

ARTICLE OPEN



Assessment of skin barrier function using skin images with topological data analysis

Keita Koseki^{1,2,3,7}, Hiroshi Kawasaki^{1,2,4,7}, Toru Atsugi^{1,5}, Miki Nakanishi^{1,5}, Makoto Mizuno^{1,5}, Eiji Naru^{1,5}, Tamotsu Ebihara^{1,2}, Masayuki Amagai^{2,4} and Eiryo Kawakami^{1,2,6}✉

Recent developments of molecular biology have revealed diverse mechanisms of skin diseases, and precision medicine considering these mechanisms requires the frequent objective evaluation of skin phenotypes. Transepidermal water loss (TEWL) is commonly used for evaluating skin barrier function; however, direct measurement of TEWL is time-consuming and is not convenient for daily clinical practice. Here, we propose a new skin barrier assessment method using skin images with topological data analysis (TDA). TDA enabled efficient identification of structural features from a skin image taken by a microscope. These features reflected the regularity of the skin texture. We found a significant correlation between the topological features and TEWL. Moreover, using the features as input, we trained machine-learning models to predict TEWL and obtained good accuracy ($R^2 = 0.524$). Our results suggest that assessment of skin barrier function by topological image analysis is promising.

npj Systems Biology and Applications (2020)6:40; <https://doi.org/10.1038/s41540-020-00160-8>

INTRODUCTION

The skin provides an effective barrier between the external environment and the body, preventing the entry of pathogens and microorganisms, as well as restricting water loss¹. Skin barrier research has gained huge momentum after the discovery of the filaggrin mutation (*FLG*) in atopic dermatitis (AD) patients². Filaggrin is an epidermal structural protein critical for skin barrier formation, and the *FLG* mutation is a major risk factor for AD³. Studies have identified an association between the *FLG* mutation and asthma and food allergies, such as peanut allergies, even in the absence of AD³. Accumulating evidence, including the results of murine studies, suggests that an altered skin barrier is involved in atopic diseases, including AD^{4,5}.

Transepidermal water loss (TEWL) is the amount of water that evaporates from the body surface and is most widely used to evaluate skin barrier function⁶. Since healthy skins have the capacity of retaining water, high and low TEWL are indicative of skin barrier dysfunction and intact skin or recovered skin barrier, respectively. Recent studies suggest that TEWL not only reflects the current state of the skin barrier, but it is also a subclinical biomarker for AD and food allergy^{7,8}. Since TEWL is sensitive to environmental factors, such as temperature and humidity, examinees are required to wait in the test environment, where temperature and humidity are controlled for a certain period of time (~20 min) before measurement⁶. Therefore, direct measurement of TEWL is not easily implemented as a method to estimate skin barrier function in daily clinical practice, and more practical alternative methods are necessary.

Instead of the direct measurement of TEWL, we considered assessing the skin barrier function by image analysis of the skin surface. Since it is simple to take microscopic pictures of the skin surface, skin barrier assessment by image analysis would be beneficial in daily clinical practice and in subclinical skin care of healthy people. Recently, convolutional neural networks (CNNs) have been reported

as very effective in the field of medical image analysis for extracting important features and predicting clinical characteristics⁹. However, CNNs require a vast number of images, as well as significant computational resources and fine-tuning of parameters. In addition, the learning process for interpretation of extracted features is rather difficult¹⁰. Therefore, we applied topological data analysis (TDA) instead to extract features representing the shape of skin surfaces. TDA is a collection of methods for identifying topological structures in data¹¹ and is now considered to be an effective tool to analyze various data in many areas including material science¹², engineering¹³, and biology¹⁴. Moreover, TDA has also been applied in medicine to the quantification of tumor shapes^{15–17}, finding patterns in genetic data of cancer patients¹⁸, and characterizing brain artery networks¹⁹. In dermatology, TDA has been applied to segmenting and classifying skin lesions^{20–23} and quantifying the connectivity of epidermal cells²⁴. TDA detects the number of topological features, such as connected components, holes, and cavities, and demonstrates their robustness and magnitude. This information facilitates quantification of the shape and regularity of the skin surface. A study that examined 350 healthy adult women showed that there is a significant difference between populations with high and low TEWL in terms of the number of skin ridges²⁵. These findings suggest that structural features in the skin surface contain information associated with skin barrier function and supports our notion that skin barrier function may be assessed from skin images.

In this study, we propose a new skin barrier assessment method using skin images with TDA. We extracted features representing the regularity of skin surface patterns from images of 244 healthy people and predicted their TEWL using machine learning using identified features as predictive variables. We found that the assessment of skin barrier function from skin images using TDA holds promise in improving the accuracy of TEWL prediction.

¹Medical Sciences Innovation Hub Program, RIKEN, Yokohama, Kanagawa 230-0045, Japan. ²Department of Dermatology, Keio University School of Medicine, Shinjuku-ku 160-0016 Tokyo, Japan. ³School of Medicine, Yokohama City University, Yokohama 236-0004 Kanagawa, Japan. ⁴Laboratory for Skin Homeostasis, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045 Kanagawa, Japan. ⁵Dermatology and Cosmeceuticals Sec, KOSÉ Corporation, Kita-ku 114-0005 Tokyo, Japan. ⁶Artificial Intelligence Medicine, Graduate School of Medicine, Chiba University, Chiba 260-8670 Chiba, Japan. ⁷These authors contributed equally: Keita Koseki, Hiroshi Kawasaki. ✉email: eiryo.kawakami@riken.jp

