Appendix

In this supplementary material, we report the details of the data augmentations used to train the U-Net based mire segmentation, show more heatmap examples for qualitative comparison between SmartKC, Keratron and *SmartKC++*, and report the standard deviations of the experiments shown in Tables 1, 3 and 5.

Details on Data Augmentation used for training U-Net

Table 6 shows the details of the augmentations used in training the U-Net based Mire Segmentation model on the labels generated by the fingerprint detection algorithm [12]

More examples on qualitative improvements of heatmaps

Figure 3 shows more examples for qualitative comparison of heatmaps generated by SmartKC [9] (left), Keratron (middle) and *SmartKC++* (right). Although the SmartKC++ heatmap still looks slightly different from the ground truth Keratron heatmaps, based on feedback from three ophthalmologists, SmartKC++ heatmaps are much more similar to Keratron heatmaps compared to SmartKC heatmaps, that will eventually result in more accurate diagnosis by the ophthalmologists.

Standard Deviations of Tables 1, 3 and 5

Tables 7, 8 and 9 report the standard deviation of the results reported in Tables 1, 3 and 5, respectively. Standard deviations are reported for multiple runs/ splits of the Phase-1 success/failure sets.

For Table 7, Agreement between the predicted Sim-K values and those obtained from Keratron are shown. Standard deviation is calculated from 5 separate runs. Each run consists of a separate train/test split from the success set, and results in a different segmentation model. Results on the success set are on the 5 test splits, as mentioned earlier. The results on the failure set are on a single test set, hence have only 1 run for SmartKC. For SmartKC++, the standard deviation on the failure set is calculated for the 5 different segmentation models trained on the 5 splits, as mentioned earlier.

For Table 8, accuracy of automated diagnosis obtained by thresholding the Sim-K values. Standard deviation is calculated from 5 separate runs. Each run consists of a separate train/test split from the success set, and results in a different segmentation model. Results from thresholding Keratron Sim-K values are from a single run on the combined dataset. Results on the success set are on the 5 test splits, as mentioned earlier. The results on the failure set are on a single test set, hence have only 1 run for SmartKC. For SmartKC++, the standard deviation on the failure set is calculated for the 5 different segmentation models trained on the 5 splits, as mentioned earlier.

For Table 9, results not involving the U-Net based mire segmentation have only 1 run, whereas the U-Net based mire segmentation has 5 runs

