Title

Enhancing Meibography Image Analysis Through Artificial Intelligence-Driven Quantification and Standardization for Dry Eye Research.

Journal

Translational Vision Science & Technology, 13(6)

Authors

Yeh, Chun-Hsiao
Graham, Andrew
Yu, Stella
et al.

Publication Date

2024-06-03

DOI

10.1167/tvst.13.6.16

Peer reviewed

tvst

Artificial Intelligence

Enhancing Meibography Image Analysis Through Artificial Intelligence–Driven Quantification and Standardization for Dry Eye Research

Chun-Hsiao Yeh[1–3], Andrew D. Graham[1,3], Stella X. Yu[2,4], and Meng C. Lin[1,3]

[1] Vision Science Group, Herbert Wertheim School of Optometry and Vision Science, University of California, Berkeley, Berkeley, CA, USA
[2] Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA, USA
[3] Clinical Research Center, Herbert Wertheim School of Optometry and Vision Science, University of California, Berkeley, Berkeley, CA, USA
[4] Computer Science and Engineering, University of Michigan, Ann Arbor, MI, USA

Correspondence: Meng C. Lin, Vision Science Group, Herbert Wertheim School of Optometry and Vision Science, University of California, Berkeley, Berkeley, CA 94720, USA.
e-mail: mlin@berkeley.edu

Received: December 11, 2023
Accepted: May 15, 2024
Published: June 21, 2024

Keywords: artificial intelligence; deep learning; dry eye; eyelid detection; tarsal plate segmentation; meibography; Meibomian gland dysfunction; unsupervised feature learning

Citation: Yeh CH, Graham AD, Yu SX, Lin MC. Enhancing meibography image analysis through artificial intelligence–driven quantification and standardization for dry eye research. Transl Vis Sci Technol. 2024;13(6):16, https://doi.org/10.1167/tvst.13.6.16

Purpose: This study enhances Meibomian gland (MG) infrared image analysis in dry eye (DE) research through artificial intelligence (AI). It is comprised of two main stages: automated eyelid detection and tarsal plate segmentation to standardize meibography image analysis. The goal is to address limitations of existing assessment methods, bridge the curated and real-world dataset gap, and standardize MG image analysis.

Methods: The approach involves a two-stage process: automated eyelid detection and tarsal plate segmentation. In the first stage, an AI model trained on curated data identifies relevant eyelid areas in non-curated datasets. The second stage refines the eyelid area in meibography images, enabling precise comparisons between normal and DE subjects. This approach also includes specular reflection removal and tarsal plate mask refinement.

Results: The methodology achieved a promising instance-wise accuracy of 80.8% for distinguishing meibography images from 399 DE and 235 non-DE subjects. By integrating diverse datasets and refining the area of interest, this approach enhances meibography feature extraction accuracy. Dimension reduction through Uniform Manifold Approximation and Projection (UMAP) allows feature visualization, revealing distinct clusters for DE and non-DE phenotypes.

Conclusions: The AI-driven methodology presented here quantifies and classifies meibography image features and standardizes the analysis process. By bootstrapping the model from curated datasets, this methodology addresses real-world dataset challenges to enhance the accuracy of meibography image feature extraction.

Translational Relevance: The study presents a standardized method for meibography image analysis. This method could serve as a valuable tool in facilitating more targeted investigations into MG characteristics.

Introduction

Meibography, an infrared imaging technique for visualizing the Meibomian glands (MGs) within the eyelids,[1,2] has gained prominence due to its relevance in the study of dry eye (DE), a prevalent ocular condition affecting numerous individuals worldwide.[3] DE is often tied to MG dysfunction (MGD).[4] The MGs in the eyelids secrete a lipid-rich meibum that covers the aqueous tears in a thin film and inhibits evaporation.[5] When these glands fail to secrete a sufficient amount of meibum or meibum with suboptimal biophysical properties, tear film instability and breakup can occur, ultimately leading to the symptoms of DE.[6,7] Manual assessment of meibography images to visually estimate the area of gland atrophy[8] has long been the standard and virtually only method for quantifying meibography image features. This quantification, however, addresses only one aspect of

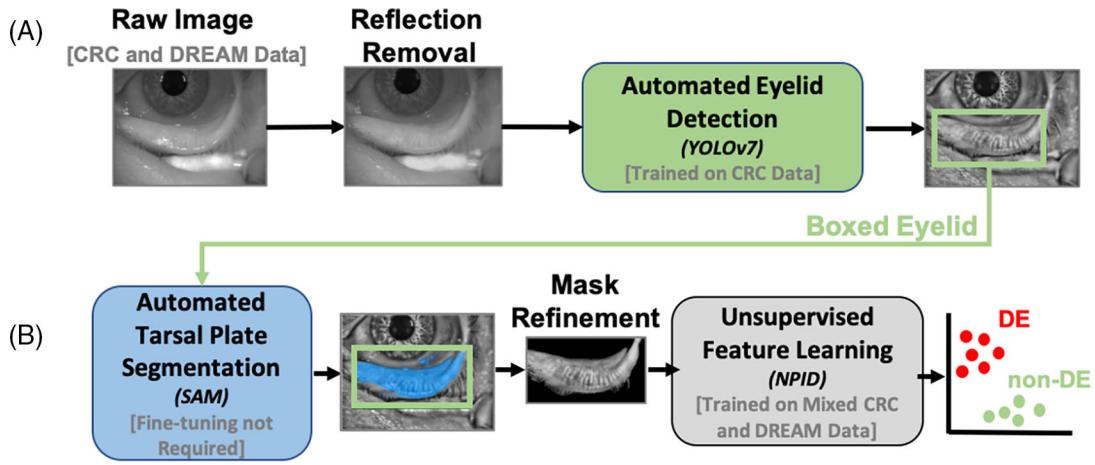


Figure 1. Overview of the unified framework. Depicted is a standardized two-phase method to extract and quantify meibography image features. (**A**) Automated eyelid detection. An eyelid detector (YOLOv7) is boosted from learning with annotated CRC data and is applied on real-world DREAM data to identify the area of interest specified by a bounding box. (**B**) Automated tarsal plate segmentation. A fine-tuning free segmentation model (SAM) refines the detected eyelid area from the previous stage. This results in a segmented tarsal plate mask that highlights the area of interest in a more precise manner. To test the method, an unsupervised feature learning model (NPID) is trained on both refined CRC and DREAM data and then applied to quantitatively compare meibography image features from normal and DE subjects.

MGD (area of atrophy), does not consider the finer morphological structures that are visible in meibography images, and reports of its relationships to downstream signs and symptoms have been equivocal. Researchers have recently turned their attention to fine-grained analysis of the entire meibography image[9] and have begun exploring the potential of artificial intelligence (AI) to advance the analysis of meibography images,[10,11] distinguish meibography images from diseased and healthy eyes using encoded image features not visible to clinicians, and ultimately facilitate more accurate MGD- and DE-related diagnoses.[12]

One approach using unsupervised feature learning has been shown to successfully encode meibography image features and outperform trained clinicians in assigning a meiboscore.[13] Although this approach has demonstrated efficacy in classifying MG characteristics like the meiboscore,[14] it is primarily tailored to the well-curated dataset collected by the University of California, Berkeley, Clinical Research Center (UCB-CRC).[10] Real-world scenarios are often confronted with datasets such as that collected during the Dry Eye Assessment and Management (DREAM) study.[9] The DREAM study was a multisite clinical trial of Ω-3 fatty acid supplements for moderate-to-severe DE patients, during which meibography imaging was conducted at 13 different locations within the United States. The DREAM study images are similar to real-world scenarios in which images are taken by many different observers with differing levels of skill, and they often lack the same level of curation and clinician annotation.

