Feasibility Assessment of a Novel Deep Learning Tool for automatic Evaluation of Meibomian Glands in Three Ocular Imaging Devices

PATRYCJA PIWOWARCZYK [*1], DOROTA SZCZĘSNA-ISKANDER [*2]

[1] *Wrocław University of Science and Technology, Faculty of Fundamental Problems of Technology, Department of Biomedical Engineering, Wybrzeże Stanisława Wyspiańskego 27, 50-370 Wrocław*

[2] *Wrocław University of Science and Technology, Faculty of Fundamental Problems of Technology, Department of Optics and Photonics, Wybrzeże Stanisława Wyspiańskego 27, 50-370 Wrocław*

[*] *Corresponding author: patrycja.piwowarczyk@pwr.edu.pl*

A deep learning tool for automated segmentation of meibomian glands (MG) was validated on three devices: Oculus Keratograph 5M (trained), Medmont Meridia, and Topcon slit lamp. Among corrected images, manual adjustment was needed in 13–15% of upper eyelids and 10–24% of lower eyelids, most frequently for Topcon. Agreement in assessing MG morphology between two eyelid eversion conditions was better for Oculus and Medmont images but lower for Topcon. The automated tool reliably evaluates MG morphology, but manual verification remains important regardless of the device.

Keywords: Meibomian glands automatic assessment, meibography, artificial intelligence

INTRODUCTION

Meibomian dysfunction (MGD) is a leading cause of dry eye disease (DED) [1]. Reliable assessment of meibomian glands (MG) morphology is crucial in the diagnosis and management of MGD [2]. Recent advances in artificial intelligence have enabled the development of automated tools for MG morphology analysis, offering quantitative assessment and potential to improve objectivity in clinical practice [3]. The primary aim of this study was to verify the performance of a novel deep learning–based tool for analyzing MG imaged by three different devices: the Oculus Keratograph 5M (K5M), on which the tool was trained, the Keratoscope Medmont Meridia (MD) and the Topcon slit lamp, SL-D701 (TPC). The secondary aim was to assess whether different eyelid eversion conditions affect MG segmentation, and if they introduce systematic variability across imaging devices.

METHODS

Retrospective analysis included images acquired with three devices was performed, including data from 54 patients (median age 64 ± 15 years) with a clinically healthy ocular surface or mild DED according to current diagnostic guidelines [1]. Available clinical data included subjective symptoms evaluated using the OSDI (76% of participants) or/and the DEQ-5 (43% of participants) questionnaires, noninvasive tear breakup time (NIBUT) assessed by K5M, MG loss graded with Arita scale [4], and tear film lipid layer thickness (LLT) graded using scale by Romeseiro et al. [5]. For each patient, two high-quality images of the upper and lower eyelids (UE; LE) were taken, suitable for automated analysis, meaning they were in focus, well-centered, and fully everted. Consequently, incomplete patient overlap resulted in the inclusion of data from 39 patients per device. The deep learning tool was used for the automatic segmentation of the tarsus and MG morphological parameters extraction. For each gland, its length and width were reported, and for each eyelid, the mean length and width of all included MG and the percentage loss of MG as the ratio of the measured area occupied by the glands to the tarsal area, were calculated [3]. All 234 UE and 234 LE images were processed. Verification of automated annotations and manual corrections were performed.

Segmentation errors were classified into the following categories listed in Table 1. Their examples are shown in Figure 1. After corrections, MG parameters were reanalysed and statistics were performed separately for each device. Among the corrected images comparisons of above parameters before and after manual correction were performed. For each eyelid, two images acquired under different eyelid eversion conditions were paired. Agreement between the model's raw annotations was assessed for each pair using the intraclass correlation coefficient (ICC). After amendments were made, the agreement was re-evaluated.

RESULTS

The mean score in DEQ-5 was 5.1 ± 3.2, and in OSDI was 15.5 ± 17.6. The mean NIBUT was 11.8 ± 6.0 s. The mean grade in MG loss was 1 ± 1 in both eyelids, and in LLT was 2 ± 1. The proportion of images requiring correction differed depending on the device. For UE, correction was required in 13% (K5M), 15% (MD), 13% (TPC) of images, for LE in 10% (K5M), 12% (MD), and 24% (TPC).