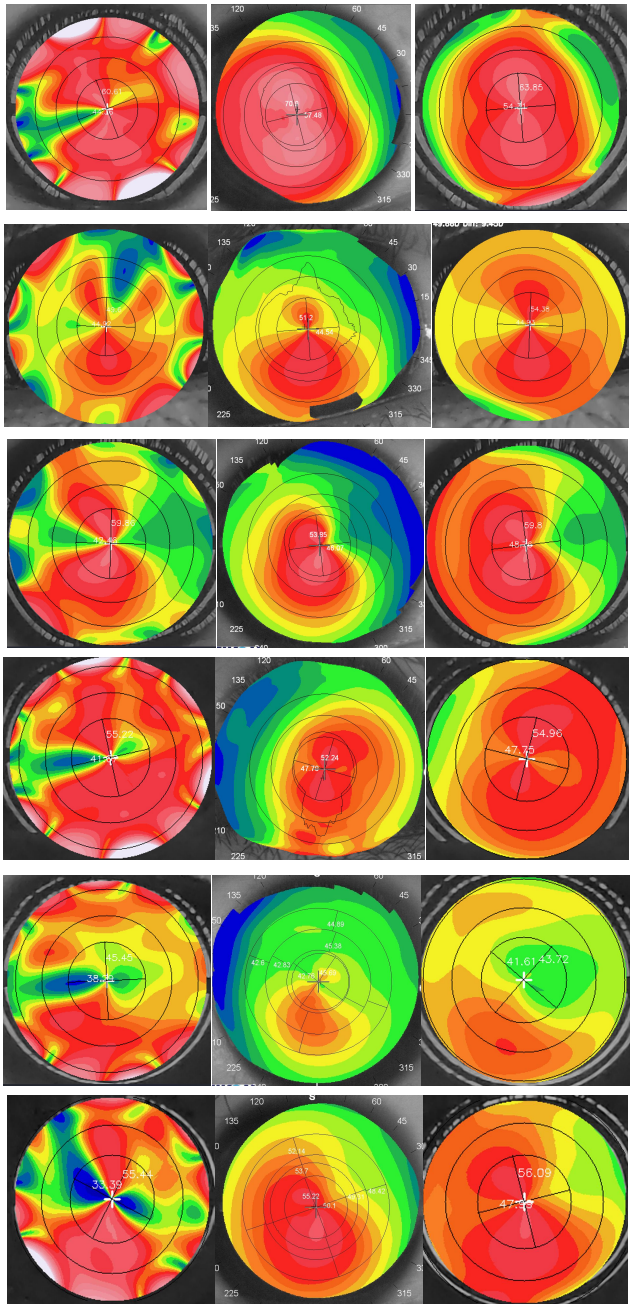

Figure 3. Qualitative comparision of heatmaps generated by SmartKC (left), Keratron (middle) and SmartKC++(right)

Type of Augmentation	Specific Augmentation Applied
Additive Noise	Gaussian, Laplacian, Poisson, Salt, Pepper and Salt+Pepper
Image Sharpness	GaussianBlur, GammaContrast, LogContrast, SigmoidContrast and LinearContrast
Spatial Augmentations	Warp, Rotate, Scale, Translate and Flip

Table 6. Different types of data augmentations applied when training the U-Net. All 3 types of augmentation are applied simultaneously, with the specific choice made randomly at each epoch

		Sim-K1			Sim-K2		
		MAE	**MAPE**	**Corr.**	**MAE**	**MAPE**	**Corr.**
Success Set	**SmartKC**	0.322	0.536	0.076	0.287	0.341	0.147
	SmartKC++ (ours)	0.287	0.542	0.090	0.178	0.347	0.151
Failure Set	**SmartKC**	-	-	-	-	-	-
	SmartKC++ (ours)	0.054	0.105	0.001	0.100	0.223	0.007

Table 7. Standard Deviation of Sim-K prediction on Phase-1 dataset. Agreement between the predicted Sim-K values and those obtained from Keratron are shown. Standard deviation is calculated from 5 separate runs. Each run consists of a separate train/test split from the success set, and results in a different segmentation model. Results on the success set are on the 5 test splits, as mentioned earlier. The results on the failure set are on a single test set, hence have only 1 run for SmartKC. For SmartKC++, the standard deviation on the failure set is calculated for the 5 different segmentation models trained on the 5 splits, as mentioned earlier.

Dataset	**Device**	**KT Condition**	**Acc.**	**Sens.**	**Spec.**	**Prec.**	**Recall**	**F1**
Combined	**Keratron**	K1 >49.995 or K1 - K2 >1.523	-	-	-	-	-	-
Success Set	**SmartKC**	K1 >44.55 or K1 - K2 >2.644	1.82	3.33	2.14	5.07	3.33	3.46
	SmartKC++ (ours)		1.02	0.001	1.36	4.08	0.001	3.74
Failure Set	**SmartKC**	K1 >44.55 or K1 - K2 >2.644	-	-	-	-	-	-
	SmartKC++ (ours)		4.52	4.99	6.67	4.06	4.99	3.74

Table 8. Standard Deviation of Automated diagnosis on Phase-1 dataset: Accuracy of automated diagnosis obtained by thresholding the Sim-K values. Standard deviation is calculated from 5 separate runs. Each run consists of a separate train/test split from the success set, and results in a different segmentation model. Results from thresholding Keratron SimK values are from a single run on the combined dataset. Results on the success set are on the 5 test splits, as mentioned earlier. The results on the failure set are on a single test set, hence have only 1 run for SmartKC. For SmartKC++, the standard deviation on the failure set is calculated for the 5 different segmentation models trained on different splits.

UbMS	**CbML**	**Acc.**	**F1**	**MAE K1**	**MAPE K1**	**MAE K2**	**MAPE K2**
✗	✗	-	-	-	-	-	-
✓	✗	2.857	3.00	0.157	0.306	0.062	0.102
✗	✓	-	-	-	-	-	-
✓	✓	4.517	3.738	0.054	0.105	0.100	0.223

Table 9. Standard Deviation of *SmartKC++* with and without various components on the Phase-1 failure dataset. UbMS: U-Net based Mire Segmentation, CbML: Clustering based Mire Localization. There is only 1 failure set - hence, results not involving the U-Net based mire segmentation have only 1 run, whereas the U-Net based mire segmentation has 5 runs