The motivation behind this study lies in addressing the discrepancies (depicted in Fig. 2) between curated datasets (e.g., UCB-CRC) and real-world datasets (e.g., DREAM). The objective is to establish a unified framework that can be boosted from annotated UCB-CRC data and be applied to real-world DREAM data to standardize the quantification of meibography images. To achieve this objective, first the annotated CRC dataset is utilized to train an automated eyelid detector (Fig. 1A). This helps identify regions of interest from non-curated DREAM data. In the second phase, automated tarsal plate segmentation (Fig. 1B) enhances the visualization of MG structure and meibography image features. This refinement improves the ability of the algorithm to differentiate between meibography images of normal and DE subjects.

This approach encompasses the quantification and classification of meibography image features and extends to the standardization of the process. This includes harmonizing diverse datasets, effectively addressing imbalances in DE and non-DE sample sizes, and refining the methodology pipeline to encompass additional steps such as reflection removal and segmented mask refinement. The methodology addresses the challenges from real-world data by bootstrapping the model from annotated CRC data, leading to the development of a robust model capable of performing effectively on previously unseen sources,

thereby improving the reliability and consistency of its outputs.

Methods

The standardized framework presented here encompasses two key phases to retrieve meibography image features: automated eyelid detection and tarsal plate segmentation. In the automated eyelid detection phase, we employ a detection model[15] to automatically identify the eyelid area. For automated tarsal plate segmentation, the Segment-Anything Model (SAM)[16] is employed. SAM utilizes a promptable segmentation approach to refine the initially detected eyelid area and generate more precise tarsal plate masks. Specular reflection removal and mask refinement are also applied to enhance the quality of the images and enable accurate analysis of meibography image features. The details of each phase are presented below.

Data Collection

To conduct the study, two distinct datasets were collected: curated and annotated meibography images collected at the UCB-CRC and meibography images collected during the DREAM study, a multisite clinical trial.[17] All UCB-CRC and DREAM study meibography images were captured with the OCULUS Keratograph 5M (OCULUS, Wetzlar, Germany).[18] In the DREAM dataset, all images were collected from moderate-to-severe DE subjects by design. The UCB-CRC dataset contains images from both DE and non-DE subjects. All images from UCB-CRC subjects who did not meet the DREAM study eligibility criteria were labeled non-DE. To qualify for the DREAM study, subjects had to have an Ocular Surface Disease Index (OSDI) score $>$ 23 and at least two clinical signs of DE, including corneal staining $\geq$ grade 4 in either eye, conjunctival staining $\geq$ grade 1 in either eye, non-invasive tear breakup time $\leq$ 7 seconds in either eye, and/or a Schirmer test strip wetted length $\leq$ 7 mm in either eye. Data collection resulted in a total of 2669 DE images from the DREAM dataset, along with 164 DE images and 1399 non-DE images from the UCB-CRC dataset. To ensure a robust evaluation of our classifier, we divided the collected data into three subsets: 70% for training, 15% for validating, and 15% for testing, for each dataset separately.

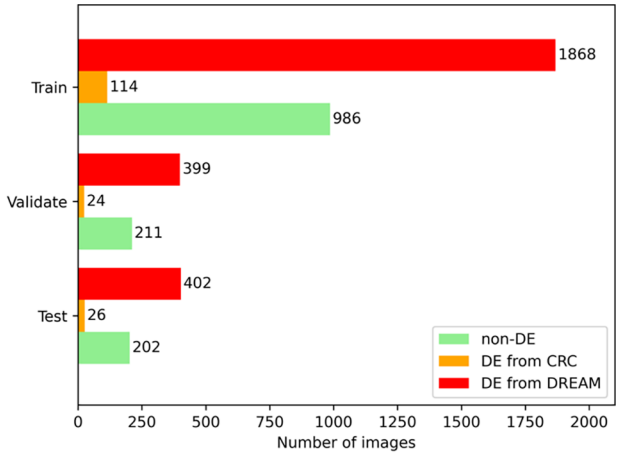
Presented here (see Fig. 2A) is a comparative analysis of the UCB-CRC and DREAM datasets, revealing substantial variations in sample distribution (with a notable imbalance between DE and non-DE samples), image resolution, and DE severity levels. Furthermore, the distribution of meibography image samples is depicted across training, validation, and test sets, clearly indicating that DE samples are significantly less prevalent in the UCB-CRC dataset than in the DREAM dataset (see Fig. 2B).

Data Acquisition for the Automated Eyelid Detector

The dataset acquisition process involved obtaining 2 $\times$ 2 grid screenshot output files from OCULUS meibography scans of both the upper and lower

	UCB-CRC	DREAM
Images, *N*	1563	2669
Image Resolution *[Width, Height]*	[564, 425]	[1360, 1024]
DE Severity	normal to moderate	moderate to severe
Annotation	eyelid, atrophy area	none
Disease Distribution, *n* (%)		
non-DE	1399 (89.5%)	0 (0 %)
DE	164 (10.5%)	2669 (100 %)

(A)



(B)

Figure 2. (**A**) Significant disparities exist between the UCB-CRC and DREAM datasets, particularly in the distribution of DE and non-DE samples (which is highly imbalanced), the resolution of the images, and the severity of DE. (**B**) The distribution of meibography image samples across the training, validation, and test sets, highlighting that the UCB-CRC dataset contains far fewer DE samples compared to the DREAM dataset.

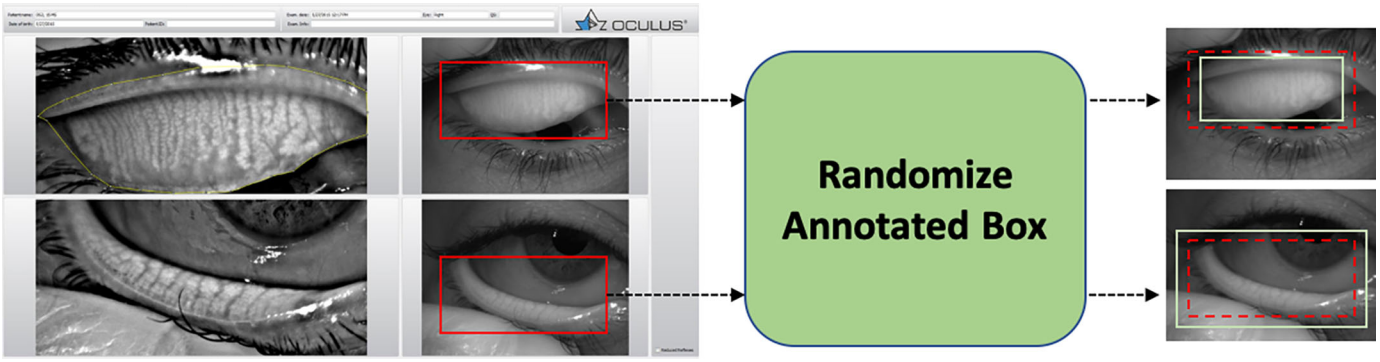


Figure 3. UCB-CRC data preprocessing. Variations were introduced in the size and location of the annotated eyelid area (*green rectangle*) compared to the fixed eyelid area (*red rectangle*) identified by clinicians using the OCULUS machine. This strategy enabled the model to learn and adapt to a range of diverse eyelid area characteristics.

eyelids of 399 DE and 235 non-DE subjects. In the 2 × 2 screenshots (Fig. 3, left), two images are in their raw form, and the other two are enhanced with the proprietary OCULUS software. In order to train the detector, the raw images from the 2 × 2 screenshots were extracted from the annotated UCB-CRC datasets. During the screening process conducted by clinicians on the OCULUS machine, precise control of the joystick was exercised to identify the eyelid area. The eyelid area was visually delineated using a fixed scale and location, represented by a red rectangle, as shown in Figure 3. This red rectangle served as the annotation for the eyelid area for training the eyelid detector to identify the area of interest in previously unseen meibography images. To ensure the robustness of the detection model and its applicability in real-world scenarios, controlled variations in the size and location of the annotated rectangles were introduced while maintaining a fixed center. This strategy enabled the model to learn and adapt to diverse eyelid area characteristics encountered during clinical examinations.