Table 1. Summary of corrections. Numbers indicate how many images were affected by each error type.

	Eyelid	K5M	MD	TPC
Total number of analysed images	Upper	78	78	78
	Lower	78	78	78
Total number of images requiring corrections from the following category (parentheses denote the percentage of total images)	Upper	10 (~13%)	12 (~15%)	10(~13%)
	Lower	8 (~10%)	9 (~12%)	18 (~24%)
a) major errors in tarsus annotation (parentheses denote the number of images in which the error affected accurate gland)	Upper	1 (1)	2 (1)	7 (1)
	Lower	1 (1)	0	14 (10)
b) omission of an entire gland or a substantial part of it, or false-positive gland annotation	Upper	6	9	1
	Lower	8	9	4
c) merging of non-collinear and spatially separated glands	Upper	1	0	0
	Lower	0	0	0
d) merging of two clearly distinct, crossed glands into a single X-, Y-, or V-shaped structure	Upper	1	2	1
	Lower	0	0	0
e) incorrect annotation of tortuous glands as thickened, amorphous areas	Upper	1	3	1
	Lower	0	0	0

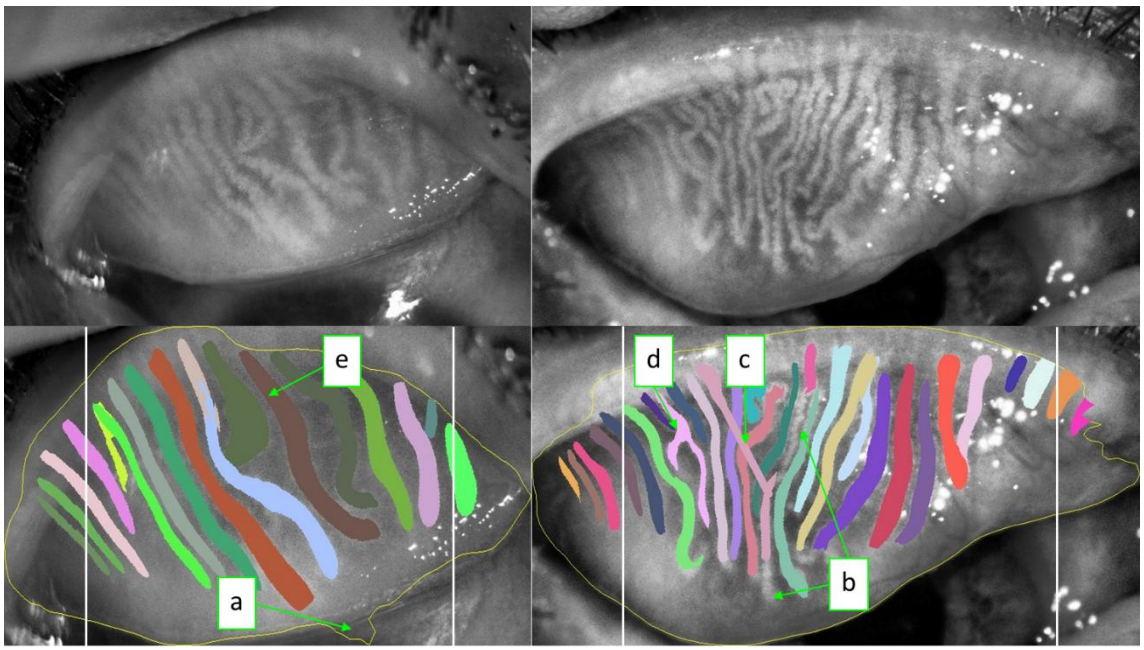

Figure 1. Examples of annotation errors within the area of analysis marked by white lines.