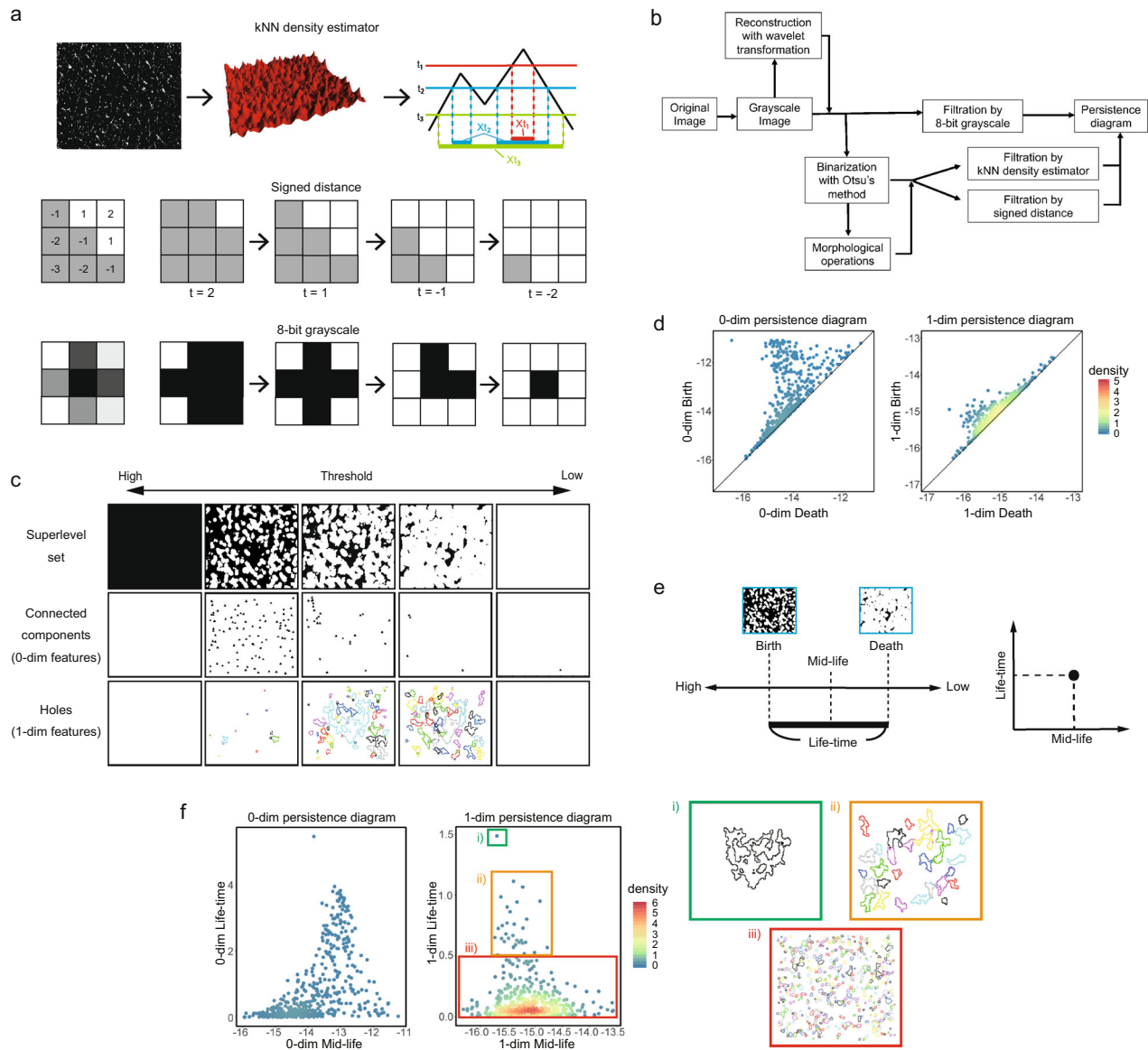


Fig. 1 An illustration of feature extraction using TDA. **a** Three filtration functions are illustrated. In the upper row, the density of white pixels is calculated using the kNN density estimator, and elevated regions in the 3D map represent a higher density. As the threshold decreases from t_1 to t_3 , the superlevel set (the domain above the threshold) grows from X_{t_1} to X_{t_3} . In the middle and lower rows, the signed distance and the 8-bit grayscale are, respectively, used as filtration functions. As the threshold decreases, the superlevel set (white region) expands. **b** The flowchart of our method. An original image is transformed into a grayscale image. For the kNN density estimator or signed distance as the filtration function, the image is binarized with Otsu's method. Wavelet transformation and morphological operations can be applied. Following TDA, the persistence diagram is obtained. **c** A schematic diagram of thresholds and superlevel sets. In the upper row, as the threshold decreases, the superlevel set (white region) grows larger. The middle row shows the birthplace of connected components (0-dim topological features), and the lower row shows the holes (1-dim topological features) in the corresponding superlevel set. **d** The information on the log-scale thresholds where the feature appears ("birth") and disappears ("death") is summarized in two diagrams called the persistence diagrams. The axes represent the log-scale thresholds where the feature appears ("birth") and disappears ("death"). The 0-dim persistence diagram represents information of the connected components, and the 1-dim persistence diagram represents that of holes. The color scale represents the density of points in the persistence diagram calculated by the function "kde2d" of the R package MASS. **e** For interpretability and handiness, we transform "birth" and "death" into their mean (mid-life) and their difference (life-time) and plot them. **f** The persistence diagrams with the axes mid-life and life-time. The color scale represents the density of points in the persistence diagram. The three small panels on the right show the holes in each corresponding domain of the 1-dim persistence diagram. The life-time roughly corresponds to the size of holes.

RESULTS

Evaluating skin patterns with TDA

Several TDA algorithms have been proposed in terms of thresholding methods called filtration functions, including the k -nearest neighbor (kNN) density estimator, the signed distance, and the 8-bit grayscale value (Fig. 1a). The kNN density estimator calculates the density of white pixels by measuring the distance from the

considered pixel to the k th nearest white pixel for a fixed integer k ²⁶. The signed distance method assigns the Manhattan distance from the border between black and white areas with a positive sign to a white pixel and a negative sign to a black pixel^{27,28}. The 8-bit grayscale value reflects the brightness of a pixel ranging from 0 (black) to 255 (white)¹⁶. For each threshold value, the superlevel set is defined as the image region where the value of the filtration

function is larger than the threshold. TDA analyzes how the shape of the superlevel set changes with a gradually changing threshold.

According to the choice of filtration functions, there are several options for image preprocessing (Fig. 1b). Here, we illustrate the TDA procedure for the case using the kNN density estimator as the filtration function. First, we transformed the images into grayscale and binarized them with Otsu's method²⁹. Wavelet transformation and morphological operations can be applied to remove brightness disproportion and noise.

After preprocessing, we quantified patterns of the images using TDA. There are two kinds of topological features in 2D image analysis: connected components and holes. Connected components (0-dim topological features) are continuously connected regions of white pixels. Holes are continuous loops through a white region surrounding a black region. In our procedure, each filtration function assigns large values to hollow spots, such as sulci cutis, and small values to protuberant spots, such as cristae cutis. Intuitively, holes represent sulci surrounding cristae, and connected components represent connected regions made of sulci. As the threshold decreases, the superlevel set spreads gradually (Fig. 1c, upper row). Connected components appear and merge, and finally, there is only one connected component (Fig. 1c, middle row). Likewise, as the threshold decreases, holes appear, fill up, and finally, there is no hole (Fig. 1c, lower row). We recorded the thresholds in log-scale at the point where the connected components and holes appeared and disappeared, described as *birth* and *death*, respectively.