Automated Eyelid Detection

The automated eyelid detector aims to alleviate the tedious and subjective nature of manually locating eyelid areas with a joystick by leveraging AI. The ultimate goal is to develop an AI model capable of automatically classifying images as DE or non-DE, and this preliminary stage serves to isolate the area of interest with an eyelid detector, ensuring that the model can focus exclusively on the clinically relevant parts of the image while disregarding extraneous elements such as the surrounding skin, eyelashes, ocular surface, or the investigator's thumb or swab.

To achieve this, the Single Shot Detector (SSD) algorithm[19] was compared with the You Only Look Once (YOLOv7) detection model[15] using labeled data from the UCB-CRC. The SSD algorithm generates a feature map from an input image with a single pass through a convolutional network. YOLOv7 is a state-of-the-art object detection algorithm that excels in real-time performance and accuracy. It employs a single neural network to simultaneously predict bounding boxes and class probabilities for multiple objects within an image. Ultimately, the YOLOv7 model was selected due to its superior performance in detecting the eyelid area in meibography images (see Results).

The trained YOLOv7 model was then tested on the previously unseen DREAM dataset,[9] which is a larger dataset without annotations. This evaluation allowed us to assess the generalization capability of the model and its effectiveness in detecting the eyelid area in real-world, non-curated meibography images. Figure 4 presents visualization examples that demonstrate the identification of the eyelid areas in meibography images.

Automated Tarsal Plate Segmentation

Tarsal plate segmentation plays an important role in the overall approach, as it allows the clinician or researcher to refine the initially detected eyelid area and obtain a more precise segmented tarsal plate mask. This refined mask increases the likelihood of the model heavily weighting areas within the image that show the MGs in order to distinguish DE from non-DE, as opposed to relying on clinically irrelevant areas of the image.

SAM was employed to automate the segmentation of the tarsal plate.[16] SAM is a state-of-the-art image segmentation model that utilizes a prompt-based

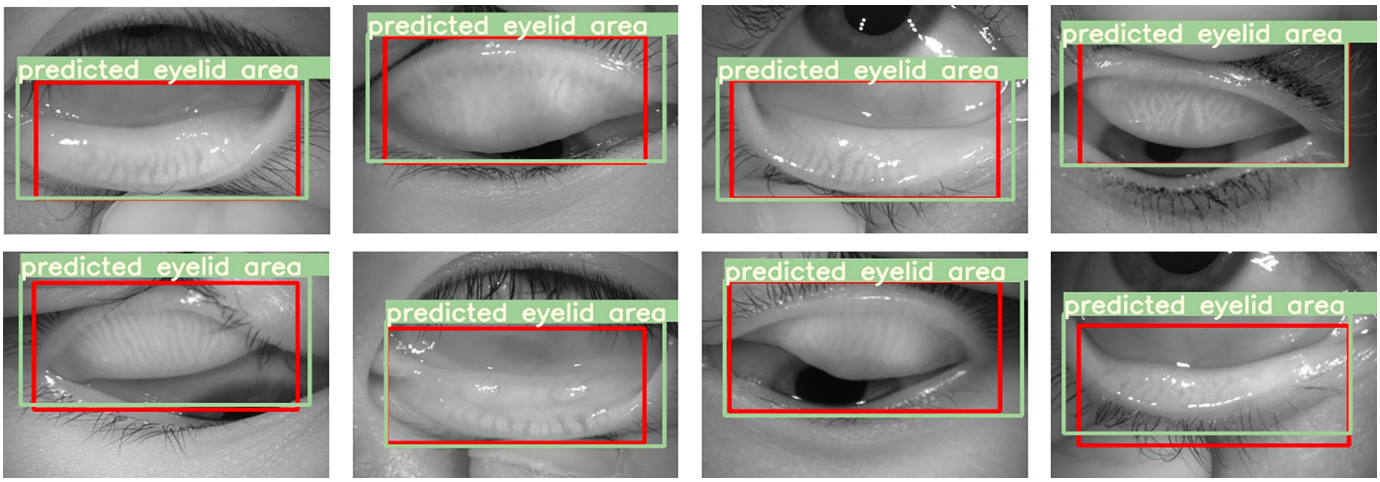


Figure 4. Automated eyelid detection. This figure presents visual examples that illustrate the accuracy of the automated eyelid detection method. The ground truth bounding boxes are indicated in *red*, representing the reference eyelid areas, and the *lime green* bounding boxes depict the predictions of the model.

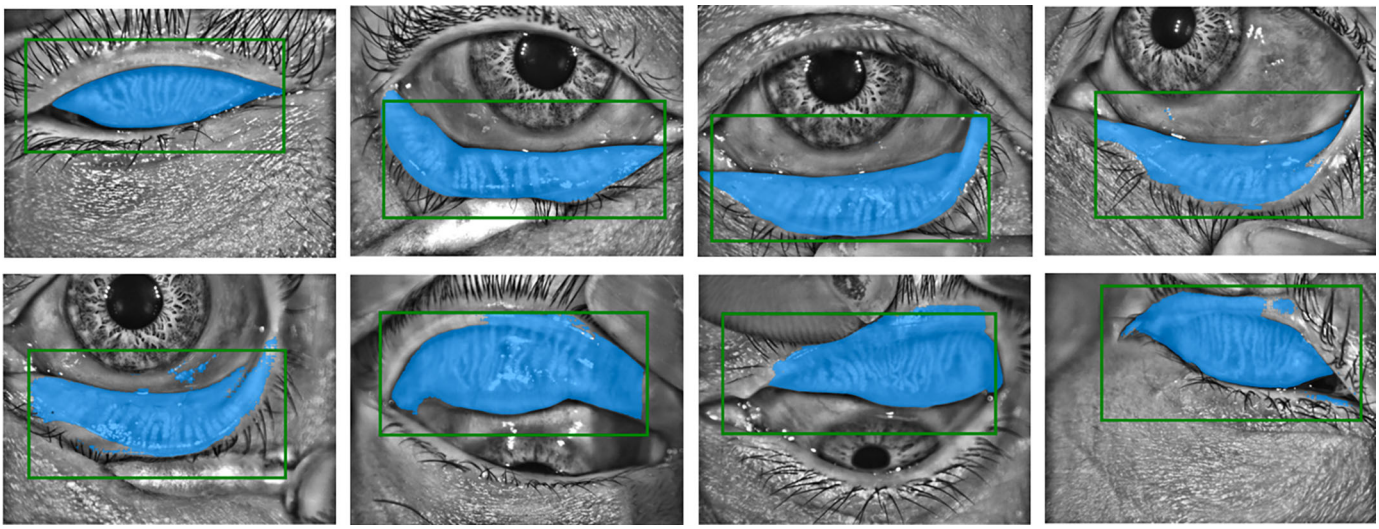


Figure 5. Segmented tarsal plate masks from the DREAM dataset. The *green bounding box* represents the box prompt generated by the automated eyelid detector in phase 1 and used by the SAM in phase 2, and the *blue area* signifies the prediction of the model.

approach, trained on an extensive dataset of over 1 billion masks. By utilizing SAM, the tarsal plate within the eyelid area can be delineated. The model design and training methodology of SAM make it highly adaptable for various tasks, including zero-shot learning.[20]

To illustrate the effectiveness of the tarsal plate segmentation, depictions of the segmented tarsal plate masks from the DREAM dataset are presented in Figure 5. These DREAM results demonstrate the capability of the model to accurately delineate the boundaries of the tarsal plate and generate refined masks in real-world, non-curated meibography images.

Segmented Meibography Image Feature Refinement

Segmented tarsal plate mask refinement is an important step in the analysis as the initial segmentation masks often contain missing or discontinuous parts of the tarsal plate, which in turn can introduce inaccuracies in subsequent measurements such as total tarsal plate area, percent area of atrophy, MG density, and gland morphological characteristics. Therefore, it is essential to develop an algorithm that can refine the segmentation mask and effectively excise clinically irrelevant areas and fill gaps in the images.

