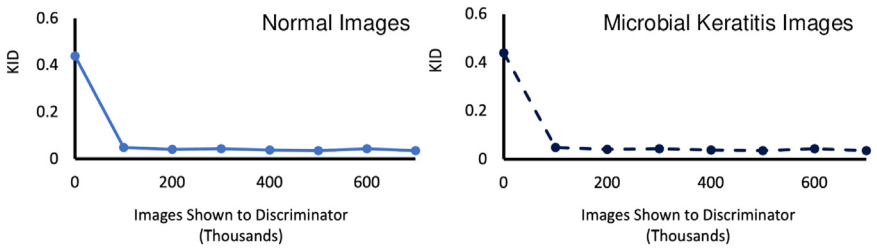

Figure 3. Kernel Inception Distance measured during the training of StyleGAN2-ADA. Lower values indicate a more similar distribution between training data and synthetically generated images.

Table 3. Results of Turing Test

Metrics	Fellow 1	Fellow 2	Fellow 3
Good quality (%)			
Real	92.5	97.5	92.5
Synthetic	77.5	100	72.5
Healthy/diseased ratings (includes real and synthetic)			
Accuracy (%)	98.8	95	95
Sensitivity (%), healthy	97.5	97.5	92.5
Specificity (%), diseased	100	92.5	97.5
Real/synthetic ratings			
Accuracy (%)	82.5	100	95
Sensitivity (%), real	97.5	100	97.5
Specificity (%), synthetic	67.5	100	92.5

Experts were presented with 80 images and underwent 3 experiments. The dataset for rating contained an equal distribution of real healthy, health diseased, synthetic healthy, and synthetic diseased. Experts were not told the distribution of the dataset. Experts were not told that the dataset contained synthetic images until they were asked to rate images as real or synthetic.
Fliess free-marginal kappa for real and synthetic ratings (95% confidence interval) was 0.72 (0.60–0.84), indicating intermediate to good agreement.

keratitis in SLP when trained on a large dataset of 13 557 images (6055 keratitis, 2777 cornea with other abnormalities, and 4725 normal cornea) from 7988 individuals. The AUROC achieved using this large, publicly unavailable dataset was 0.998 (95% CI, 0.996–0.999).[16] In our study, we utilized a dataset that was almost 1000 times smaller and achieved a similar AUROC using synthetic data augmentation of model training (0.93). Hu et al[19] utilized a dataset of 2757 slip-lamp images that included normal cornea, viral keratitis, fungal keratitis, and bacterial keratitis. The authors found that of 6 different deep learning algorithms, EficientNetV2-M showed the best classification ability with an AUROC of 0.85. Tiwari et al[20] trained the VGG-16 architecture for convolutional neural network (CNN) on 1313 photos of corneal ulcers and 1132 photos of corneal scars and tested the CNN on 2 different patient populations: those in eye clinics in India and the Byers Eye Institute at Stanford University. The AUROC for Indian patients was 0.97, and the AUROC for patients in Northern California was 0.95. These findings show that GAN-generated synthetic image augmentation of CNN training yields an AUROC that is comparable to training on much larger datasets.

Previous studies that utilized StyleGAN2-ADA to generate synthetic data with limited data utilized larger datasets (few hundred to few thousand).[14] Our study adds to

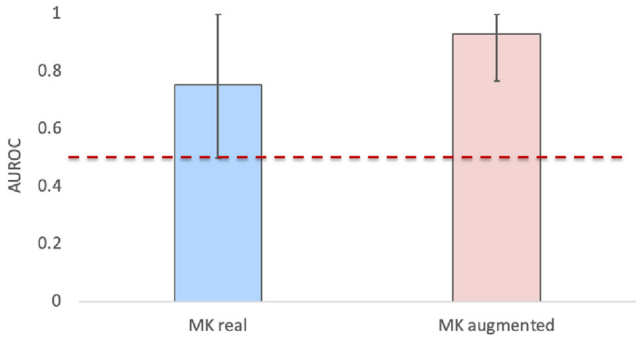

Figure 5. The dotted red line marks 0.5, the value at which the classification is as good as random.