In the corrected images from TPC, the predominant issue, particularly in the LE, was a) (~75% cases). In K5M and MD, the most common error was b) (~80% cases). Errors c), d) e) occurred only in rare cases of UE images across all devices. Although statistically significant differences before and after

corrections were observed for UE images in length in K5M ($p = 0.043$) and in width in MD ($p = 0.004$), the magnitude of these differences was small (mean diff. ± SD: 0.25 ± 0.28 mm and 0.014 ± 0.013 mm, respectively) and therefore considered clinically insignificant. Among corrected images of the LE, significant differences were observed in the number of glands for the K5M ($p = 0.043$; mean diff. ± SD: 0.75 ± 0.66) and TPC ($p = 0.001$; mean diff. ± SD: 1.6 ± 1.6), as well as in the percentage of MG loss for all three devices: K5M ($p = 0.018$; mean diff. ± SD: 8.1 ± 7.1%), MD ($p = 0.003$; mean diff. ± SD: 4.9 ± 2.5%), TPC ($p = 0.005$; mean diff. ± SD: 14.7 ± 18.2%). The agreement of MG raw segmentation between two different UE eversion conditions in TPC and MD was good for MG number, length, width and percentage loss (ICC > 0.75), in K5M it was good for width (ICC = 0.76) and percentage loss (ICC = 0.81), whereas it was moderate for MG number (ICC = 0.58, mean diff. ± SD: 2.8 ± 2.1) and length (ICC=0.66; mean diff. ± SD: 0.47 ± 0.41mm). For LE, overall agreement was lower across all devices. K5M had the lowest agreement overall, with ICCs ranging from 0.44 (length) to 0.60 (width). MD performed slightly better, with ICCs ranging from 0.49 (width) to 0.82 (length), reflecting moderate to good agreement. TPC, while performing moderately for most parameters with ICCs between 0.57 (number) to 0.76 (length), showed the lowest agreement for percentage loss (ICC = 0.43). However, differences in length and width between eversions were ≤ 0.5 mm and clinically negligible for all devices. The mean difference for MG number in K5M and TPC was about 2 ± 2, in MD 1 ± 1. The mean difference for percentage loss in K5M and TPC was about 10 ± 10%, in MD 6 ± 8%. Agreement between eversions after amendments changed significantly only for LE length in TPC, remaining at a moderate level (ICC = 0.63).

CONCLUSIONS

For UE images, automated analysis achieved approximately 85–90% correct MG segmentations on all devices, and manual corrections were clinically insignificant. For LE, major corrections were needed only for a small proportion of atypical glands in K5M and MD. In contrast, a higher proportion of TPC images required correction, highlighting the need for careful verification of automated annotations. This is likely due to uneven target illumination in the TPC, which causes underexposure of peripheral areas, which reduces the accuracy of tarsus and MG segmentation. Consequently, raw segmentations in the LE showed the lowest agreement between eyelid eversions for percentage loss, but this also resulted in reduced agreement when the optometrist corrected the MG. The impact of eyelid eversion conditions on MG characterization was generally minor for UE. Due to the smaller surface of LE tarsus, even small changes result in substantial effect on the percentage loss. Overall, the deep learning tool reliably evaluates MG morphology across devices, although manual verification remains important to ensure accuracy in more challenging cases.

Funding: The study was supported by the National Science Center, Poland, SONATA BIS grant 2021/42/E/ST7/00345.

REFERENCES

[1] Wolffsohn JS, et al. TFOS DEWS III: Diagnostic methodology. *Am J Ophthalmol*. 2025;279:387–450.
[2] Jones L, et al. TFOS DEWS III: Management and therapy. *Am J Ophthalmol*. 2025;279:289–386.
[3] Paściak A, et al. Enhancing meibography-based assessment of gland morphology by utilizing an image-rotating Mask R-CNN approach. *Biomed Signal Process Control*. 2025;109:108045.
[4] Arita R, et al. Proposed algorithm for management of meibomian gland dysfunction based on noninvasive meibography. *J Clin Med*. 2020;10(1):65.
[5] Remeseiro B, et al. A methodology for improving tear film lipid layer classification. *IEEE J Biomed Health Inform*. 2014;18(4):1485–1493.
[6] Muzyka-Woźniak M, et al. Diagnostic agreement between invasive and non-invasive assessments for preoperative dry eye disease screening in patients with cataracts. *Ophthalmol Ther*. 2026;1–16.