The popular way to express this extracted information is drawing *persistence diagrams* which show the relationship between birth and death (Fig. 1d)³⁰. We plotted the means of birth and death (*mid-life*) and the difference between birth and death (*life-time*) of each topological feature instead of directly plotting birth and death (Fig. 1e, f) since they were easy to interpret. The mid-life indicates the threshold at which the feature exists. Features with a large life-time can be regarded as important structures because it is likely that these are not noise³¹. Also, the life-time of holes roughly correlates with the size of corresponding holes (Fig. 1f, right three panels). On the other hand, the distribution of mid-life is related to the regularity of the image. Since features visible in the original image appear at the early stage with high mid-life, if the image has a regular ring-shaped structure, many holes appear convergently at the range of high mid-life.

Relationships between TEWL and persistence diagrams

We compared the distributions of mid-life and life-time on 0-dim and 1-dim topological features. For illustration, we show six typical samples of cheek skin images (Fig. 2a). We found clear differences in their persistence diagrams. In particular, the distribution of 1-dim mid-life for case A had a sharper peak in the higher value range than that for case E (Fig. 2b). Because case A had a regular texture, holes appeared intensively at the range of high mid-life. On the other hand, case E showed a wavy pattern and very little texture, and holes appeared loosely at the range of low mid-life.

To investigate the correlations between TEWL and topological features, we performed linear regression analysis. Although linear regression only captures linear relationships between variables, it enables us to intuitively understand whether two variables have a significant relationship, how closely they are correlated, and whether the correlation is positive or negative³². As the comparison of case A and case E suggested, there was a strong positive correlation between TEWL and the standard deviation of 1-dim mid-life (Fig. 2c, d) and strong negative correlation between TEWL and the mean of 1-dim mid-life (Fig. 2c, e). These correlations suggest that TDA can detect regularity of skin texture patterns which are associated with skin barrier integrity. Interestingly, the correlations between TEWL and some topological

features were stronger than those of background factors, such as age and sex, and environmental factors, such as temperature and humidity. This indicates that skin images include substantial information associated with skin barrier function. The mean and standard deviation had an inverse linear relationship, and a clear trend for TEWL changes could be seen along the regression line (Fig. 2f), suggesting that this line may be considered an indicator of the regularity of skin surfaces. From these results, we concluded that the topological features extracted from skin images using TDA provide essential information associated with skin barrier function.

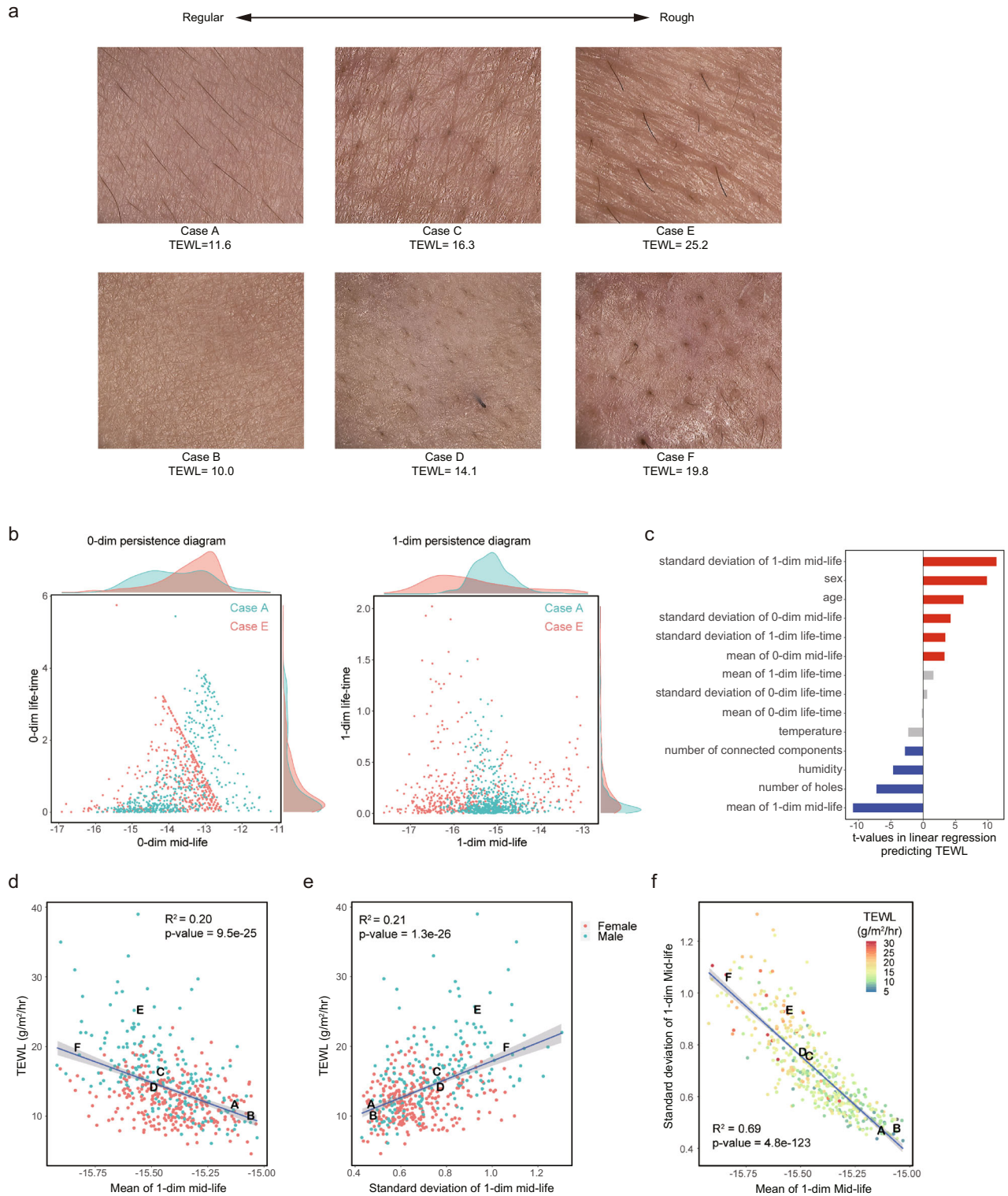
We also analyzed the relationships between the moisture content of the stratum corneum, and the features extracted from skin images (Supplementary Fig. 1a, b). The moisture content was measured using two devices, namely, the Corneometer (Model CM825; Courage & Khazaka Electronic, Cologne, Germany) and the Skicon (Model 200EX-USB; YAYOI, Tokyo, Japan). We compared these two devices owing to differences in their mechanisms. The Corneometer uses electrical capacitance, whereas the Skicon uses high-frequency conductance to assess the level of hydration^{33,34}. Reports suggest that the Skicon is more sensitive to hydration dynamics compared with the Corneometer; however, the latter is useful for the measurement of very dry skins³⁵. In contrast to TEWL, the moisture content was not strongly related to skin images with both devices. Instead, the moisture content was more associated with the environmental condition.