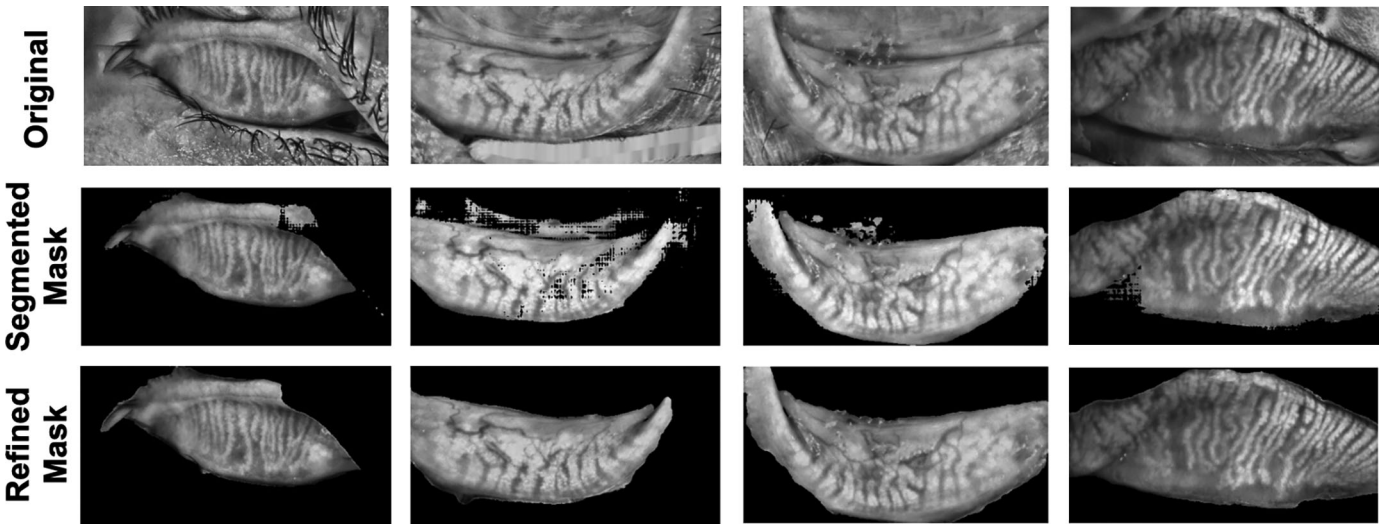


Figure 6. Examples of segmented tarsal plate mask refinement. This figure depicts the visual enhancement of meibography images through the mask refinement algorithm. *Row 1* displays the original raw images, *row 2* illustrates the initial state of the segment masks, and *row 3* displays the enhanced state achieved after the application of the mask refinement algorithm.

To address this issue, the CascadePSP algorithm[21] was employed for refining segmentation masks. The algorithm leverages a cascade framework consisting of multiple stages, each performing global and local refinement steps. At each stage, the global refinement is first applied to incorporate high-level contextual information to refine the initial segmentation mask, capturing global patterns and context within the image. In the subsequent local refinement, the algorithm focuses on enhancing the local details and fine-grained boundaries of the tarsal plate, ensuring its accurate delineation.

The refinement process of the segmented tarsal plate mask enables a more thorough exploration and analysis of meibography image features. Figure 6 presents visual examples of tarsal plate masks depicting the initial state before the refinement process (row 2) and the improved state after applying the mask refinement algorithm (row 3). This comparison highlights the enhanced quality and completeness of the tarsal plate masks obtained through the refinement process.

Specular Reflection Removal

Specular reflections in images of the MG can occur due to various factors such as the lighting conditions during imaging or the surface properties of the palpebral conjunctival tissues.[22] These reflections can obscure the visibility of the MG and affect the accuracy of analysis. Therefore, it was important to develop an automated algorithm to effectively remove these reflections.

Given a raw meibography image $I(x, y)$, a binary threshold is applied to the grayscale image to create a mask that identifies the specular reflections:

$$M(x, y) = \begin{cases} 1, & if \quad I(x, y) > \tau \\ 0, & otherwise \end{cases} \tag{1}$$

where the threshold value τ is empirically determined to distinguish the reflections from the rest of the image, and $M(x, y)$ is the binary mask representing the area of specular reflections in the meibography image $I(x, y)$. Here, x and y are the spatial coordinates that define the position of each pixel in the image.

To include neighboring pixels in the mask, dilation using a 3×3 kernel as default is performed. Let K be the 3×3 dilation kernel and D be the dilated mask:

$$D(x, y) = \max_{i,j \in K} \{M(x + i, y + j)\} \tag{2}$$

where $D(x, y)$ incorporates neighboring pixels using the 3×3 kernel K. The coordinates x and y specify the position of a pixel in the original mask, M, and i and j represent the relative positions of neighboring pixels within the 3×3 kernel. This helps to expand the reflection region and ensure that all affected pixels are accounted for. To achieve this, the dilated mask, D, was utilized to guide an inpainting algorithm for its ability to fill in the reflection regions with pixel values that maintain the integrity of the original image. This inpainting process begins by carefully assessing the information from surrounding pixels to predict and fill in any gaps or damaged areas within the image. The goal is to ensure that the reconstructed segments integrate with the existing

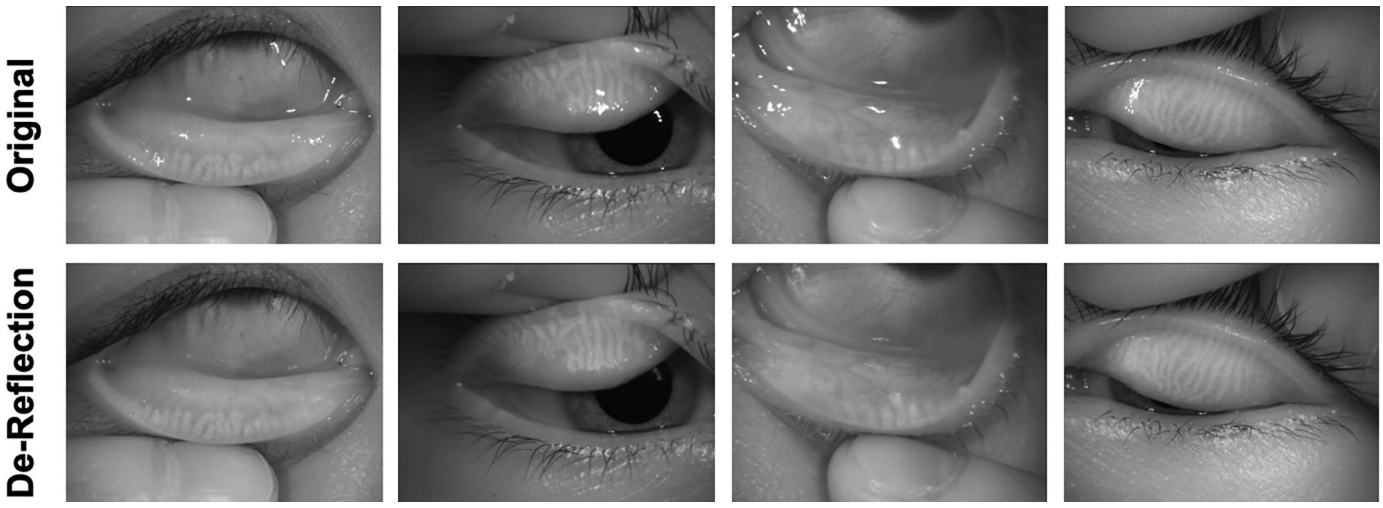


Figure 7. Examples of specular reflection removal. This figure presents a visual comparison of meibography images before (*row 1*) and after (*row 2*) the application of the reflection removal algorithm.

image features, resulting in an inpainted image that maintains visual consistency with the original meibography image. Visual examples of meibography images before (row 1) and after reflection removal (row 2) are presented in Figure 7. By mitigating the loss of pixels potentially encoding diagnostic information caused by the presence of specular reflections, the quality of the images is enhanced to facilitate more accurate analysis of meibography image features.

