AI-based automated Meibomian gland segmentation, classification and reflection correction in infrared Meibography

Ripon Kumar Saha\textsuperscript{a}, A. M. Mahmud Chowdhury\textsuperscript{c}, Kyung-Sun Na\textsuperscript{b}, Gyu Deok Hwang\textsuperscript{b}, Youngsub Eom\textsuperscript{c}, Jaeyoung Kim\textsuperscript{d}, Hae-Gon Jeon\textsuperscript{\textordfiori e}, Ho Sik Hwang\textsuperscript{b}, Euiheon Chung\textsuperscript{\textordfiori ef}

\textsuperscript{a}Department of Biomedical Science and Engineering, Gwangju Institute of Science and Technology, Gwangju, Korea
\textsuperscript{b}Department of Ophthalmology, Yonsei St. Mary’s Hospital, College of Medicine, The Catholic University of Korea, Seoul, Korea
\textsuperscript{\textordfiori c}Department of Ophthalmology, Korea University College of Medicine, Seoul, Korea
\textsuperscript{\textordfiori d}Department of Ophthalmology, Chungnam National University School of Medicine, Daejeon, Korea
\textsuperscript{\textordfiori \textordfiori e}AI graduate school, Gwangju Institute of Science and Technology, Gwangju, Korea

Abstract

\textbf{Purpose:} Develop a deep learning-based automated method to segment meibomian glands (MG) and eyelids, quantitatively analyze the MG area and MG ratio, estimate the meiboscore, and remove specular reflections from infrared images.

\textbf{Methods:} A total of 1600 meibography images were captured in a clinical setting. 1000 images were precisely annotated with multiple revisions by investigators and graded 6 times by meibomian gland dysfunction (MGD) experts. Two deep learning (DL) models were trained separately to segment areas of the MG and eyelid. Those segmentation were used to estimate MG ratio and meiboscores using a classification-based DL model. A generative adversarial network was implemented to remove specular reflections from original images.

\textbf{Results:} The mean ratio of MG calculated by investigator annotation and DL segmentation was consistent 26.23% vs 25.12% in the upper eyelids and 32.34% vs. 32.29% in the lower eyelids, respectively. Our DL model achieved 73.01% accuracy for meiboscore classification on validation set and 59.17% accuracy when tested on images from independent center, compared to 53.44% validation accuracy by MGD experts. The DL-based approach successfully removes reflection from the original MG images without affecting meiboscore grading.

\textbf{Conclusions:} DL with infrared meibography provides a fully automated, fast quantitative evaluation of MG morphology (MG Segmentation, MG area, MG ratio, and meiboscore) which are sufficiently accurate for diagnosing dry eye disease. Also, the DL removes specular reflection from images to be used by ophthalmologists for distraction-free assessment.

1. Introduction

Meibomian gland dysfunction (MGD) is a chronic, diffuse abnormality of the meibomian glands (MG), commonly characterized by terminal duct obstruction and changes in glandular secretion.\textsuperscript{[1]} A diagnosis of MGD requires evaluation of subjective symptoms, morphologic changes in the eyelid margin, meibum secretion, infrared meibography, lipid layer thickness measurement, and other factors.\textsuperscript{[2]} Although examining MG secretion and identifying abnormal eyelid margins are the main methods for clinical evaluation of MGD, meibography has become widely used to assess the morphologic changes of the MG since the introduction of transillumination meibography\textsuperscript{[3]} and non-contact infrared meibography.\textsuperscript{[4, 5]} Non-contact infrared meibography uses infrared light and an infrared camera to provide MG images with reduced discomfort for the patient. After obtaining images, the MG morphology can be assessed, including MG dropout (MG loss) area, width, length, and tortuosity.\textsuperscript{[6]} Among these parameters, the MG dropout area is one of the most critical.\textsuperscript{[4, 7, 8, 9, 10, 11]}

For semi-quantitative analysis, different methods for grading MG dropout have been proposed. Pluflgfelder et al.\textsuperscript{[12]} used a grading system (Grade 0: no gland dropout, Grade 1: <33% gland dropout; Grade 2: 33%–66% gland dropout; Grade 3: >66% gland dropout). Nichols et al.\textsuperscript{[13]} used meibography image grading scales (Grade 1: no partial gland, Grade 2: <25% partial glands; Grade 3: 25%–75% partial glands; Grade 4: >75% partial glands). Arita et al.\textsuperscript{[9]} defined a meiboscore based on 4 grades, Grade 0: no dropout, Grade 1: dropout of less than 1/3, Grade 2: dropout of more than 1/3 and less than 2/3, and Grade 3: dropout of more than 2/3. Pult et al.\textsuperscript{[13]} used degree 0 = no partial glands; 1 = <25% partial glands; 3 = 25%-50% partial glands; 3 = 50%-75%, and 4: >75% partial glands.