Prediction of TEWL with machine learning

In the previous section, we showed that topological features on skin images have a significant correlation with TEWL. Given this, we used the features in the persistence diagrams to predict TEWL with machine-learning algorithms. The process of prediction is shown in Fig. 3a. First, for cross-validation, we partitioned all images into 70% training data (997 images of 170 people) and 30% test data (431 images of 74 people). We investigated the relationships between TEWL and summarized values (means and standard deviations of the mid-life and life-time) of persistence diagrams; however, these values omit a significant amount of information that may have important relationships with skin conditions. Since we cannot apply machine-learning algorithms directly to persistence diagrams due to the space of persistence diagrams lacking the vector space structure (e.g., each persistence diagram has a different number of points), we had to vectorize the persistence diagrams before applying machine-learning algorithms.

Various vectorization methods for persistence diagrams have been proposed. For example, the persistence landscape embeds persistence diagrams into a Banach space made of piecewise-linear functions³⁶; kernel methods were used to apply kernel-based machine-learning methods and statistical concepts to persistence diagrams^{37,38}; functions defined using tropical geometry were also utilized to represent persistence diagrams without loss of information³⁹. Among many vectorization methods, we applied the *persistence image* to vectorize persistence diagrams⁴⁰ because the persistence image embeds each persistence diagram into a finite-dimensional Euclidean space that is easy to handle. Besides, another approach exists that directly combines TDA and neural networks⁴¹. The combination of the persistence image and machine learning enabled us to easily analyze which regions of persistence diagrams were important in the prediction and where important features (connected components and holes) resided in the original images²⁷.

After vectorization of the persistence diagrams, we combined each generated vector with age, sex, temperature, and humidity to make the feature vector for each subject and applied several machine-learning methods to construct TEWL prediction models. Since high-dimensional data that is composed of highly correlated variables often causes inefficient prediction due to



multicollinearity and overfitting, we applied a principal component analysis (PCA) to the features extracted from images and extracted the most important components whose proportions of variance were larger than 0.01. Among the several machine-learning algorithms and linear regression model, the random forest regression model with PCA was the best prediction model according to the coefficient of determination (R^2) (Supplementary Table 1). Furthermore, the choices of several filtration functions,

preprocessing methods, and vectorization methods have been considered and summarized in Supplementary Tables 2 and 3. The preprocessing procedure with the best performance was (1) signed distance as the filtration function, (2) persistence image with a standard deviation of 0.1 as the vectorization method, and (3) without wavelet transformation or morphological operations in the preprocessing. With this procedure, the random forest regression model predicted TEWL of the test data with high

Fig. 2 Relationships between TEWL and features of skin images extracted using the kNN density estimator. **a** Typical examples of skin images. Cases A and B show relatively regular skin texture and low TEWL, and cases E and F show relatively rough skin surfaces and high TEWL. The written consent was obtained for publication of the photographs. **b** The persistence diagrams of cases A and E with marginal distributions. It is remarkable that the mid-life of case A has a sharper peak with a higher value range than that of case E. **c** The t -value of each variable is calculated using simple linear regression predicting TEWL. As explanatory variables, we used the mean and standard deviation of mid-life and life-time for 0-dim and 1-dim topological features, the number of all connected components and holes, age, and sex. Variables with a false discovery rate (FDR) larger than 0.01 are colored red (if its t -value is positive) or blue (if its t -value is negative). **d, e** The results of simple linear regression predicting TEWL from the mean and the standard deviation of 1-dim mid-life with the coefficient of determination (R^2) and p -value. Since sex is the most important variable in determining TEWL other than the topological features of the images, each point is colored red or blue according to the subject's sex. The points representing cases A–F are also labeled. The shaded areas represent the 95% confidence intervals of the regression. **f** The mean and the standard deviation of 1-dim mid-life are linearly related, and a clear observed TEWL trend change can be seen along the regression line. The shaded area represents the 95% confidence interval of the regression.

accuracy ($R^2 = 0.524$; Fig. 3b). Combining the three feature vectors obtained by the three filtration functions (each under the optimal combination of other methods) did not improve the prediction accuracy ($R^2 = 0.512$). This is likely because all three filtration functions extract quite similar features of the skin image and, therefore, the information obtained using each method overlaps.

The variable importance of each region in the persistence diagrams was calculated using the random forest regression model (Fig. 3c). The regions that contributed significantly to the regression of TEWL had high importance. For three subjects, we extracted only those features in the regions of highest importance and drew their locations over the original images (Fig. 3d). The locations of connected components were represented by their birth positions (i.e., the locations where they appear), and those of holes were represented by their death positions (i.e., the locations where they are filled in)²⁷. We confirmed that large structures such as pores and hairs were not included in the important features.

We also predicted the moisture content of the stratum corneum with the random forest regression model from the same variables as for TEWL (Supplementary Fig. 1c, d). The choices of several filtration functions, preprocessing methods, and vectorization methods were considered and are summarized in Supplementary Tables 4 and 5. Although the accuracy was lower than the prediction of TEWL, the moisture content can also be predicted with $R^2 = 0.219$ for the Corneometer and $R^2 = 0.364$ for the Skicon. This is consistent with the linear regression results, where the moisture content measured with the Skicon had a stronger association with the topological features and also with the environmental condition.

DISCUSSION

In this study, we predicted TEWL from skin images with good accuracy by combining TDA and machine learning. An advantage of this method is that it takes far less time than the direct measurement of TEWL. In contrast to the direct measurement of TEWL, which requires subjects to wait in the test environment with controlled temperature and humidity for about 20 min, our method requires only a short time for taking a picture of the subject and a few seconds for analyzing the image. Therefore, we believe that our method can be applied in daily clinical practice and to the skin care of healthy people. Another advantage of our method is that it requires a relatively small number of images. In this study, we used only 997 images of 170 subjects to train the prediction model. Typically studies using CNNs in dermatology use 10,000–100,000 skin images to train models^{42–44}, which is sometimes problematic in clinical image analysis where not so many images of the same standard are available for training. TDA extracts predefined essential information on the skin surface structures and does not require learning for feature extraction. Because of this characteristic, our method based on TDA can be easily transferred to new projects which utilize images of different standards. Feature extraction methods such as TDA are suited for

situations where a small number of images is available and clear features exist; conversely, deep learning is suited for situations where many images are available and defining specific features is difficult. Therefore, it is essential to use deep-learning and non-learning feature extraction methods in parallel when dealing with medical images of various types and sizes.