Unsupervised Meibography Image Feature Learning

The non-parametric instance discrimination (NPID) method[23] was employed to differentiate between DE and non-DE meibography images. An unsupervised instance discriminator was pretrained on ImageNet[24] and then trained on encoded meibography image features after eyelid detection, tarsal plate segmentation, and mask refinement. The primary objective of this discriminator was to establish a representation space in which similar features from the same category (DE or non-DE) are brought into closer proximity while features from distinct categories are mapped farther apart. This enables the learned feature representation to effectively encapsulate the underlying patterns and characteristics that distinguish DE and non-DE meibography images.

Results

Experiments were conducted to demonstrate the performance of the proposed method. In these experiments, the automated eyelid detector, tarsal plate segmentation, and mask refinement steps were applied prior to inputting the resulting image to the unsupervised instance discriminator to distinguish meibography images from DE and non-DE subjects.

Eyelid Detection Results

The automated eyelid detection model was trained using the annotated UCB-CRC dataset. The performance evaluation was conducted on the UCB-CRC validation split to assess the effectiveness of the model in detecting eyelid areas. Two detection models were employed, SSD[19] and YOLOv7,[15] and they were compared in terms of their performance (Table 1). The YOLOv7 model consistently achieved higher mean average precision (mAP), indicating its superior performance in detecting the eyelid area.

To further enhance the performance of the detection model, variations to the annotations were intro-

Table 1. Performance of Two Eyelid Area Detection Models

	mAP at 0.5	mAP at 0.5:0.95
SSD	73.2%	31.5%
YOLOv7	95.3%	51.2%
YOLOv7**	98.6%	73.1%

Two detection models, SSD and YOLOv7, were evaluated in terms of mAP at 0.5 and mAP at 0.5:0.95, and YOLOv7 achieved better accuracy. To further improve the accuracy of the detection model, annotation variations were introduced for scale and position adaptability, along with application of a contrast-enhancing high-pass filter. These strategies notably enhanced the detection performance (shown as YOLOv7**).

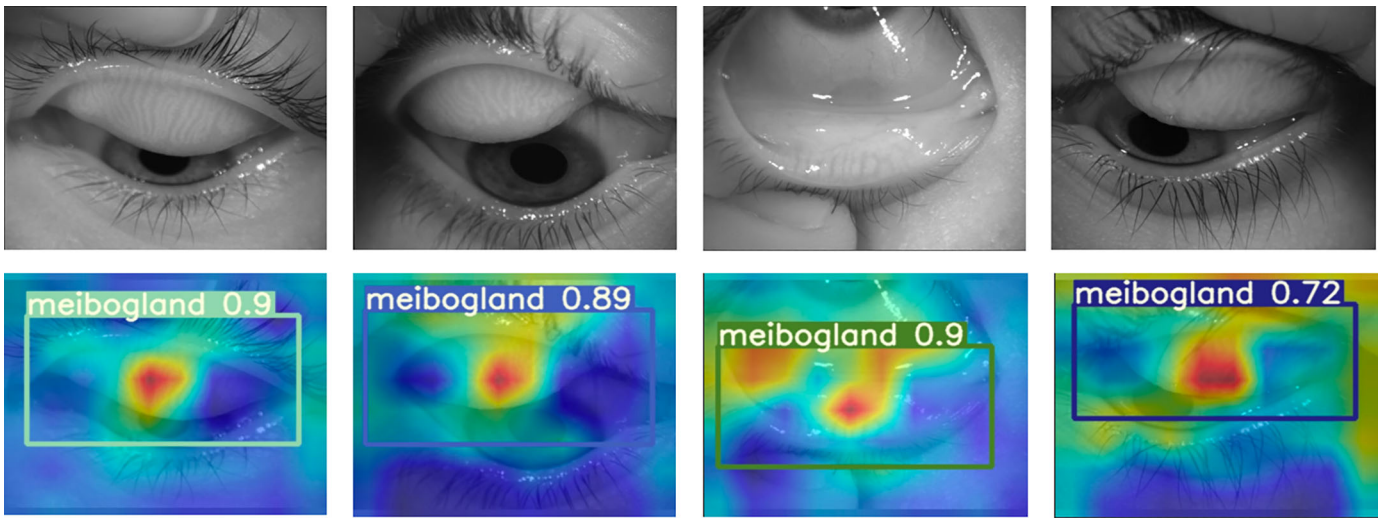


Figure 8. Examples of Grad-CAM visualization. Grad-CAM was employed to gain visual insight into the regions of the meibography images that played a significant role in the final detection process. The Grad-CAM analysis highlights that the identified important visual regions consistently correspond to the eyelid areas containing the Meibomian glands.

duced, allowing the model to adapt to different scales and positions of the gland areas encountered in real-world scenarios. A high-pass filter was also applied to the input images to enhance contrast, further improving detection accuracy.

Gradient-weighted class activation mapping (Grad-CAM)[25] was utilized to provide visual understanding of the key areas in meibography images that played a significant role in the detection process. This adaptation was achieved by applying Grad-CAM to the final convolutional layer outputs of YOLOv7, focusing on the feature maps that are actively involved in predicting the class and location of the object. The visual evidence from Grad-CAM, illustrated in Figure 8, qualitatively confirms that the attention of the model was consistently focused on MG regions essential for the diagnosis, ensuring the reliance of the model on pertinent features within the tarsal plate area. This approach underscores the ability of the model to concentrate on diagnostically relevant aspects of the image while disregarding potential distractions from background elements (e.g., eyelashes) or other non-relevant parts of the image (e.g., pupil). Such clarity in the focal areas of the model is critical, enhancing its trustworthiness and interpretability in the context of medical imaging, where the rationale behind diagnostic predictions is of paramount importance.

Tarsal Plate Segmentation Results

After eyelid detection had been performed by the trained model on all UCB-CRC and DREAM study meibography images, the next step was to perform a more detailed segmentation of the tarsal plate area containing the MG. In the case of the DREAM dataset, which does not provide mask annotations for tarsal plates, a human assessment strategy was devised to evaluate the segmentation performance. The principal clinical investigator (MCL), alongside a team of laboratory members, examined 634 segmentation outcomes and developed a qualitative categorization to classify these results, as depicted in Figure 9. The evaluation process categorized the segmentation results into four distinct categories, each capturing different aspects of the segmentation outcome:

1. *Accurate and precise segmentation*—Tarsal plate segmentation is well executed, accurately delineating the entirety of the visible tarsal plate without including surrounding tissues.
2. *Tarsal plate with eyelashes and/or skin*—Segmentations encompass not only the tarsal plate but also surrounding eyelashes and/or skin.
3. *Tarsal plate with ocular surface*—The tarsal plate is properly segmented, but the segmentation area extends to include parts of the ocular surface.
4. *Other segmentation challenges*—The segmentation process leads to incomplete or inaccurately segmented tarsal plates, due to factors such as poor focus or alignment of the original images, image artifacts, partially everted eyelids, or the inclusion of foreign objects such as fingers or cotton swabs in the meibography images.