For quantitative analysis, some studies measured the MG dropout ratio using image software such as ImageJ.\textsuperscript{[14, 15, 16, 17]} In these studies, however, only the contour of normal MGs except the dropout area was defined and MG dropout ratio was measured. They didn’t measure the area of the individual MGs. Arita et al.\textsuperscript{[9]} objectively evaluated the MG area using custom-developed software with noninvasive meibography. Lid borders were determined, and the MG area was automatically discriminated by enhancing the contrast and reducing image noise. They succeeded in calculating the ratio of the total MG area relative to the total analysis area in the upper and lower eyelids with their software.\textsuperscript{[9]} As described in their discussion, how-
conventional and DL segmentation with MG area comparisons between meiboscores obtained by MGD experts ages from a different test center by several methods, including comparisons between meiboscores obtained by MGD experts and DL segmentation with MG area/ratio.

2. Patients and methods

2.1. Patients

This retrospective study was approved by the Institutional Review Board of Yeouido St. Mary’s Hospital (SC20RISE0063) and Korea University Ansan Hospital (2022AS0015) and adhered to the tenets of the Declaration of Helsinki. The study included 572 eyes of 320 patients who visited Yeouido St. Mary’s Hospital for a dry eye examination.
create images consisting of MG/eyelid information to feed into the DL model (Fig. 2).

The eyelid area was defined vertically by the line of the lid margin to the edge of the tarsal fold. This was somewhat subjective for the lower lid because the superior edge of the fold is not defined by a discrete and clearly demarcated tarsal plate. The eyelid area was defined horizontally by the lacrimal punctum to the lateral canthus. The MG and eyelid area in the upper and lower eyelid were calculated from the processed segmentations. The MG ratio in the upper and lower eyelids was defined by the ratio of the MG area and the corresponding eyelid area. The mean and standard deviation (SD) of the MG area and ratio in the upper and lower eyelids were obtained. To validate the annotation results obtained by the investigators, we calculated the correlation between the MG area/ratio and age, because it is well known that MG dropout increases with age.

2.3. Meiboscore grading by 2 MGD experts

To train/validate the DL model, 2 MGD experts graded meiboscores of 1000 images of the ‘MGD-1K’ according to Arita’s criteria. We used Labelbox for the meiboscore collection. Before sending the images to the MGD experts, all image names and personal identifiers were removed. All 1000 images were graded 3 times for meiboscore by 2 MGD experts separately; thus, 6 times per image (Fig. 1). All images were randomly ordered in each round. There was at least a 1-week interval between each round so that the previous grading would not influence the next round of grading. Also, the 2 MGD experts (H.S.H., K.S.N.) who served as graders were not allowed to share their results with each other. Besides providing meiboscores, we included an option for the graders to select any image as ungradable. A total of 59 images were indicated as ungradable in at least 1 round of grading. Thus, the remaining 941 images with corresponding meiboscore were used to train/validate DL. The mean meiboscore was calculated by averaging meiboscore from all graders. However, we got 139 cases (out of 1,000) with average meiboscore 0.5, 45 cases with meiboscore 1.5, and 8 cases with average meiboscore 2.5, and we have rounded up those mean to the nearest integer meiboscore 1, 2 and 3, respectively.

We examined the variation in the 3 gradings by each MGD expert. Also, to determine the grading consistency of the 2 MGD experts, the correlation of the grades by the 2 MGD experts was calculated using the confusion matrix approach. With this approach, the mean meiboscore assigned by the MGD experts was defined as the closest integer to the average of 3 gradings by the MGD expert. Simple accuracy was calculated based on the same score prediction. Inter- and intra-rater agreement and disagreement were calculated based on kappa scores.

2.4. Deep learning model and training procedure: Segmentation

While it is common to use pre-processed images in conventional deep neural networks, it is now possible to utilize more features from original raw images than processed images with advanced DL. Thus, we used original infrared MG images with investigators’ annotated segmentations to train DL.

We designed an encoder-decoder based DL model (Fig. 2) with skip connections and residual connections (ResNet 34) on the encoder part. We have experimented with different encoders (Original U-Net encoder, ResNet 18, ResNet 34, and ResNet 50 encoder) and found that ResNet 34 was the best fit to handle full-size images and provided high accuracy in case of segmentation. The input and output size were single-color channel grayscale images of resolution 1280 × 640 pixels. We used image augmentation methods for image transformation such as ‘shear’ and ‘horizontal flip up to 15%’ without changing the original input image size. In total, we used 1000 images (800 for training and 200 for validation) paired with the corresponding 1000 images of MG segmentations and 1000 images of eyelid segmentation.

We initially trained our DL model with the 800 input images (385 upper eyelids, 415 lower eyelids) and 800 MG segmentation annotated by investigators. The same architecture was then trained with the original input images with eyelid segmentations. The DL model was made with the PyTorch framework and then trained on a machine with 1 GPU (GEFORCE RTX 2080 Ti) for 200 epochs with the MGD-1K dataset, which took approximately 24 hours. Training time was not a limitation in our case because, for prediction, the model needs to be trained only once.