Our results showed that TDA can quantify the regularity of skin surfaces, which correlates with TEWL. A previous study also applied TDA to detect the patterns of the microsurface structure of the gastrointestinal tract; images were classified according to their patterns into three groups with variable risk for cancer (oval, tubular, and irregular patterns with no, low, and high risk, respectively). Approximately 90% of the classification matches were performed by medical doctors¹⁵. Moreover, in cardiac image analysis, a study applying TDA to computed tomography images successfully extracted the shape of the trabeculae, the fine muscle columns on the ventricular walls which had been missed by previous methods⁴⁵. These results and ours suggest that TDA is suitable for the image analysis of organs with fine structures. Since there are many organs with fine structures in the body, such as the lung, liver, and brain, there are many potential applications of TDA in the field of medicine. In dermatology, several studies applied TDA to the malignancy classification of melanomas using skin images taken by dermatoscopes or stereomicroscopes^{20–22}. Another study proposed a method of applying TDA to classify seven skin diseases including melanomas and basal cell carcinomas²³. These studies have shown that skin image analysis using TDA is useful for the qualitative assessment of skin diseases. Our application of TDA, on the other hand, allowed us to quantitatively evaluate the skin phenotype. This will lead not only to the stratification of existing skin diseases but also to the prediction of pathological changes in chronic skin diseases such as AD. Quantification of skin structures using TDA may lead to the establishment of objective diagnostic criteria for skin diseases. In this study, the moisture content of the stratum corneum did not correlate strongly with topological features of the skin. This is probably because the moisture content reflects the state of the deeper layer of the stratum corneum, which does not appear in the surface⁴⁶. The Raman microspectrometer (Model 3510; River Diagnostics BV, Rotterdam, The Netherlands) has been used to determine molecular concentration profiles in the deep skin^{47,48}. Combining topological features of the skin surface with such deep skin information may allow for more detailed skin condition monitoring and stratification.

Historically, dermatology has evolved through the observation of body surfaces by specialists⁴⁹. However, recent developments in molecular biology and genomic sciences have revealed mechanisms and causative genes of skin diseases, and many biologics have appeared⁴⁹. For precision medicine considering these mechanisms, it is necessary to objectively evaluate the skin condition and systematically select treatment that is suitable for each individual patient. Although the present study is exclusively designed for healthy people, the methodology may be

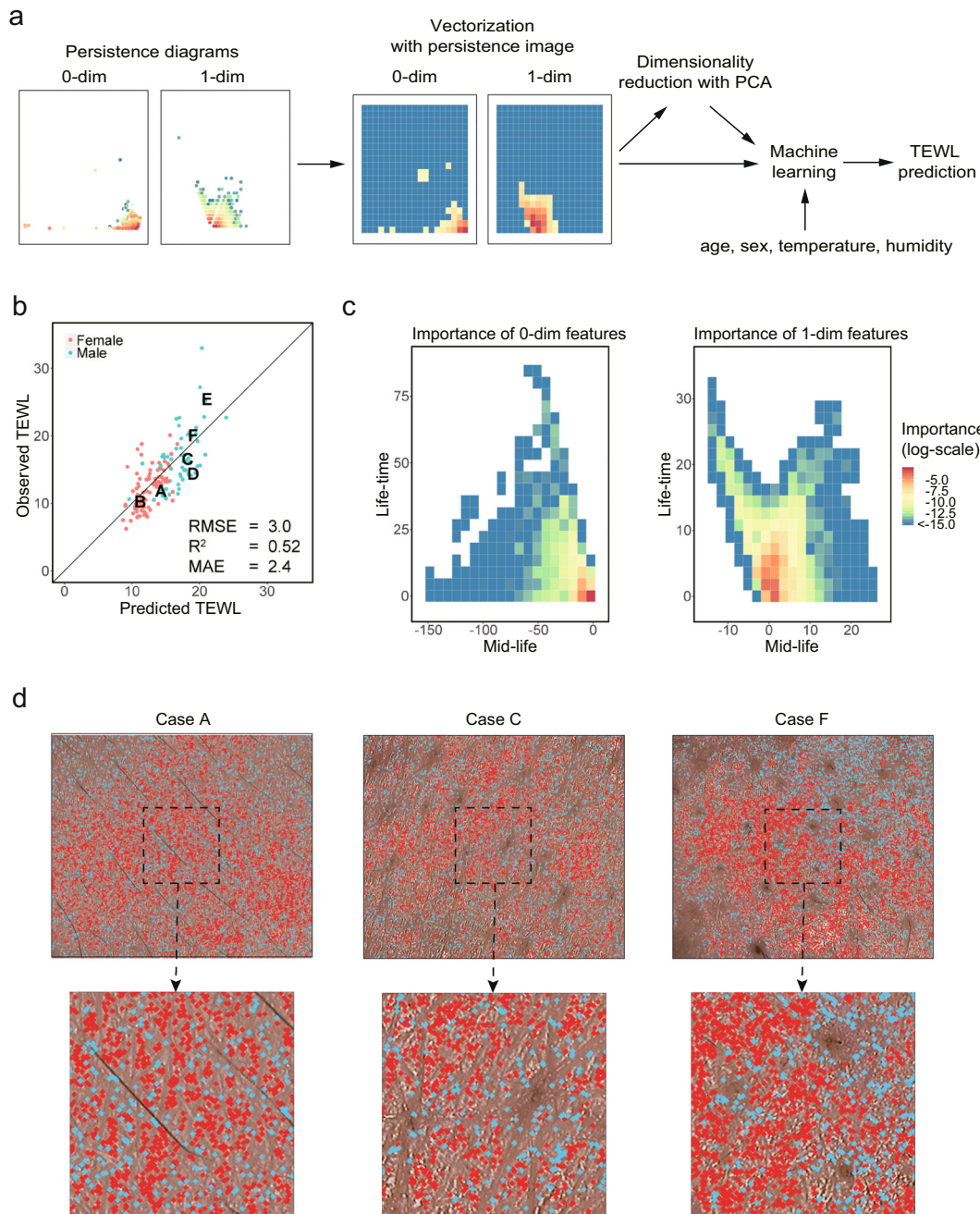


Fig. 3 Prediction of TEWL. **a** The protocol for TEWL prediction using machine learning. Each persistence diagram was vectorized by the persistence image. Then, when using principal component analysis (PCA), each vector was reduced to a low-dimensional vector with PCA and combined with age, sex, temperature, and humidity to create a feature vector. Using the feature vectors, several machine-learning analyses were performed to predict TEWL. **b** Predicted vs. observed plots of TEWL produced by random forest regression. The values of cases A–F in Fig. 2a are also labeled. As indicators of accuracy, the coefficient of determination (R^2), the root mean squared error (RMSE), and the mean absolute error (MAE) are shown. **c** The variable importance of each region in the persistence diagrams is calculated by the random forest regression model according to how much the variable contributed to the regression of TEWL. **d** For cases A, C, and F in Fig. 2a, the birth position of connected components (red) and the death position of holes (blue) in the most important regions of the persistence diagrams are drawn (upper row) and enlarged (lower row).

extrapolated to dermatological disease research for precision medicine. More samples may be needed to reflect the diversity of the patients; in addition to phenotypes, genetic factors such as *FLG* mutations should also be analyzed to improve understanding of the disease pathology. Deep clinical phenotyping based on TDA can provide a basis for precision medicine in dermatology by quantifying clinical characteristics and associating them with molecular biological knowledge.