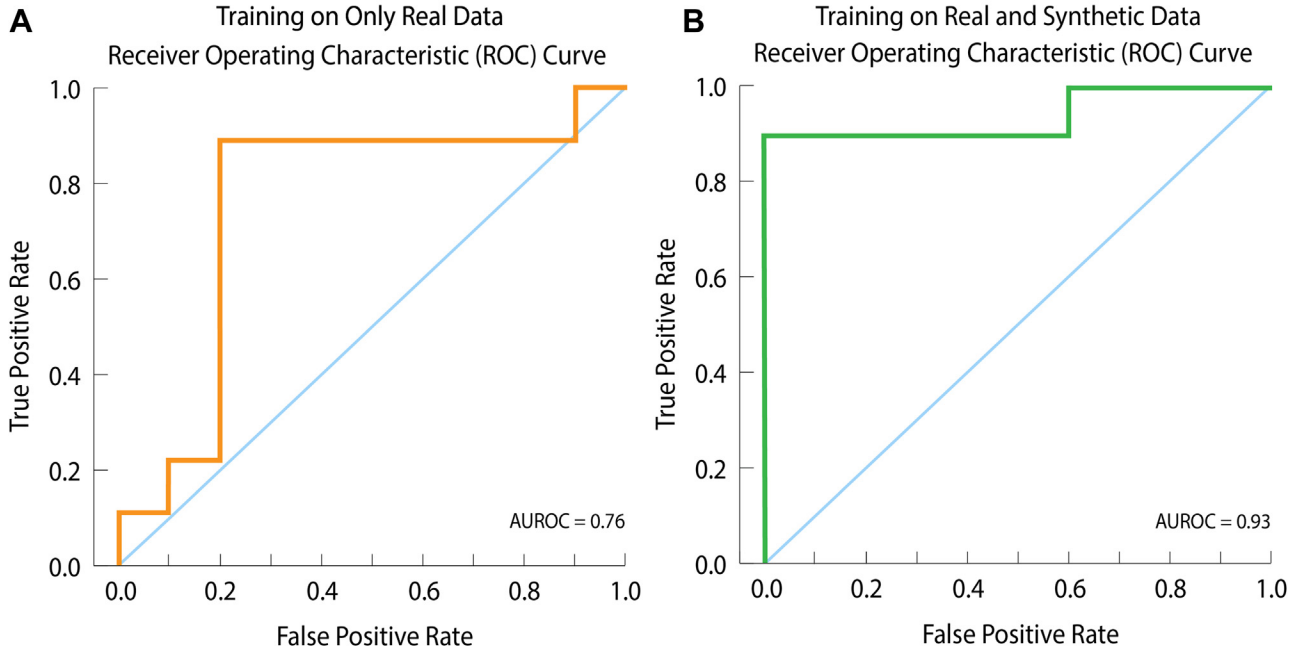

Figure 4. **A,** The AUROC curve for DenseNet121 trained with only real data: 0.76. **B,** The AUROC curve for DenseNet121 with augmented training on synthetic and real data: 0.93. AUROC = area under the receiver–operator characteristic.

this body of work by showing that this GAN is able to train on even more limited datasets (27 and 57 images). This shows that the novel data augmentations developed by Karras et al can augment GAN training to allow stable development of synthetic images with extremely limited datasets. Additionally, no studies have utilized GANs to generate synthetic data for external eye pathology using SLP. Prior studies have focused on generating synthetic OCT, angiography, and fundus images, with the large majority utilizing images of the retina. Previous studies utilizing GANs to generate synthetic images of the cornea have created topographical images and corneal nerve segmentation maps of confocal microscopy images, but none have applied GANs to SLP.[21,22] Additionally, though we focused specifically on MK, other groups such as Li et al[16] have shown that DenseNet121 was able to distinguish between MK and other causes of keratitis; although this is promising, the ability of data augmentation to aid in distinguishing among different forms of keratitis remains an open area for research.

Burlina et al[12] have previously shown that training CNNs with only synthetic data generated by GAN has a modest decrease in performance compared to CNN trained on only real data. In that study, the authors placed synthetic data in both the training and validation sets. Synthetic data in the validation set could be concerning as the model may overtrain on artifacts found in synthetic images and worsen performance on real data. In contrast, we utilized only real data in our validation set, and our resulting synthetic image-augmented CNN was found to have improved AUROC compared to CNNs trained on only real data. Additionally, Burlina et al utilized over 100 000 age-related macular degeneration fundus photographs from the Age-Related Eye Disease Study and applied a progressive growing of GAN architecture for generating synthetic images. We utilize a more sophisticated GAN architecture specifically designed to have good performance on smaller datasets and a much smaller dataset of SLPs.

Another previous study by Burlina et al[9] found that using a synthetic data approach to rebalance a dataset of fundus photographs of diabetic retinopathy improved parity in the accuracy of diagnosing images of patients with darker pigmentation. This was done by increasing the number of fundus images with darker pigmentation in the dataset. In this way, the authors showed synthetic data may improve classification performance by amplifying certain underrepresented features in the data distribution. This could explain how the application of GAN-generated synthetic data augmentation in our study improved DenseNet121 performance. By learning features in the real data distribution and creating representative images with different combinations of features, synthetic data provide more opportunities for classification models to learn important features of anatomy and pathology that assist in differentiating between image classes.