After our DL model was trained, it could segment out each MG and eyelid from 200 new unseen images (82 upper eyelids, 118 lower eyelids) without any additional help from investigators or MGD experts. First, the MGD experts directly observed the results to check whether there were areas of insufficient or excess markings by DL. Next, MG area and eyelid area were calculated from the MG segmentation and eyelid segmentation. Then MG ratio was calculated from the ratio of MG area and eyelid area.

For validation of the DL model based on MG area/ratio, we compared the mean MG area/ratio obtained by investigators and
that obtained by the DL using a paired t-test (only 91 left eye images). The correlation between the MG area/ratio by DL model and the MG area/ratio by annotation of the investigators was analyzed (linear regression). We further performed a Bland-Altman analysis. To validate the DL output, we analyzed the correlation between the MG area/ratio and age in the upper and lower eyelids (linear regression), because it is well known that MG dropout increases with age.

To validate the DL mode, we analyzed the correlation between the DL segmentation-based MG area/ratio and meiboscore determined by the MGD experts. For 189 meibography images, we assigned the arithmetic mean of all 6 rounds of meiboscores. In this way, we obtained a meiboscore with intermediate values (0, 0.17, 0.33, 0.50, \ldots 2.83, 3.00). We found the correlation between the DL-based MG area/ratio and the arithmetic mean of all 6 rounds of meiboscores.

2.5. Segmentation accuracy and meiboscore prediction

We evaluated the segmentation accuracy of the MG and eyelid area at the pixel level based on Intersection over Union (IoU). The IoU measures how much the DL-predicted binary images correspond to the ground truth images by dividing the IoU of these 2 images. Similarly, the IoU can be used to measure the match between 2 investigators. The principle of the IoU accuracy matrix in our system is represented in (Supplement Fig. 1). We tried to determine the performance of our DL model compared with that of the investigators. For this, all 200 images of the validation dataset were labeled twice by investigators. We computed the accuracy of the DL by taking the IoU between the ‘DL-predicted images’ and ‘images annotated by any 2 investigators’ (for each 200 validation images, the investigator was randomly selected). Investigator accuracy was determined by calculating the IoU of each marked image pair of the validation set. The mean accuracy (IoU) of the investigators was calculated by taking the mean of all accuracies.

The meiboscore prediction part of the model was based on the classification model. Three images were used as input: the original image, segmented MG, segmented eyelid (obtained as output from the segmentation model), and the meiboscore was provided as input. Using this classification architecture, it is also possible to estimate meiboscore directly from original images without those segmentation steps (MG/eyelid segmentation image). However, adding those MG and eyelid segmentation provide more insight to the classification architecture and significantly improves meiboscore prediction. The meiboscore output was given a grade of 0, 1, 2, or 3. To compare the meiboscore prediction accuracy of the DL model, we used the confusion matrix. Mean meiboscore by 2 MGD experts (ground truth) was defined as the closest integer to the mean of the 6 gradings by the MGD experts. Meiboscore prediction accuracy was calculated based on the same prediction score.

Besides 1,000 images, another dataset of 600 images from Korea University Ansan Hospital was used to test the meiboscore prediction part of the deep learning model. Another 2 MGD experts (Y.S.E, J.Y.K.) graded meiboscores of 600 images according to Arita’s criteria. All 600 images were labeled 3 times for meiboscore by 2 MGD experts separately; thus, 6 times per image.

2.6. Correction of specular reflections by the generative adversarial network model

Specular reflection is a distracting factor for MGD experts when diagnosing MG from original images. Specular reflection often creates confusion for MG assessment because of the white appearance with saturated pixels. So, we found that besides providing important parameters like MG area/ratio and meiboscore, it is also helpful to remove specular reflection from original images for MGD experts. Different DL-based image restoration model have emerged emerged. [38][39][40] We also applied the DL approach to correct specular reflections from MG images. Specular reflections are usually represented by a high light intensity, i.e., a high pixel value, in the images. Based on the principle of global thresholding, we detected specular reflection parts with realistic pixel combinations.
The GAN architecture comprises a generator and a discriminator component. A variation of the model was implemented to fill the detected specular reflection component of the MG images with possible gland/eyelid portions based on an understanding of surrounding information provided by the whole image. This model takes 2 images as input (original image where the reflection portion was filled by a solid color and a specular reflection mask image) and provides 1 output layer—the predicted images without reflection (Fig. 3). The model was pre-trained with different types of publicly available image datasets to gain the knowledge to restore any missing parts of those images. The generator fills the specular reflection component, and the discriminator verifies the final output repeatedly until the output reflection-free image achieves the best algorithmic performance. This approach remove specular reflection and fill that with understanding of image to make meibography images less distracting to be graded by ophthalmologist. A paired t-test was calculated based on the 109 validation images (only left eye) to represent the difference in the MG area/ratio predicted from a DL algorithm with original images as input vs. specular reflection-corrected images as input.