METHODS

Subjects and measurement

We recruited 244 healthy subjects between the ages of 0 and 64 years. Among them, 143 were women, and 101 were men; 132 were from Akita, and 112 were from Tokyo, Japan (Supplementary Table 6). We only chose subjects who did not use external topical medicines and did not have any skin diseases or other factors such as warts and stains in the measured region. For each subject, measurements were performed twice, in June and

December 2018. We obtained informed consent for analysis and publication of measured data and skin photographs from all participants and ethical approval from the Kenshokai ethical committee for data acquisition [IRB no. 20180810-2 and H30-044] and from Keio University School of Medicine Ethics Committee for data analysis [IRB no. 20160191-6] in accordance with the Declaration of Helsinki. The procedure of the measurement performed in June and December involved certain steps. Before measurement, subjects washed their face using makeup remover and facial wash. For infants who could not wash their faces themselves, we wiped their faces with damp cotton wool twice. Subjects waited for more than 15 min in the test room with temperature and humidity kept to 20.0 °C and 50%, respectively. We took three pictures of the left cheek (the intersection point of the horizontal line through the inferior margin of the nose and the vertical line through the left edge of the left eye) using a microscope and measured TEWL and the moisture content of the stratum corneum of the same region. We used a digital microscope (Model KH-8700; Hirox, Tokyo, Japan) and a vapometer (Model SWL5001JT; Delfin, Technologies Ltd, Kuopio, Finland) to measure TEWL and the Corneometer (Model CM825; Courage & Khazaka Electronic, Cologne, Germany) and Skicon (Model 200EX-USB; YAYOI, Tokyo, Japan) to measure the moisture content of the stratum corneum. We measured TEWL three times and the moisture content of the stratum corneum five times per subject and used their medians in the subsequent analysis.

Image processing

We processed skin images using the Python packages OpenCV⁵⁰ and PyWavelets⁵¹. We trimmed the images into 1400 × 1200 pixels to delete scale bars and transformed them into grayscale using the OpenCV function “cv2.cvtColor” and “cv2.COLOR_BGR2GRAY.” In the wavelet transformation, the grayscale image was decomposed into levels from 0 (coarsest) to 10 (finest). Wavelet coefficients at coarse resolutions represent large structures of the image, including disproportionate light intensity. The image was then reconstructed using only some of these levels (Supplementary Fig. 2).

The images were binarized using Otsu’s method. In the morphological operations, the eroding operation was applied using the OpenCV function “cv2.erode” to expand the black region with the structuring element obtained by the OpenCV function “cv2.getStructuringElement(cv2.MORPH_CROSS,(3,3)).”

Application of TDA using the kNN density estimator

The kNN density estimator was applied as the filtration function using the R package TDA⁵². The specific process was as follows: We set a grid spacing of 10 pixels on an image. On each grid point, the density of white pixels was estimated by measuring the *k*th nearest white pixel with the parameter *k* set to 100. We gradually decreased the threshold from ∞ to −∞ and recorded the thresholds in log-scale at the point where the connected components and holes appeared and disappeared as birth and death. We calculated the mean (mid-life) and the difference (life-time) of these and plotted them to draw the persistence diagrams. We calculated the means and standard deviations of the mid-life and life-time, respectively.

Linear regression analysis

After the skin images had been processed by grayscale transformation, wavelet transformation using levels 4–10, Otsu’s method, eroding operation for five times, and TDA with the kNN density estimator, we performed 14 separate simple linear regression analyses using the R lm function to predict TEWL from 14 explanatory variables, such as sex, age, temperature, humidity, the number of all connected components and holes, and the means and standard deviations of mid-life and life-time of connected components and holes. To investigate the influences of environmental factors on TEWL, we included the temperature and humidity as explanatory variables; these were the averages of the daily temperature and humidity over one month. We calculated the *t*-value and the two-sided *p*-value of each regression. The false discovery rate was calculated using the R p.adjust function with the method of Benjamini and Hochberg⁵³.

Comparison of machine learning models for predicting TEWL

First, we performed machine learning using the R package caret to investigate which algorithm performs best⁵⁴. We removed the meaningless

dimensions with very low variances from the count data of the partitioned persistence diagrams using the caret nearZeroVar function. We performed PCA using the R prcomp function. The eight most important components, namely, PC1–PC8, with contribution ratios larger than 0.01 were used. Then, we created two feature vectors for each sample. One was made of the count data together with age, sex, temperature, and humidity, and the other was made of PC1–PC8 together with age, sex, temperature, and humidity. We used each feature vector, respectively, to predict TEWL and compared their accuracies. We split all data into 70% training data and 30% test data using the R sample function to evaluate the accuracy of prediction. We constructed several models for predicting TEWL using the caret methods “rf” (random forest), “svmRadial” (support vector machine with Gaussian kernel), “enet” (elastic net), “xgbLinear” (gradient boosting using linear functions), “xgbTree” (gradient boosting using tree models), “nnet” (neural network), and “lm” (linear model). Parameter tuning of each model was performed using 10 separate 10-fold cross-validations. The best parameters were chosen according to the root mean square error (RMSE). Each trained model predicted TEWL from each image. Because we took three pictures of each subject in each measurement, we chose the median of the three predicted values as the genuine predicted TEWL of the subject. We evaluated the prediction models by calculating RMSE, the coefficient of determination (R^2), and the mean absolute error (MAE). Since the random forest performed best, we used it as the prediction algorithm of TEWL in the following procedures. To apply the random forest afterwards, we used the RandomForestRegressor function of the Python package scikit-learn to predict TEWL because it is easier to speed up by parallelization than caret⁵⁵.