This study was limited by the availability of SLPs. Because patients do not routinely receive SLPs for normal eyes or for MK, images found on chart review may depict eyes that have more severe disease or other pathology. Images of normal eyes were collected by selecting images that were taken of the opposite eye as a baseline when 1 eye contained pathology or eyes with disease that had been treated and were being imaged for surveillance with no abnormal external examination findings. Physical examination findings in the patient chart were reviewed to ensure that the corneal surface and surrounding conjunctiva were documented as anatomically normal. This study was constrained in scope to training a classification model that would serve as a triaging tool for general care settings. The model was only trained to differentiate between normal eyes and eyes with any kind of MK. This type of tool would be envisioned for use in a nonsubspecialty setting. In the setting of a small dataset, another limitation is the need for bootstrapping. Bootstrapping provides a robustness check on our results, but generalizability may be limited without the ability to measure performance on an external dataset. The classification model may have limits to generalizability given the imbalance of the datasets, as the normal images were slightly older and had a larger proportion of White patients. Additionally, both the model trained on only real data and the model trained on real and synthetic data misclassified the same 1 MK image as normal. This image depicted a peripheral ulcer. In reviewing the training and validation set, there were no images that contained peripheral ulcers. This shows that while synthetic data may be able to assist the model in learning features in the dataset, features outside of the distribution may still be unknown to the model.

In this study, we showed that StyleGAN2-ADA was able to utilize extremely limited datasets (27 and 57 images, respectively) to generate synthetic SLPs of MK and normal eyes that were assessed to be good quality for grading and contained pertinent anatomy and pathology. We then demonstrated how utilizing these synthetic images in the training set for an AI-based MK screening model resulted in a model that had improved AUROC performance. These results demonstrate the potential of the use of GANs to create improved classification tools for settings with extremely limited data availability.

Footnotes and Disclosures

Originally received: August 7, 2024.
Final revision: November 21, 2024.
Accepted: December 6, 2024.
Available online: December 20, 2024. Manuscript no. XOPS-D-24-00287.

[a] Duke University School of Medicine, Durham, North Carolina.

[b] Ophthalmology, University of North Carolina School of Medicine, Chapel Hill, North Carolina.

[c] Omni Eye Specialists, Fort Collins, Colorado.

[d] Ophthalmology, Duke University School of Medicine, Durham, North Carolina.

[e] Biostatistics & Bioinformatics, Duke University, Durham, North Carolina.

[f] Biomedical Engineering, Duke University, Durham, North Carolina.

Presented at the 2024 AAO Annual Meeting Poster Discussion.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The authors made the following disclosures:

R.P.M.: Grants − DoD grants involving ophthalmic imaging (no overlap with this paper); Research contract − Johnson & Johnson Vision about choroidal imaging, which was paid to Duke University (no overlap with this paper); Royalties − Leica Microsystems for software IP related to OCT image processing and analysis licensed by Duke University to Leica Microsystems; Patents − ophthalmic imaging, but none related to microbial keratitis.

R.G.: Receipt of equipment − Alcon Vision LLC, food and bev: 135.20; ABBVIE INC., food and bev: 132.91.

A.N.K.: Grants − DoD grants involving ophthalmic imaging (no overlap with this paper); Research contract − Johnson & Johnson Vision about choroidal imaging, which was paid to Duke University (no overlap with this paper); Royalties − Leica Microsystems for software IP related to OCT image processing and analysis licensed by Duke University to Leica Microsystems; Patents − ophthalmic imaging, but none related to microbial keratitis; Travel expenses − Michigan Society of Eye Physicians and Surgeons.

B.S.: Receipt of equipment − Glaukos Corporation, food and bev: 139.98, Sight Sciences, Inc., food and bev: 127.55, Dompe US, Inc., food and bev: 115.05, Johnson & Johnson Surgical Vision, Inc., food and bev: 108.51, Oyster Point Pharma, Inc., food and bev: 29.19, Sight Sciences, Inc., food and bev: 15.65.

This work was supported by NIH P30EY005722, NIH R01-EY035534, NIMH K01MH127309, and a Research to Prevent Blindness Unrestricted Grant to the Duke Eye Center. The sponsors or funding organizations had no role in the design or conduct of this research.

HUMAN SUBJECTS: Human subjects were included in this study. This study was approved by the Institutional Review Board of Duke Health and followed the tenets of the Declaration of Helsinki and the Health Insurance Portability and Accountability Act. Informed consent was obtained through a signed written form for each prospectively recruited patient. The requirement of informed consent was waived for retrospective data review.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Wang, Sklar, Tian, Gabriel, Engelhard, McNabb, Kuo

Data collection: Wang, Sklar, Tian, Gabriel, Engelhard, McNabb, Kuo

Analysis and interpretation: Wang, Sklar, Tian, Gabriel, Engelhard, McNabb, Kuo

Obtained funding: N/A. Study was performed as part of regular employment duties at Duke University. No additional funding was provided.