For statistical analysis, GraphPad Prism 8.4.2 (GraphPad Corp., San Diego, CA) was used. A p-value of less than 0.05 was considered significant.

3. Results

3.1. Patients and dataset of MGD-1K

The demographic information of the patients included in the study is presented in Table 1. The mean (SD) age of the 320 patients was 54.6 (20.2) years with 107 men and 213 women (Supplement Fig. 2).

3.2. MG annotation by investigators and meiboscore grading by 2 MGD experts

It took approximately 12–25 minutes for the investigators to precisely annotate each MG image with Photoshop. The mean MG area was 33.80 ± 9.48 mm² in the upper eyelid and 32.04 ± 10.87 mm² in the lower eyelid, and the mean MG ratio was 26.23 ± 5.83% in the upper eyelid and 32.34 ± 7.45% in the lower eyelid. There was a significant negative correlation between the MG area and age in the upper (p < 0.0001) and lower eyelids (p = 0.0002), and between the MG ratio and age in the upper (p < 0.0001) and lower eyelids (p < 0.0001); (Supplement Fig. 3).

The scoring was notably inconsistent within and between the 2 MGD experts (Fig. 4). The color-coded full data representation revealed all grades of 0, 1, 2, and 3 for the same meibography image. Intergrading accuracy between MGD Expert 1 and 2 are represented by kappa score of 0.20 representing slight agreement. Moreover, the intra-grading accuracies of MGD Expert 1 and 2 were Kappa scores of 0.71 and 0.54, respectively.
We found a 53.63% inter-grading similarity between the MGD experts 3 and 4 (Fig. 5).

### 3.3. DL model and training: Segmentation

After the DL-based segmentation model was trained with the input images, any grayscale MG infrared image could be transformed into images containing the segmented labels of the MG or eyelid in 0.3s. MGD experts checked the results to determine whether there were areas of insufficient or excess markings, and found no obvious incorrectly marked areas in any of the 1000 images. A representative example of the original infrared image and images overlaid by the investigator’s annotation and DL segmented output for the upper and lower eyelid is shown in (Fig. 6).

A comparison of MG Area/ratio by investigators and DL-based model is shown with paired t-test based on left eyes in (Table 2). In the upper eyelids, a paired t-test showed no significant difference between the annotated results obtained by the investigators and the DL results (Table 2). In the upper-lower eyelid, the DL-based MG area/MG ratio was slightly underestimated overestimated compared with the investigator annotated MG area/MG ratio ($p = 0.048$). The MG area and MG ratio of both eyelids obtained by DL and the investigators had a good positive correlation ($p < 0.0001$, Fig. 7). A Bland-Altman plot representing the MG area (mm$^2$) and MG ratio (%) calculated from the MG area by investigators and by DL is shown in (Fig. 8). The plot shows that the 95% confidence interval of the differences in the MG area and MG ratio in the upper and lower eyelids represented no positive or negative bias. Regarding the effect of age on the MG area and MG ratio, there was a significant negative correlation of the DL-based and investigator-based MG areas and MG ratios in the upper and lower eyelids with respect to age (Fig. 9). As the arithmetic mean of all 6 rounds of meiboscores increased, the MG area and MG ratio measured by investigators and DL segmentation in the upper and lower eyelids decreased (Fig. 10). In all 4 cases, the investigator and DL showed similar tendencies with similar linear regression lines.

Figure 5: Meiboscore intergrading accuracy and accuracy of the deep learning. The X-axis and Y-axis represent the category of the meiboscore by the 2 meibomian gland dysfunction (MGD) experts (left side) or deep learning and the MGD expert (right side). The inside values represent the match within those categories. The diagonal line represents the same category matched between the X- and Y-axes.

We found a 53.63% inter-grading similarity between the MGD experts 3 and 4 (Fig. 5).

Figure 6: Annotation and prediction. Investigator's segmentation and deep learning output segmentation results are shown for both meibomian glands and eyelids. The area is also shown based on the segmentation. The red transparent color represents the annotation by investigators, and the blue color represents the output segmentation by the deep learning model.

Figure 7: Meibomian gland (MG) area of investigator vs. deep learning model. Correlation of the MG area (mm$^2$) and MG ratio (%) calculated from the annotations of the MG area by investigators and deep learning.
Figure 8: Bland-Altman plot representing the meibomian gland (MG) area (mm²) and MG ratio (%) calculated from the annotations of the MG area by investigators and deep learning. The 95% confidence interval of the differences in the MG area and MG ratio in the upper and lower eyelids of patients was -13.56 to 9.371 and -9.187 to 8.044 in the MG area, respectively, and -11.07 to 8.930 and -9.073 to 8.651 in the MG ratio, respectively. The confidence interval does not reveal any positive or negative bias.