Comparison of vectorization methods, filtration functions, and preprocessing methods

Next, two vectorization methods of persistence diagrams were considered: counting points in each region and persistence image. The dynamic range of persistence diagrams was partitioned into 20 × 20 regions. In the persistence image, for each point in the persistence diagram, we associated a Gaussian distribution centered at the point with the standard deviation set to 0.1 or 1. Then, the distribution was multiplied by linear weighting function which is 0 at the *x*-axis (i.e., where life-time equals 0) and 1 at the maximum life-time of all persistence diagrams. Finally, all the weighted distributions for all points in the persistence diagram were added and integrated over each region to obtain a 400-dimensional vector. We implemented the persistence image by modifying the Python package persim of scikit-tda⁵⁶.

Furthermore, the choice of algorithms of TDA was considered. In addition to the kNN density estimator, the signed distance and 8-bit grayscale were considered as the filtration functions. The signed distance and 8-bit grayscale were applied using the Python package HomCloud (http://www.wpi-airm.tohoku.ac.jp/hiraoka_lab/homcloud-english.html).

Finally, the choice of preprocessing methods was considered. We assessed the performance of prediction with or without the wavelet transformation and morphological operations. Additionally, to investigate the effect of parameters in the wavelet reconstruction, seven combinations of levels were examined.

Calculation of the variable importance

The variable importance was calculated using the random forest regression model, which had the vectorized data of persistence diagrams, age, sex, temperature, and humidity as explanatory variables. We obtained an average of the variable importance calculated 10 times with different test data selected randomly. When using PCA, the importance of the regions of persistence diagrams was calculated from the importance of principal components. The information of birth positions and death positions was obtained using the HomCloud function “homcloud.interface.draw_birthdeath_pixels_2d.”

DATA AVAILABILITY

Except for skin images, all data used in this study are included in the GitHub repository (https://github.com/kosekei/skin_TDA). The skin images used in this study are available from the authors upon request.

CODE AVAILABILITY

All scripts used in this study are included in the GitHub repository (https://github.com/kosekei/skin_TDA).

Received: 23 December 2019; Accepted: 15 October 2020;