Overall responsibility: Wang, Sklar, Tian, Gabriel, Engelhard, McNabb, Kuo

Support for Open Access publication was provided by Duke University.

Abbreviations and Acronyms:
AI = artificial intelligence; **AUROC** = area under the receiver−operator characteristic; **CI** = confidence interval; **CNN** = convolutional neural network; **GAN** = generative adversarial network; **MK** = microbial keratitis; **SLP** = slit-lamp photography.

Keywords:
Artificial intelligence, Microbial keratitis, Generative adversarial network, Slit-lamp photography, Cornea.

Correspondence:
Daniel Wang, BA, Duke Eye Center, DUMC Box 3802, Durham, NC 27710. E-mail: dw325@duke.edu.

References

1. Abràmoff MD, Lavin PT, Birch M, et al. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med*. 2018;1:39.
2. Rajesh AE, Davidson OQ, Lee CS, Lee AY. Artificial intelligence and diabetic retinopathy: AI framework, prospective studies, head-to-head validation, and cost-effectiveness. *Diabetes Care*. 2023;46:1728−1739.
3. Flaxman SR, Bourne RRA, Resnikoff S, et al. Global causes of blindness and distance vision impairment 1990-2020: a systematic review and meta-analysis. *Lancet Glob Health*. 2017;5:e1221−e1234.
4. Cabrera-Aguas M, Watson SL. Updates in diagnostic imaging for infectious keratitis: a review. *Diagnostics*. 2023;13:3358.
5. Khan SM, Liu X, Nath S, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit Health*. 2021;3:e51−e66.
6. Gao C, Killeen BD, Hu Y, et al. Synthetic data accelerates the development of generalizable learning-based algorithms for X-ray image analysis. *Nat Mach Intell*. 2023;5:294−308.
7. Chen RJ, Lu MY, Chen TY, et al. Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Eng*. 2021;5:493−497.
8. Veturi YA, Woof W, Lazebnik T, et al. SynthEye: investigating the impact of synthetic data on artificial intelligence-assisted gene diagnosis of inherited retinal disease. *Ophthalmol Sci*. 2023;3:100258.
9. Burlina P, Joshi N, Paul W, et al. Addressing artificial intelligence bias in retinal diagnostics. *Transl Vis Sci Technol*. 2021;10:13.
10. Zheng C, Xie X, Zhou K, et al. Assessment of generative adversarial networks model for synthetic optical coherence tomography images of retinal disorders. *Transl Vis Sci Technol*. 2020;9:29.
11. Chen JS, Coyner AS, Chan RVP, et al. Deepfakes in ophthalmology. *Ophthalmol Sci*. 2021;1:100079.
12. Burlina PM, Joshi N, Pacheco KD, et al. Assessment of deep generative models for high-resolution synthetic retinal image generation of age-related macular degeneration. *JAMA Ophthalmol*. 2019;137:258−264.
13. Kossale Y, Airaj M, Darouichi A. Mode collapse in generative adversarial networks: an overview. In: *2022 8th International Conference on Optimization and Applications (ICOA)*. Marrakesh, Morocco: IEEE; 2022:1−6.

14. Karras T, Aittala M, Hellsten J, et al. *Training Generative Adversarial Networks with Limited Data*. *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada*. 2020.
15. Hosna A, Merry E, Gyalmo J, et al. Transfer learning: a friendly introduction. *J Big Data*. 2022;9:102.
16. Li Z, Jiang J, Chen K, et al. Preventing corneal blindness caused by keratitis using artificial intelligence. *Nat Commun*. 2021;12:3738.
17. Bińkowski M, Sutherland DJ, Arbel M, Gretton A. *Demystifying MMD GANs*. *The Sixth International Conference on Learning Representations (ICLR 2018), Vancouver, Canada*. 2021.
18. Efron B. Bootstrap methods: another look at the jackknife. *Ann Stat*. 1979;7:1−26.
19. Hu S, Sun Y, Li J, et al. Automatic diagnosis of infectious keratitis based on slit lamp images analysis. *J Pers Med*. 2023;13:519.
20. Tiwari M, Piech C, Baitemirova M, et al. Differentiation of active corneal infections from healed scars using deep learning. *Ophthalmology*. 2022;129:139−146.
21. Jameel SK, Aydin S, Ghaeb NH, et al. Exploiting the generative adversarial network approach to create a synthetic topography corneal image. *Biomolecules*. 2022;12:1888.
22. Yıldız E, Arslan AT, Yıldız Taş A, et al. Generative adversarial network based automatic segmentation of corneal subbasal nerves on in vivo confocal microscopy images. *Transl Vis Sci Technol*. 2021;10:33.