Figure 9: Correlation of the meibomian gland (MG) area/MG ratio with patient age. The relation is represented for both the upper eyelid and lower eyelid by investigators and deep learning. The correlation demonstrates that older people more commonly have MG dysfunction. Usually, the area estimation of deep learning is slightly lower compared to the investigator by a certain threshold.

Figure 10: Graph representing the correlation of mean meiboscore from MGD experts and MG area (collected from investigators and deep learning). Investigators and deep learning show a similar correlation in the lower eyelid, but the output area from deep learning is slightly lower for the upper eyelid.

Supplementary Figure 4: Meiboscore intergrading accuracy and accuracy of the deep learning. The X-axis and Y-axis represent the category of the meiboscore by the meibomian gland dysfunction (MGD) experts 1 and 2 (left side) or deep learning and the MGD expert (right side). The inside values represent the match within those categories. The diagonal line represents the same category matched between the X- and Y-axes.

3.4. Segmentation accuracy and meiboscore prediction

The accuracy of the investigators was computed by taking the IoU between the 2 investigators. The accuracy of the investigators was 64.03% (Fig. 11). With the 200 validation images, the DL model’s accuracy (mean IoU score) was 67.63% for the MG output.

Regarding meiboscore prediction, we found a 73.01% agreement between the DL-predicted meiboscore and the mean of 6 gradings by MGD experts 1 and 2 (Supplement Fig. 5). The intergrading agreement between MGD expert 1 and 2 was 53.44%. However, when the pre-trained model was used to predict meiboscore from an independent dataset of 600 images with meiboscore from MGD expert 3,4, the agreement was not as high as the original dataset (Fig. 5). The intergrading agreement between MGD expert 3 and 4 was 53.67%. (Supplement Fig. 5) shows deviation of MGD experts and deep learning prediction of Meiboscore.
Figure 11: Segmentation accuracy of investigators and deep learning model. The accuracy (IoU) distribution of 50 MG images from 200 images of the validation set is displayed on the left side. In most cases, the deep learning model showed better accuracy segmenting the MG. The relative segmentation accuracy of the deep learning model was determined by representing the difference of Accuracy of Model (IoU) and Accuracy of investigators (IoU) for all 200 validation images (represented on the right side). All 200 validation images were labeled twice by investigator comparison. The ophthalmologists had 64.03% agreement between those 2 segmented sets (SD 6.453%). When the deep learning model (trained on different images) predicted segmentation images of those 200 unseen images, there was a 67.63% match between those predicted images and images marked by the ophthalmologists (SD: 6.635%). In 167 of 200 images, the machine learning algorithm had higher accuracy segmenting those glands compared with the investigators.

Supplementary Figure 5: Deviation of meiboscore prediction from deep learning compared to MGD experts. The Y-Axis represent percentage of 4 meiboscores by each MGD experts or deep learning. The deep learning model was only trained on 800 images with meiboscore from MGD Expert No 1 and 2 and tested on both validation images (from MGD Expert 1,2) and on 600 images from an independent center (from MGD Expert 3,4) without retraining again. We see that the meiboscores from different MGD experts were not consistent and high variability can be seen with meiboscore 0, 2 and 3. Deep learning seems to predict more number of meiboscore 1 compared to 0,2,3.

3.5. Correction of specular reflection region based on DL

The DL-based approach was able to remove reflection regions from the original MG images and fill that based on the understanding of the whole image. The output seems natural (Fig. 12). Also, we didn’t lose any image details when using this specular reflection process. When we tested original images and reflection-removed images to find the MG area/ratio, there was no significant difference in the MG area (p = 0.689) and MG ratio (p = 0.255) between the segmented output from the original and reflection-free MG images (Table 3). This justifies that ML-based reflection removal provides not only natural-looking images but also reliable prediction results without deviation.

4. Discussion

In present study, we demonstrated that DL can automatically detect all individual MGs and quantify the MG area and MG area ratio. Other studies have evaluated methods to quantify MG dropout in meibography.[14, 15, 16, 17, 18, 20, 21, 22] Among various types of commercial equipment, Idra (SBM Systemi, Orbassano, Italy) shows meibography and the quantified dropout ratio. Recently, Vigo et al.[44] evaluated the diagnostic performance of Idra, including measurement of MG loss, but they did not describe how they quantified and validated MG loss. We could find no other literature referencing the Idra system.

This study demonstrated the capability of DL to automatically detect all individual MGs and calculate the MG area and MG ratio. When quantifying MG dropout, it may be more meaningful to calculate the whole area of individual MGs rather than to calculate on the basis of a drawn contour of the MG area without dropout, as done in previous papers.[14, 15, 16, 17] For example, in (Fig. 13), Comparison of the meibomian gland (MG) area and the MG ratio between contoured MG and individual MG analyses in atrophic MG (A) and normal MG (B). The MG area of atrophic MGs and normal MGs calculated by contoured MG analysis was 53.1 mm2 and 89.3 mm2, respectively. The MG area of atrophic MGs and normal MGs calculated by individual MG analysis was 22.9 mm2 and 48.5 mm2.
MG ratio in several ways. Sixth, we have evaluated the
meiboscore in this study. Until now, in the semi-quantitative grading systems, a grade of 0 is defined as no gland dropout. By definition, the proportion of those eyelids with a meiboscore of 0, or no dropout, may not exist in a quantitative grading approach. Therefore, it might be better to designate the intervals equally with grade 1: <25%, grade 2: 25% & <50%, grade 3: >50% & <75%, and grade 4: >75%.