Published online: 18 December 2020

REFERENCES

- Segre, J. A. Epidermal barrier formation and recovery in skin disorders. *J. Clin. Invest.* **116**, 1150–1158 (2006).
- Palmer, C. N. et al. Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. *Nat. Genet.* **38**, 441–446 (2006).
- McAleer, M. A. & Irvine, A. D. The multifunctional role of filaggrin in allergic skin disease. *J. Allergy Clin. Immunol.* **131**, 280–291 (2013).
- Goleva, E., Berdyshev, E. & Leung, D. Y. Epithelial barrier repair and prevention of allergy. *J. Clin. Invest.* **129**, 1463–1474 (2019).
- Kubo, A., Nagao, K. & Amagai, M. Epidermal barrier dysfunction and cutaneous sensitization in atopic diseases. *J. Clin. Invest.* **122**, 440–447 (2012).
- Akdeniz, M., Gabriel, S., Lichtenfeld-Kottner, A., Blume-Peytavi, U. & Kottner, J. Transepidermal water loss in healthy adults: a systematic review and meta-analysis update. *Br. J. Dermatol.* **179**, 1049–1055 (2018).
- Kelleher, M. et al. Skin barrier dysfunction measured by transepidermal water loss at 2 days and 2 months predates and predicts atopic dermatitis at 1 year. *J. Allergy Clin. Immunol.* **135**, 930–935.e1 (2015).
- Kelleher, M. M. et al. Skin barrier impairment at birth predicts food allergy at 2 years of age. *J. Allergy Clin. Immunol.* **137**, 1111–1116.e8 (2016).
- Greenspan, H., van Ginneken, B. & Summers, R. M. Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. *IEEE Trans. Med. Imaging* **35**, 1153–1159 (2016).
- Zhang, Q.-S. & Zhu, S.-C. Visual interpretability for deep learning: a survey. *Front. Inf. Technol. Electron. Eng.* **19**, 27–39 (2018).
- Carlsson, G. Topology and data. *Bull. Am. Math. Soc.* **46**, 255–308 (2009).
- Hiraoka, Y. et al. Hierarchical structures of amorphous solids characterized by persistent homology. *Proc. Natl Acad. Sci. USA* **113**, 7035–7040 (2016).
- de Silva, V. & Ghrist, R. Coverage in sensor networks via persistent homology. *Algebraic Geom. Topol.* **7**, 339–358 (2007).
- Chan, J. M., Carlsson, G. & Rabadan, R. Topology of viral evolution. *Proc. Natl Acad. Sci. USA* **110**, 18566–18571 (2013).
- Dunaeva, O. et al. The classification of endoscopy images with persistent homology. *Pattern Recognit. Lett.* **83**, 13–22 (2016).
- Kaiser, T. et al. Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features. *Med Image Anal.* **55**, 1–14 (2019).
- Crawford, L., Monod, A., Chen, A. X., Mukherjee, S. & Rabadán, R. Predicting clinical outcomes in glioblastoma: an application of topological and functional data analysis. *J. Am. Stat. Assoc.* 1–12, <https://doi.org/10.1080/01621459.2019.1671198> (2019).
- Nicolau, M., Levine, A. J. & Carlsson, G. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc. Natl Acad. Sci. USA* **108**, 7265–7270 (2011).
- Bendich, P., Marron, J. S., Miller, E., Pieloch, A. & Skwerer, S. Persistent homology analysis of brain artery trees. *Ann. Appl. Stat.* **10**, 198–218 (2016).
- d'Amico, M., Ferri, M. & Stanganelli, I. Qualitative asymmetry measure for melanoma detection. In *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821)*, Vol. 2, 1155–1158 (IEEE, Arlington, VA, USA, 2005).
- Ferri, M. & Stanganelli, I. Size functions for the morphological analysis of melanocytic lesions. *Int. J. Biomed. Imaging* **2010**, 621357 (2010).
- Ferri, M., von Tomba, L., Visotti, A. & Stanganelli, I. A feasibility study for a persistent homology-based k -nearest neighbor search algorithm in melanoma detection. *J. Math. Imaging Vis.* **57**, 324–339 (2017).
- Chung, Y.-M., Hu, C.-S., Lawson, A. & Smyth, C. Topological approaches to skin disease image analysis. In *2018 IEEE International Conference on Big Data (Big Data)*, 100–105 (IEEE, Seattle, WA, USA, 2019).
- Binchi, J., Merelli, E., Rucco, M., Petri, G. & Vaccarino, F. jHoles: a tool for understanding biological complex networks via clique weight rank persistent homology. *Electron. Notes Theor. Comput. Sci.* **306**, 5–18 (2014).
- Arakawa, N., Ohnishi, H. & Masuda, Y. Development of quantitative analysis for the micro-relief of the skin surface using a video microscope and its application to examination of skin surface texture. *J. Soc. Cosmet. Chem. Jpn.* **41**, 173–180 (2007).
- Mack, Y. & Rosenblatt, M. Multivariate k -nearest neighbor density estimates. *J. Multivar. Anal.* **9**, 1–15 (1979).
- Obayashi, I., Hiraoka, Y. & Kimura, M. Persistence diagrams with linear machine learning models. *J. Appl. Comput. Topol.* **1**, 421–449 (2018).
- Robins, V., Saadatfar, M., Delgado-Friedrichs, O. & Sheppard, A. P. Percolating length scales from topological persistence analysis of micro-CT images of porous materials. *Water Resour. Res.* **52**, 315–329 (2016).
- Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**, 62–66 (1979).
- Edelsbrunner, H. & Harer, J. Persistent homology—a survey. In (eds Goodman, J. E., Pach, J. & Pollack, R.) *Surveys on Discrete and Computational Geometry: Twenty Years Later*, Vol. 453, 257–282 (American Mathematical Soc., Snowbird, UT, USA, 2008).
- Edelsbrunner, H., Letscher, D. & Zomorodian, A. Topological persistence and simplification. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, 2000, 454–463 (IEEE Computer Society, Redondo Beach, CA, USA, 2002).
- McDonald, J. H. *Handbook of Biological Statistics*, Vol. 2 (Sparky House Publishing, Baltimore, MD, 2009).
- Heinrich, U. et al. Multicentre comparison of skin hydration in terms of physical, physiological- and product-dependent parameters by the capacitive method (Corneometer CM 825). *Int. J. Cosmet. Sci.* **25**, 45–53 (2003).
- Tagami, H. Electrical measurement of the hydration state of the skin surface in vivo. *Br. J. Dermatol.* **171**, 29–33 (2014).
- Hashimoto-Kumasaka, K., Takahashi, K. & Tagami, H. Electrical measurement of the water content of the stratum corneum in vivo and in vitro under various conditions: comparison between skin surface hygrometer and corneometer in evaluation of the skin surface hydration state. *Acta Derm. Venereol.* **73**, 335–339 (1993).
- Bubenik, P. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.* **16**, 77–102 (2015).
- Reininghaus, J., Huber, S., Bauer, U. & Kwitt, R. A stable multi-scale kernel for topological machine learning. In *Proc. IEEE Conference on Computer Vision Pattern Recognition*, 4741–4748 (IEEE, San Juan, Puerto Rico, USA, 2015).
- Kwitt, R., Huber, S., Niethammer, M., Lin, W. & Bauer, U. Statistical topological data analysis—a kernel perspective. In (eds Cortes, C., Lee, D. D., Sugiyama, M., Garnett, R.) *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, MIT Press, 55 Hayward St., 3070–3078 (Cambridge, MA, USA, 2015).
- Monod, A., Kališnik, S., Patiño-Galindo, J. A. & Crawford, L. Tropical sufficient statistics for persistent homology. *SIAM J. Appl. Algebra Geom.* **3**, 337–371 (2019).
- Adams, H. et al. Persistence images: a stable vector representation of persistent homology. *J. Mach. Learn. Res.* **18**, 1–35 (2017).
- Carrière, M. et al. PersLay: a neural network layer for persistence diagrams and new graph topological signatures. *Proc. Mach. Learn. Res.* **108**, 2786–2796 (2020).
- Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
- Haenssle, H. et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* **29**, 1836–1842 (2018).
- Brinker, T. et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *Eur. J. Cancer* **111**, 148–154 (2019).
- Wu, P. et al. Optimal topological cycles and their application in cardiac trabeculae restoration. In *International Conference on Information Processing in Medical Imaging. IPMI 2017. Lecture Notes in Computer Science*, Vol. 10265 (eds Niethammer, M. et al.) 80–92 (Springer, 2017).
- Fluhr, J. W. et al. Comparative study of five instruments measuring stratum corneum hydration (Corneometer CM 820 and CM 825, Skicon 200, Nova DPM 9003, DermaLab). Part I. In vitro. *Skin Res. Technol.* **5**, 161–170 (1999).
- Caspers, P., Lucassen, G., Bruining, H. & Puppels, G. Automated depth-scanning confocal Raman microspectrometer for rapid in vivo determination of water concentration profiles in human skin. *J. Raman Spectrosc.* **31**, 813–818 (2000).
- Egawa, M., Hirao, T. & Takahashi, M. In vivo estimation of stratum corneum thickness from water concentration profiles obtained with Raman spectroscopy. *Acta Derm. Venereol.* **87**, 4–8 (2007).
- Griffiths, C. E. M., Barker, J., Bleiker, T., Chalmers, R. & Creamer, D. *Rook's Textbook of Dermatology* 9th edn (Wiley-Blackwell, West Sussex, 2016).
- Bradski, G. The OpenCV Library. *Dr. Dobbs's J. Softw. Tools* **25**, 122–125 (2000).
- Lee, G. R., Gommers, R., Wasilewski, F., Wohlfahrt, K. & O'Leary, A. PyWavelets: a Python package for wavelet analysis. *J. Open Source Softw.* **4**, 1237 (2019).
- Fasy, B. T., Kim, J., Lecci, F. & Maria, C. Introduction to the R package TDA. Preprint at <https://arxiv.org/abs/1411.1830v2> (2015).
- Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).

54. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **28**, 5 (2008).
55. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
56. Saul, N. *Scikit-TDA: Topological Data Analysis for Python* (Zenodo, 2019).

ACKNOWLEDGEMENTS

This work was supported by RIKEN Hub for predictive and preventive precision medicine driven by big data in JST Support program for starting up innovation hub (ihub), SECOM Science and Technology Foundation (to E.K.) and AMED under grant numbers JP17gm5010003, JP19gk0110043 (to E.K.), JP18ek0410046, and JP19ek0410058 (to H.K., E.K., T.E., and M.A.).

AUTHOR CONTRIBUTIONS

K.K., H.K., and E.K. designed the study and prepared the manuscript. K.K. coded the computational model and analyzed the data. T.A., M.N., M.M., and E.N. designed the measurement and collected the data. T.E. discussed the results and commented on the manuscript. M.A. and E.K. supervised the study.

COMPETING INTERESTS

This work was partially supported by KOSÉ Corporation. T.A., M.N., M.M., and E.N. are employed by KOSÉ Corporation.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at <https://doi.org/10.1038/s41540-020-00160-8>.

Correspondence and requests for materials should be addressed to E.K.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020