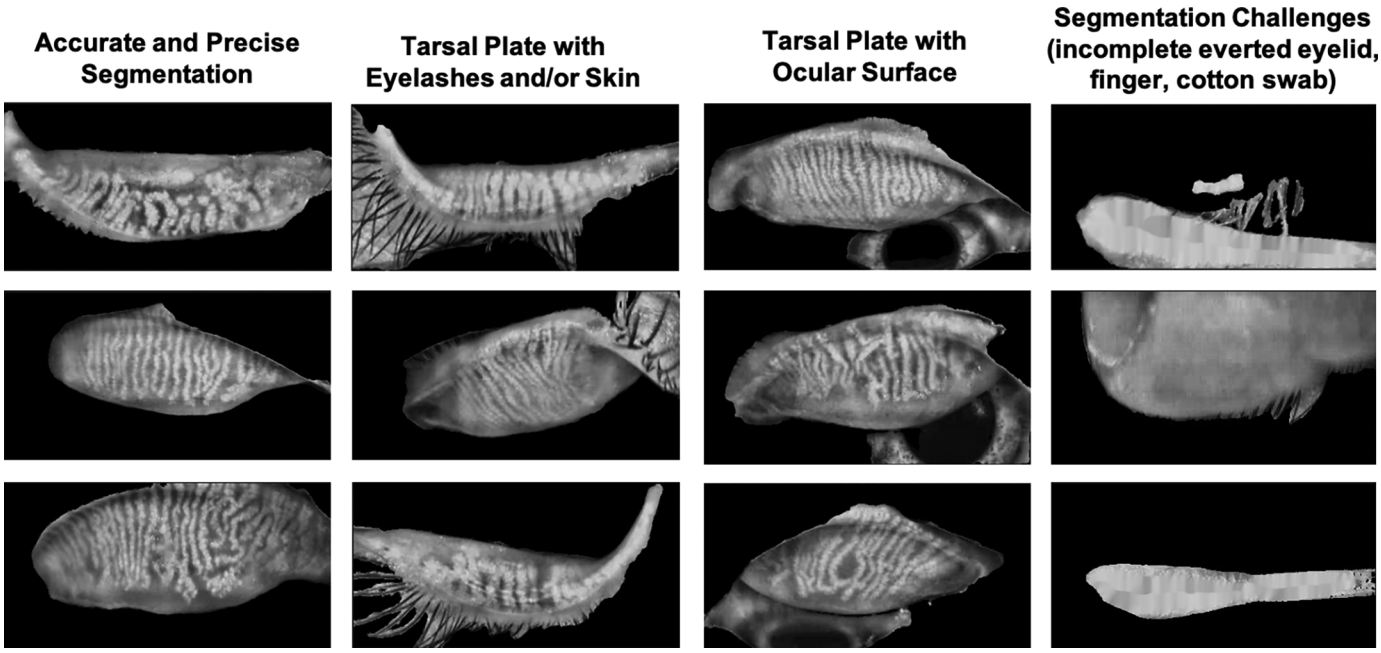


Figure 9. Segmentation outcome categorization. This figure presents the qualitative classification of masked and segmented images into four categories, detailing the varying levels of segmentation precision and characteristics.

Table 2. Evaluation for Tarsal Plate Segmentation

	CRC (DE)	CRC (Non-DE)	DREAM (DE)	Total
Accurate and precise segmentation	15 (63%)	124 (59%)	197 (49%)	336 (53%)
Tarsal with eyelash and/or skin	1 (4%)	18 (9%)	85 (21%)	104 (17%)
Tarsal with ocular surface	7 (29%)	61 (29%)	66 (17%)	134 (21%)
Usable images	23 (96%)	203 (97%)	348 (87%)	574 (91%)
Other segmentation challenges	1 (4%)	8 (3%)	51 (13%)	60 (9%)
Tarsal $>$ 50% (success)	23 (96%)	202 (96%)	333 (83%)	558 (88%)
Tarsal $<$ 50% (failure)	1 (4%)	9 (4%)	66 (17%)	76 (12%)
Total	24 (100%)	211 (100%)	399 (100%)	—

This evaluation for tarsal plate segmentation includes a four-category assessment and an additional evaluation to determine whether the tarsal plate area constitutes more than 50% of the image.

It is important to note that images in the first three categories are all usable for meibography evaluation, as they all include the majority of the tarsal plate area with the MG in clear focus. The fourth-category images are generally not usable for meibography evaluation, and their identification serves as a screening step to eliminate flawed images. Upon review by the lead clinical investigator and the study team, there were no images that were considered usable for MG analysis that nevertheless were judged to be in category 4 due to flaws in detection, segmentation, or mask refinement.

In addition to the four-category qualitative assessment, all images were also inspected to determine whether the represented tarsal plate (the area of interest) constituted more than approximately 50% of the masked and segmented image. This evaluation resulted in either success or failure outcomes. In Table 2, summarizing the tarsal plate segmentation outcomes in this way reveals specific insights and challenges due to the diverse datasets analyzed. The row for "Tarsal $>$ 50% (success)" highlights that in 88% of images processed, more than half of the resulting final image depicts the tarsal plate area of interest, affirming the efficacy of the model in delineating the area of interest. However, the images categorized as "Other" represent complex cases often influenced by factors such as

partially everted eyelids or foreign objects, and they present a collective challenge, ranging from 3% to 13% of images, with the highest number (13%) coming from the non-curated, real-world data. Similarly, the "Tarsal with ocular surface" outcomes in the DREAM (DE) dataset (17%) suggest opportunities for refining the segmentation algorithms to precisely isolate the tarsal plate without extending into ocular surface components.

Meibography Image Feature Learning Details

Having detected the eyelid area, refined the mask, and segmented the tarsal plate area, a model was next trained to learn features of the resulting images and attempt to discriminate meibography images from DE and non-DE subjects. We employed the NPID approach to quantitatively analyze the extracted image features. For this purpose, ResNet-50[26] was employed as the backbone network, which encoded the outputs into 128-dimensional vectors throughout all experimental setups. The model training process utilized stochastic gradient descent[27] with a momentum value of 0.9, employing a batch size of 16 and an initial learning rate of 0.005.

The network model was trained over 200 epochs, building upon the foundation of the pretrained ImageNet model.[24] To augment the data for robust training,[28] various techniques were applied to each image feature sample. Random cropping of 224 × 224 pixel regions from the images served to introduce spatial diversity into the training data, allowing the model to better capture different gland configurations and orientations. Color jittering was employed to enhance the ability of the model to generalize by introducing variations in grayscale color intensity in the infrared images, emulating real-world variations in image acquisition conditions. Horizontal flipping was used to further diversify the training data, ensuring that the model could effectively recognize MG features regardless of their orientation. These techniques collectively aimed to enrich the training dataset, promoting the ability of the model to generalize and perform effectively across a wide range of MG feature variations commonly encountered in practice.

Meibography Image Feature Learning Results

Initially, results on the validation set based on the NPID model pretrained on ImageNet and fine-tuned on UCB-CRC + DREAM data achieved surprisingly high accuracy. The instance-wise accuracy was 92.8%, and the DE and non-DE class accuracies were 92.5% and 92.9%, respectively. This initial result was surprising because DE is a multifaceted disease with many interrelated factors ultimately contributing to patient symptoms, so the model would not be expected to achieve such high classification accuracy using only meibography images as the sole input. These results raised concerns about the reliability of the classification model and extracted meibography image features because all non-DE samples originated from the UCB-CRC dataset, and all DREAM study subjects were moderate-to-severe DE by design. Given this imbalanced data distribution and the initial results, it was plausible to suspect that the prediction of the model might be influenced by the source of each dataset, rather than accurately capturing the intrinsic characteristics of the MG features. Therefore, the source-wise accuracy for correctly predicted DE samples was examined, and a substantial difference in accuracy was found between the UCB-CRC (DE) and DREAM (DE) datasets. The source-wise accuracy for UCB-CRC (DE) samples was found to be only 16.6%, whereas it was 97.5% for DREAM (DE) samples. This observation suggested that the classifier might have been relying on features that reflected the dataset source rather than the characteristics of the MG.

To address the issue of dataset source correlation, two strategies were implemented. First, the image samples in the DREAM dataset were resized to have a scale similar to that of the UCB-CRC images. This resizing aimed to align the scales of samples from different sources, reducing the potential bias introduced by source-specific image characteristics. Second, the scale range for the random resize cropping augmentation was adjusted from 0.7 to 1, ensuring that the cropping process remained focused on the image areas containing the MG. Finally, an additional random rotation of ±5° was added in order to improve the robustness of the model in handling meibography images with slight rotational variations in real-world scenarios, and empirical fine-tuning was performed to improve source-wise accuracy.

By implementing these strategies in the data and model training scheme, a substantial improvement in source-wise accuracy was observed for UCB-CRC (DE) samples, increasing from 16.6% to 70.8%. This adjustment resulted in more balanced source-wise accuracies between UCB-CRC (DE) and DREAM (DE) samples, as shown in Table 3. However, it is important to note that further work is required to prevent the model from relying on dataset source and instead prioritize the classification based on the characteristic MG features indicative of DE and non-DE conditions.