Regarding the meiboscore grading by experts, the consistency was evaluated by 4 MGD experts in 3 rounds for the same 941 meibography images (Fig. 4) as well as from another different test center. Comparison of the intraobserver grades revealed an inter-grading agreement of only 53.67% (Fig. 5). Previous studies described similar inconsistencies and weak intergrading correlation in meibography. On the other hand, the DL-based meiboscore prediction achieved an accuracy of 73.01% on images annotated by MGD Expert 1.2 and 59.17% on images annotated by MGD Experts of different test center.

In this study, the DL performance was validated by various methods. First, MGD experts directly observed the results to check whether there were areas of insufficient or excess markings. Second, the mean DL-based MG area and MG ratio were compared with those measured by investigators. The correlation between the MG area/MG ratio by the DL model and the MG area/MG ratio by annotation of the investigators was analyzed (linear regression) (Fig. 7). Also, Bland-Altman analysis was performed (Fig. 9). Third, the correlation between the MG area/MG ratio and age was analyzed by the investigator and DL results (Fig. 9). Fourth, the correlation between the MG area/MG ratio with the meiboscopes determined by the MGD experts and DL segmentation was compared (Fig. 10). The findings of each of these methods indicated a good DL performance.

Furthermore, for the first time, we applied DL to correct specular reflection from the meibography images. In previous automated detection systems, specular reflections were mistaken as an MG. Instead of excluding all reflections, our DL model could restore reflection regions without noticeable artifacts (Fig. 12), resulting in specular reflection-free images that retained all gland information for easy and distraction-free images for clinicians. We found no significant differences in the MG area/MG ratio due to specular reflection removal (Table 3), implying minimal data loss.

Compared with other studies, the present study has the following advantages. First, we used 1600 images. This was possible only because DL segmentation was completely automated and not manual. Second, unlike previous studies, the analysis was conducted not only on the upper eyelid but also on the lower eyelid. Third, our designed DL model could produce a segmentation mask in 0.3 second directly from original images without any further input instruction and works for both the upper and lower eyelids. Fourth, because images from Lipiview, a widely used apparatus, are used, they can be commercialized as is if our newly developed software is installed on the apparatus. It is also good for use by other researchers with the same kind of data. Fifth, validation was performed for DL-segmented MG area/MG ratio in several ways. Sixth, we have evaluated the

respectively. The actual MG area calculated by individual MG analysis was much smaller than the area calculated by contour MG analysis. Furthermore, the MG area of normal MGs (89.3 mm2) was only 1.68-fold larger than that of atrophic MGs (53.1 mm2) by contour MG analysis. The MG area of normal MGs (89.3 mm2) was only 1.62-fold larger than that of atrophic MGs (89.3 mm2) by individual MG analysis. Therefore, the individual MG analysis is more accurate and better represents MG atrophy than contour MG analysis.

We measured both the MG area and MG ratio in this study. When calculating the MG ratio, the MG area should be divided by the area where the MG should normally be (i.e., eyelid area). In the upper eyelid, the fold line of the everted eyelid has a definite tarsal plate, so the denominator of the ratio is definite. In the lower eyelid, however, it is difficult to determine exactly where the MG should normally be found. Therefore, the MG ratio is somewhat inaccurate for the lower eyelid. So, especially in the lower eyelid, we should measure not only the MG ratio but also the MG area itself.

For each meibography image, the DL generated a meiboscore prediction based on the 6 meiboscopes by the 2 MGD experts. In fact, because the DL-segmented MG ratio was calculated, the meiboscore could have been predicted according to Arita’s criteria with the MG ratio result. When grading with the DL MG ratio, however, the proportion of those with a meiboscore of 0 (no dropout) is almost 0% by definition. Due to this disad-
model based on different dataset of 600 labeled images. Seventh, we also implemented a correction of the specular reflection in the MG images using a GAN-based DL model.

In conclusion, this study describes an automated method for quantifying individual MGs to identify the MG area, eyelid area, MG ratio, and classify the meiboscore in non-contact infrared images of upper and lower eyelids using a specialized method comprising multiple DL blocks. This fully automated DL approach provides better accuracy in terms of segmentation and prediction of the meiboscore. Further, we created the MGD-1K dataset, which is fully public (https://mgd1k.github.io), as the first image-based dataset on meibomian glands with precise eyelid segmentations, gland segmentation, and meiboscores for open research on meibomian gland dysfunction.

**Financial disclosures**

The authors have no financial interest or relationship to disclose.

