S5. Dimensionality reduction and image categorization

Using our approach each Meibomian image can be described by 30 independent features. Finding regularities in such a high-dimension data can be difficult and requires proper data analysis procedure. To get more important information out of the whole set of data we firstly used Principal Component Analysis (PCA) [1,2]. This method consists of projecting each 30-dimensional data point onto new axes (principal components) while preserving as much of the data’s variation as possible. Because in most cases only the first few principal components contain most of the variance in data, the method is commonly used for dimensionality reduction.

In Fig. S6 the result of PCA analysis is shown. Scree plot in panel a) shows the variances explained by first 10 principal components. It is clear that focusing only on the first two PCA components is enough to explain nearly 50% of all variation in the whole set of data. The correlation plot between the first two PCA components is presented on panel b) for images subjectively classified as healthy, intermediate and unhealthy. Here it is worth to note that PCA belongs to the unsupervised class of machine learning algorithms meaning that the analysis does not require prior categorization of data set. The fact that the data points for each category seems separated in PCA1(PCA2) representation suggest that this kind of analysis can be used as an objective classification routine. In panel c) the PCA loading (score) vectors, for each of 30 objective features, are presented in PCA1(PCA2) representation. Projection of a loading vector for a given feature on a particular PCA axis indicates significance of this feature on this principal component (Fig. S6c).

All 30 features sorted according to their PCA1 loadings are given in Table S1. As follows from the data collected in tab.S1, an entropies and the means of distributions of intrinsic properties sensitive to variability in gland frequency and orientation ($\sigma_q$, $\sigma_0$, $C_0$, $G_q$) are the features with the most descriptive potential.

![Fig.S6](image)

*Fig.S6* a) Scree plot indicating how much variance in whole set of 30-dimentional data is explained by each of ten first PCA components (red dots). Blue squares shows the cumulative variance. Notice that first two PCA components are responsible for almost 50% of variance in whole data set. b) Correlation plot between two first PCA components. Notice that although PCA is unsupervised method (it does not make use of the a-priori knowledge about the existence of categories) it allows to separate the categories along PCA components. c) PCA score (loading) plot. Each of the 30 feature has corresponding vector (solid line) whose projection on particular PCA axis indicates its influence on this principal component value. 10 features of the highest projection on PCA1 axis are indicated with color lines and corresponding symbols.

Tab. S1 All 30 features sorted according to the highest weight they have on PCA1 component. Feature symbol includes name of distribution measure followed by an inherent quantity given in parenthesis. Two features indicated in red ($\text{Var}(\mathbf{\theta})$ and $M(q_0)$) correspond to global features (periodicity anisotropy parameter ($\kappa^{-1}$) and mean gland frequency ($\tilde{q}$), respectively) derived in our earlier work [3].
Differences in the values of the descriptive features found for each Meibomian image can be used to automatically categorize the images into subjective categories. There are many machine learning algorithms which can be utilized for this task: k-means clustering (unsupervised), support vector machine, neural networks, to mention just the most popular ones [3,4]. Here we test the performance of two widely used approaches based on PCA analysis and Linear Discriminant Analysis (LDA). Both methods are similar in a sense that they reduce dimensionality of data by projecting them on new axes. The difference is that PCA is an unsupervised method and it does not make use of the a-priori knowledge about the existence of categories. LDA on the other hand belongs to supervised class of learning algorithms and uses information about classes.

On Fig. S7 we show the result of PCA (panel a) and LDA (panel b) analysis performed on the same 30-dimensional set of data collected for all Meibomian images. After analysis the dimensionality of data can be reduced. We present our data as a correlation plot of the first two
components (PCA\(_{1,2}\) and LDA\(_{1,2}\)). It is clear that both methods allow to separate the categories along their components plane. A simple comparison of the results presented in the Fig.S6 shows that the LDA analysis allows for a better separation of data for three subjective categories which is not surprising considering that LDA is supervised method.

To use the data dispersion observed in main panels of Fig.S7 for automatic image categorization, the marginal distributions of each of the component (PCA\(_{1,2}\) and LDA\(_{1,2}\)) for each subjective category was first calculated. These distributions are shown as histograms on Fig.S7. It can be observed that for each type of analysis used, the maxima of the distribution appears in different location and shows some overlap. Next, the experimental histograms were described by normal distribution (Gaussian) to get continuous probability distribution of each component for each category (solid lines in marginals of Fig.S7). Knowing the probability distributions of a given component for each category, a simple threshold classifier were created.

We tested the classifier based on the probability that an image is parametrized by certain value of first component (e.g. PCA\(_1\)) AND certain value of the second one (e.g. PCA\(_2\)). More specifically, an image being parametrized with a pair of component values (PCA\(_1/PCA_2\) or LDA\(_1/LDA_2\)) is assigned to a category specified by the highest value of the product of corresponding probability functions (\(p(PCA_1)p(PCA_2)\) or \(p(LDA_1)p(LDA_2)\)). The classification performance of this approach is presented in Table 1 of the main manuscript. The symbols of these classifiers are: \(p(PCA_1)p(PCA_2)\) and \(p(LDA_1)p(LDA_2)\).

Notice that \(p(LDA_1)\) distributions for images classified as healthy and unhealthy barely overlap. This means that, based solely on \(p(LDA_1)\), meaning that images belonging to these two categories can be distinguish with nearly 100% accuracy. Notice also that \(p(LDA_2)\) for intermediate category is clearly moved away from probability distributions for other two categories (being close together). This information can be used to increase the effectiveness of automatic classification of images belonging to intermediate category.

Fig.S7 Correlation plots between first two components of a) Principal Component Analysis (PCA) and b) Linear Discriminant Analysis (LDA) for images classified as healthy, intermediate and unhealthy. Marginal plots shows probability distributions of corresponding components. Notice that although both analysis provide noticeable separation of categories, the LDA analysis (being supervised method) provides much better clustering of classes.
[1] Jolliffe IT. Principal Component Analysis. 2nd ed. New York: Springer-Verlag; 2002. https://doi.org/10.1007/b98835.

[2] Martínez A. M., Kak A. C. PCA versus LDA. IEEE Trans. Pattern Anal. Mach. Intell. 2001, 23(2), 228-233. https://doi.org/10.1109/34.908974.

[3] Ciężar K., Pochylski M. 2D fourier transform for global analysis and classification of meibomian gland images. Ocul Surf 2020, 18(4), 865-870. https://doi.org/10.1016/j.jtos.2020.09.005.

[4] Celik T, Lee HK, Petznick A. Bioimage informatics approach to automated meibomian gland analysis in infrared images of meibography. J Optom. 2013; 6: 194–204. https://doi.org/10.1016/j.optom.2013.09.001.