Table 3. Performance of Classification Model on the UCB-CRC and DREAM Datasets

	DE Accuracy Within UCB-CRC ($n = 24$)	DE Accuracy Within DREAM ($n = 399$)	Instance-Wise Accuracy	Non-DE Class Accuracy	DE Class Accuracy
Raw meibography image					
Baseline results	4.2%	97.7%	94.4%	92.2%	96.7%
Raw image + eyelid detection					
Baseline results	4.2%	92.5%	91.6%	87.5%	95.7%
Raw image + eyelid detection + tarsal plate segmentation					
Baseline results	16.6%	97.2%	92.8%	92.5%	92.9%
+ Rescale DREAM	41.7%	91.4%	86.6%	81.3%	91.9%
+ Adjust random crop	54.2%	88%	84.2%	76.7%	91.7%
+ Rotation and fine-tune	70.8%	91.7%	80.8%	72%	90.5%

The results indicate that the ability of the classifier to distinguish between DE and non-DE samples from tarsal plate masks can be attributed more to the inherent characteristics of the MG features rather than being solely influenced by the dataset source. This also reinforces the value of the segmentation-based approach, with the results serving as a baseline comparison for models utilizing raw meibography images without standardization and those enhanced with eyelid detection.

Meibography Image Feature Visualization

The Uniform Manifold Approximation and Projection (UMAP) visualization[29] technique was applied to visualize and analyze feature distributions from 634 validation samples of meibography images, as depicted in Figure 10. The UMAP plot distinctly separates DE samples from the DREAM dataset and non-

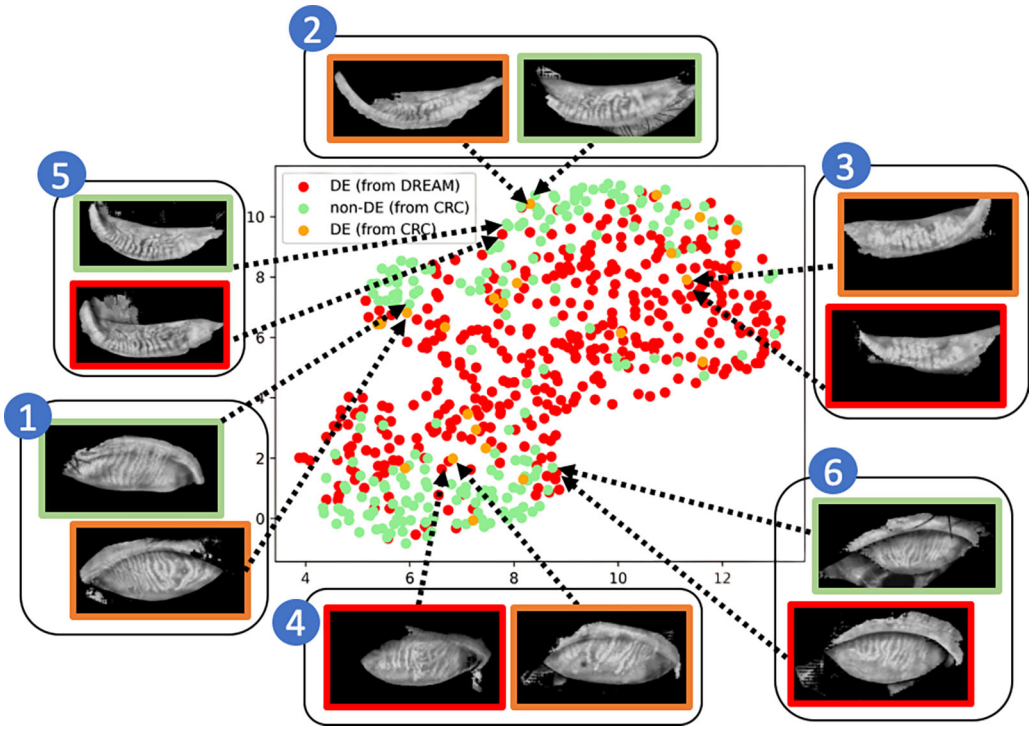


Figure 10. UMAP visualization. In examples 1 and 2, it is observed that the *orange dots* (indicating DE samples from the UCB-CRC dataset) are closer to the *green dots* (indicating non-DE samples from UCB-CRC dataset). This closeness implies that many DE samples in UCB-CRC have moderate severity, leading to similar tarsal plate morphology compared to non-DE samples. Conversely, in examples 3 and 4, the *orange dots* are nearer to the *red dots* (indicating DE samples in the DREAM dataset), suggesting that some DE samples in the DREAM dataset may also exhibit moderate severity. In examples 5 and 6, we observe that the *red dots* are close to the *green dots*. This suggests that, although images naturally cluster by the DE grade, images from different grades may be closely positioned. This proximity could stem from inconsistencies in subjective annotations or because similarities in overall morphology overshadow more subtle distinctions.

DE samples from the UCB-CRC dataset, highlighting the feature discriminative capability of our model. Moreover, non-DE samples exhibit subclustering, aligning with anatomical distinctions between the upper and lower eyelids.

Interestingly, we observe a subset of DE samples (shown in examples 1 and 2 in Figure 10) from the UCB-CRC dataset (orange dots) intermingling with non-DE samples (green dots) from the same source. This pattern indicates possible morphological similarities in the tarsal plate region between some DE and non-DE subjects within the UCB-CRC dataset. This is likely because the CRC dataset includes many DE cases that are mild or moderate, whereas the DREAM study recruited moderate-to-severe DE patients only (see Fig. 2). The similar phenomenon can also explain the closeness between CRC and DREAM DE samples illustrated in examples 3 and 4 in Figure 10. This observation could potentially affect the predictive performance of the image-based model when distinguishing between DE and non-DE conditions based solely on image features. In examples 5 and 6, we noticed that the DREAM DE samples (red dots) are positioned close to some UCB-CRC non-DE samples (green dots). This observation suggests that although images naturally cluster according to the DE grade, occasionally images from different grades appear in close proximity. This phenomenon could result from inconsistencies in subjective annotation or because the overall morphological similarities overshadow local differences. Acknowledging this limitation, future work could incorporate clinical metadata alongside meibography imaging, forming a multimodal input strategy. Such an approach, inspired by prior work,[31–33] would provide the model with a richer information set, potentially refining its predictive accuracy and enhancing its clinical utility.

Discussion

The current study presents a standardized framework that quantifies and classifies meibography image features and establishes a consistent analysis process. This framework introduces a two-stage approach involving automated eyelid detection and tarsal plate segmentation. In the first stage, an AI model trained on curated UCB-CRC data can also identify relevant meibography image features within non-curated datasets such as that from the DREAM study. The second stage employs automated tarsal plate segmentation to enhance the accuracy of gland structure characterization, thereby enabling more precise comparisons between non-DE and DE images.

The proposed framework represents a significant effort to standardize the analysis process for meibography images. Meibography images from real-world clinical settings often present unique challenges compared to curated datasets. These challenges include variations in image quality, lighting conditions, and patient positioning, resulting in a wide array of image characteristics.[30] Moreover, real-world datasets are typically less controlled, non-curated, and lacking annotations. By bootstrapping the model from curated datasets (e.g., UCB-CRC), this methodology addresses the challenges from real-world datasets (e.g., DREAM). The approach presented here ensures robustness and reliability in classifying meibography images and standardizes processing of meibography images, thus improving comparisons of meibography taken at different locations and/or by different investigators.

Although the methodology is currently based on a single set of criteria that defines DE and non-DE subjects, future endeavors will extend the classification to accommodate multiple different criteria for comparisons between normal and DE subjects.[3] This expansion could provide insights into the set of criteria that best defines DE for purposes of predictions based on meibography. Another aspect that requires consideration is tarsal plate segmentation. In the non-curated, real-world DREAM dataset, we observed that "Tarsal with eyelash and/or skin" and "Tarsal with ocular surface" outcomes accounted for 21% and 17% of processed images, respectively. This highlights the need to refine segmentation algorithms to precisely isolate the tarsal plate, minimizing interference with the eyelashes, skin, and ocular surface components. These observations underscore both the strengths and opportunities for improvement in tarsal plate segmentation, setting the stage for more robust meibography image analysis in real-world scenarios.