**Declaration of competing interest**

The authors declare no conflicts of interest.

**Acknowledgments**

The work was supported by the GIST Research Institute (GRI) research collaboration grant funded by GIST in 2022; The Brain Research Program through the N.R.F. funded by the Ministry of Science, I.C.T. & Future Planning (NRF-2017M3C7A1044964), and the Korea Medical Device Development Fund grant funded by the Korea government, the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety (Project Number: 1711138096, KMDF_PR_220200901_0076) to E.C.; the 2022 Joint Research Project of Institutes of Science and Technology to E.C.; a grant from the National Research Foundation of Korea (N.R.F.) funded by the Korean government (MEST) (NRF-2019R1A2C2086003) to E.C.; the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health & Welfare Republic of Korea (grant number: HI17C0659, HI17C0659) (H.S.H.); and the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Republic of Korea (No. 2020R1A2C1005009) (H.S.H.).

**References**

[1] J. D. Nelson, J. Shimazaki, J. M. Benitez-del Castillo, J. P. Craig, J. P. McCulley, S. Den, G. N. Foulks, The international workshop on meibomian gland dysfunction: report of the definition and classification subcommittee, Invest Ophthalmol Vis Sci 52 (4) (2011) 1930–7.

[2] G. N. Foulks, A. J. Bron, Meibomian gland dysfunction: a clinical scheme for description, diagnosis, classification, and grading, Ocul Surf 1 (3) (2003) 107–26.

[3] J. V. Jester, L. Rife, D. Nii, J. K. Luttrull, L. Wilson, R. E. Smith, In vivo biomicroscopy and photography of meibomian glands in a rabbit model of meibomian gland dysfunction, Invest Ophthalmol Vis Sci 22 (5) (1982) 660–7.

[4] R. Arita, K. Inoh, K. Inoue, S. Amano, Noncontact infrared meibography to document age-related changes of the meibomian glands in a normal population, Ophthalmology 115 (5) (2008) 911–5.

[5] T. Celik, H. K. Lee, A. Petznick, L. Tong, Bioskage informatics approach to automated meibomian gland analysis in infrared images of meibography, Journal of optometry 6 (4) (2013) 194–204.

[6] Y. W. Koh, T. Celik, H. K. Lee, A. Petznick, L. Tong, Detection of meibomian glands and classification of meibography images, J Biomed Opt 17 (8) (2012) 086006.

[7] R. Arita, J. Suehiro, T. Haraguchi, R. Shirakawa, H. Tokoro, S. Amano, Objective image analysis of the meibomian gland area, Br J Ophthalmol 98 (6) (2014) 746–55.

[8] H. Pult, B. H. Riede-Pult, J. J. Nichols, Relation between upper and lower lids’ meibomian gland morphology, tear film, and dry eye, Optom Vis Sci 89 (3) (2012) E110–5.

[9] S. C. Pfliugfelder, S. C. Tseng, O. Sanabria, H. Kell, C. G. Garcia, C. Felix, W. Feuer, B. L. Reis, Evaluation of subjective assessments and objective diagnostic tests for diagnosing tear-film disorders known to cause ocular irritation, Cornea 17 (1) (1998) 38–56.

[10] H. Pult, J. J. Nichols, A review of meibography, Optometry and Vision Science 89 (5) (2012) E760–E769.

[11] Y. Eom, J.-S. Lee, S.-Y. Kang, H. M. Kim, J.-S. Song, Correlation between quantitative measurements of tear film lipid layer thickness and meibomian gland loss in patients with obstructive meibomian gland dysfunction and normal controls, American journal of ophthalmology 155 (6) (2013) 1104–1110.

[12] W. Ngo, S. Srinivasan, M. Schulze, L. Jones, Repeatability of grading meibomian gland dropout using two infrared systems, Optom Vis Sci 91 (6) (2014) 658–67.

[13] H. Pult, B. Riede-Pult, Comparison of subjective grading and objective assessment in meibography, Cont Lens Anterior Eye 36 (1) (2013) 22–7.

[14] Y. Wu, H. Li, Y. Tang, X. Yan, Morphological evaluation of meibomian glands in children and adolescents using noncontact infrared meibography, J Pediatr Ophthalmol Strabismus 54 (2) (2017) 78–83.

[15] C. Llorens-Quintana, P. Syga, D. R. Iskander, Automated image processing algorithm for infrared meibography, in: Imaging Systems and Applications, Optical Society of America, 2018, pp. IM3B–3.

[16] C. Llorens-Quintana, L. Rico-del Viejo, P. Syga, D. Madrid-Costa, D. R. Iskander, A novel automated approach for infrared-based assessment of meibomian gland morphology, Translational vision science & technology 8 (4) (2019) 17–17.

[17] J. Wang, T. N. Yeh, K. Chakraborty, X. Y. Stella, M. C. Lin, A deep learning approach for meibomian gland atrophy evaluation in meibography images, Translational vision science & technology 8 (6) (2019) 37–37.

[18] Q. Dai, X. Liu, X. Lin, Y. Fu, C. Chen, X. Yu, Z. Zhang, T. Li, M. Liu, W. Yang, et al., A novel meibomian gland morphology analytic system based on a convolutional neural network, IEEE Access 9 (2021) 23083–23094.

[19] A. J. Bron, V. E. Evans, J. A. Smith, Grading of cornel and conjunctival staining in the context of other dry eye tests, Cornea 22 (7) (2003) 640–50.

[20] S. L. Maskin, W. R. Testa, Infrared video meibography of lower lid meibomian glands shows easily distorted glands: Implications for longitudinal assessment of atrophy or growth using lower lid meibography, Cornea 37 (10) (2018) 1279–1286.

[21] Labelbox, “labelbox,” online, 2021. [online]. available: https://labelbox.com.

[22] D. J. Hand, R. J. Till, A simple generalisation of the area under the roc curve for multiple class classification problems, Machine Learning 45 (2) (2001) 171–186.

[23] R. F. Mansour, Deep-learning-based automatic computer-aided diagnosis system for diabetic retinopathy, Biomedical engineering letters 8 (1) (2018) 41–57.

[24] P. Hansson, Fracture analysis of adhesive joints using the finite element
method, Lund Inst. Technol. 60 (February) (2002) 84–90.

[26] K. Armanious, V. Kumar, S. Abdulatif, T. Hepp, S. Galdis, B. Yang, ipa-medgan: Inpainting of arbitrary regions in medical imaging, in: 2020 IEEE International Conference on Image Processing (ICIP), IEEE, 2020, pp. 3005–3009.

[27] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[28] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[29] A. Mikolajczyk, M. Grochowski, Data augmentation for improving deep learning in image classification problem, in: 2018 international interdisciplinary PhD workshop (IIPhDW), IEEE, 2018, pp. 117–122.

[30] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch (2017).

[31] T. Falk, D. Mai, R. Bensch, Ö. Çiçek, A. Abdulkadir, Y. Marrakchi, A. Böhm, J. Deubner, Z. Jäckel, K. Seiwald, et al., U-net: deep learning for cell counting, detection, and morphometry, Nature methods 16 (1) (2019) 67–70.

[32] W. Xiong, J. Yu, Z. Lin, J. Yang, X. Lu, C. Barnes, J. Luo, Foreground-aware image inpainting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5840–5848.

[33] E. Ntavelis, A. Romero, S. Bigdeli, R. Timofte, Z. Hui, X. Wang, X. Gao, C. Shin, T. Kim, H. Son, et al., Aim 2020 challenge on image extreme inpainting (2020) 716–741.

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.

[35] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, M. Ebrahimi, Edgeconnect: Structure guided image inpainting using edge prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0.

[36] L. Vigo, M. Pellegrini, F. Bernabei, F. Carones, V. Scarcia, G. Giannacare, Diagnostic performance of a novel noninvasive workup in the setting of dry eye disease, J Ophthalmol 2020 (2020) 5804123.

[37] M. Garza-Leon, A. Gonzalez-Dibildox, N. Ramos-Betancourt, E. Hernandez-Quintela, Comparison of meibomian gland loss area measurements between two computer programs and intra-inter-observer agreement, Int Ophthalmol 40 (5) (2020) 1261–1267.

[38] M. A. K. Setu, J. Horstmann, M. E. Stern, P. Steven, Deep learning-based automatic meibomian gland segmentation and morphology assessment in infrared meibography, Scientific reports 11 (1) (2021) 1–11.

[39] M. L. McHugh, Interrater reliability: the kappa statistic, Biochemia medica 22 (3) (2012) 276–282.

[40] A. von Ahrentschildt, L. Haneberg, M. L. Robich, L. A. Jones-Jordan, S. Marx, W. Sickenberger, D. R. Powell, J. T. Kwan, S. Wong, S. Srinivasan, Morphological characteristics of meibomian glands and their influence on dry eye disease in contact lens wearers, The Ocular Surface 24 (2022) 93–99.

[41] E. Daniel, M. Pistilli, G.-s. Ying, V. Y. Bunya, M. Massaro-Giordano, P. A. Asbell, M. G. Maguire, Association of meibomian gland morphology with symptoms and signs of dry eye disease in the dry eye assessment and management (dream) study, The ocular surface 18 (4) (2020) 761–769.

[42] K. Ciezar, M. Pochylski, 2d fourier transform for global analysis and classification of meibomian gland images, The Ocular Surface 18 (4) (2020) 865–870.

[43] M. Ha, J. S. Kim, S.-Y. Hong, D.-J. Chang, W.-J. Whang, K.-S. Na, E. C. Kim, H. S. Kim, H. S. Hwang, Relationship between eyelid margin irregularity and meibomian gland dropout, The Ocular Surface 19 (2021) 31–37.