In real-world applications, the proposed framework addresses the limitations inherent in existing assessment methods. By automating the process and effectively utilizing the AI model, subjectivity is minimized and efficiency increased in analyzing meibography image features. This approach also significantly reduces the gap between the curated UCB-CRC dataset and real-world datasets, such as that from the DREAM study, which lack annotations. This integration of curated and non-curated datasets facilitates a more generalizable evaluation of meibography image features and their variations in health and disease.

This framework introduces a standardized methodology that is universally applicable. It simplifies the integration of diverse meibography images into large-scale databases and plays a central role in enhancing the comparability of research findings across

different sources by standardizing the image processing pipeline. Whether the task at hand involves disease classification, quantifying gland features, or any other meibography-related analysis, this framework optimizes the process, making it accessible to a broader research community while fostering consistency in analytical approaches.

Acknowledgments

The authors thank Dorothy Ng, Jessica Vu, Jasper Cheng, Kristin Kiang, Megan Tsiu, Fozia Khan Ram, April Myers, Shawn Tran, Michelle Hoang, and Zoya Razzak for providing annotations for the meibography images.

Supported by a grant from the National Institutes of Health (R21EY033881 to MCL and SXY); by a University of California, Berkeley, Clinical Research Center Unrestricted Fund (MCL); and by the Roberta J. Smith Research Fund (MCL).

Disclosure: **C.-H. Yeh**, None; **A.D. Graham**, None; **S.X. Yu**, None; **M.C. Lin**, None

References

1. Pult H, Nichols JJ. A review of meibography. *Optom Vis Sci*. 2012;89(5):E760–E769.
2. Butovich IA. Meibomian glands, meibum, and meibogenesis. *Exp Eye Res*. 2017;163:2–16.
3. Craig JP, Nichols KK, Akpek EK, et al. TFOS DEWS II definition and classification report. *Ocul Surf*. 2017;15(3):276–283.
4. Nichols KK, Foulks GN, Bron AJ, et al. The international workshop on Meibomian gland dysfunction: executive summary. *Invest Ophthalmol Vis Sci*. 2011;52(4):1922–1929.
5. Bron AJ, Tiffany JM, Gouveia SM, Yokoi N, Voon LW. Functional aspects of the tear film lipid layer. *Exp Eye Res*. 2004;78(3):347–360.
6. Dursch TJ, Li W, Taraz B, Lin MC, Radke CJ. Tear-film evaporation rate from simultaneous ocular-surface temperature and tear-breakup area. *Optom Vis Sci*. 2018;95(1):5–12.
7. Teo CH, Ong HS, Liu YC, Tong L. Meibomian gland dysfunction is the primary determinant of dry eye symptoms: analysis of 2346 patients. *Ocul Surf*. 2020;18(4):604–612.
8. Arita R, Itoh K, Maeda S, et al. Proposed diagnostic criteria for obstructive Meibomian gland dysfunction. *Ophthalmology*. 2009;116(11):2058–2063.
9. Asbell PA, Maguire MG, Peskin E, Bunya VY, Kuklinski EJ. Dry eye assessment and management (DREAM©) study: study design and baseline characteristics. *Contemp Clin Trials*. 2018;71:70–79.
10. Wang J, Yeh TN, Chakraborty R, Stella XY, Lin MC. A deep learning approach for Meibomian gland atrophy evaluation in meibography images. *Transl Vis Sci Technol*. 2019;8(6):37.
11. Wang J, Li S, Yeh TN, et al. Quantifying Meibomian gland morphology using artificial intelligence. *Optom Vis Sci*. 2021;98(9):1094–1103.
12. Lin MC, Graham AD, Kothapalli T, et al. Lifestyle and behaviors: predicting clinical signs and symptoms with machine learning. *Invest Ophthalmol Vis Sci*. 2023;64(8):2880.
13. Yeh C-H, Yu SX, Lin MC. Meibography image phenotyping and classification from unsupervised discriminative feature learning. *Transl Vis Sci Technol*. 2021;10(2):4.
14. Arita R, Suehiro J, Haraguchi T, Shirakawa R, Tokoro H, Amano S. Objective image analysis of the Meibomian gland area. *Br J Ophthalmol*. 2014;98(6):746–755.
15. Wang CY, Bochkovskiy A, Liao HY. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7464–7475). Piscataway, NJ: Institute of Electrical and Electronics; 2023.
16. Kirillov A, Mintun E, Ravi N, et al. Segment anything. arXiv. 2023, https://doi.org/10.48550/arXiv.2304.02643.
17. Hussain M, Shtein RM, Pistilli M, et al. The Dry Eye Assessment and Management (DREAM) extension study–a randomized clinical trial of withdrawal of supplementation with omega-3 fatty acid in patients with dry eye disease. *Ocul Surf*. 2020;18(1):47–55.
18. Markoulli M, Duong TB, Lin M, Papas E. Imaging the tear film: a comparison between the subjective Keeler Tearscope-Plus and the objective Oculus Keratograph 5M and LipiView interferometer. *Curr Eye Res*. 2018;43(2):155–162.
19. Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector. In: *Computer Vision–ECCV 2016* (pp. 21–37). Berlin: Springer International Publishing; 2016.
20. Shi P, Qiu J, Abaxi SM, Wei H, Lo FP, Yuan W. Generalist vision foundation models for medical imaging: a case study of segment anything model

on zero-shot medical segmentation. *Diagnostics*. 2023;13(11):1947.
21. Cheng HK, Chung J, Tai YW, Tang CK. CascadePSP: toward class-agnostic and very high-resolution segmentation via global and local refinement. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020* (pp. 8890–8899). Piscataway, NJ: Institute of Electrical and Electronics; 2020.
22. Saha RK, Chowdhury AM, Na KS, et al. AI-based automated Meibomian gland segmentation, classification and reflection correction in infrared meibography. arXiv. 2022, https://doi.org/10.48550/arXiv.2205.15543.
23. Wu Z, Xiong Y, Yu SX, Lin D. Unsupervised feature learning via non-parametric instance discrimination. In: *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3733–3742). Piscataway, NJ: Institute of Electrical and Electronics; 2018.
24. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). Piscataway, NJ: Institute of Electrical and Electronics; 2009.
25. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the 2017 IEEE International Conference on Computer Vision* (pp. 618–626). Piscataway, NJ: Institute of Electrical and Electronics; 2017.
26. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). Piscataway, NJ: Institute of Electrical and Electronics; 2016.
27. Ruder S. An overview of gradient descent optimization algorithms. arXiv. 2016, https://doi.org/10.48550/arXiv.1609.04747.
28. Shijie J, Ping W, Peiyi J, Siping H. Research on data augmentation for image classification based on convolution neural networks. In: Proceedings of the *2017 Chinese Automation Congress (CAC)* (pp. 4165–4170). Piscataway, NJ: Institute of Electrical and Electronics; 2017.
29. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv. 2018, https://doi.org/10.48550/arXiv.1802.03426.
30. Wise RJ, Sobel RK, Allen RC. Meibography: a review of techniques and technologies. *Saudi J Ophthalmol*. 2012;26(4):349–56.
31. Graham AD, Kothapalli T, Wang J, et al. A machine learning approach to predicting dry eye-related signs, symptoms and diagnoses from meibography images [published online ahead of print on February 15, 2024]. *Heliyon*. 2024, http://dx.doi.org/10.2139/ssrn.4724519.
32. Lin MC, Wang J, Kothapalli T, Graham AD, Yu S. AI provides deeper understanding of Meibomian gland morphology and function. *Paper presented at the American Academy of Optometry (AAOPT) Annual Meeting 2022*, San Diego, CA, October 26–29, 2022.
33. Lin MC, Graham AD, Kothapalli T, et al. Lifestyle and behaviors: predicting clinical signs and symptoms with machine learning. *Invest Ophthalmol Vis Sci*. 2023;64(8):2880–2880.