Collagen and Leather

RESEARCH Open Access

Topological data analysis with digital microscope leather images for animal species classification

Takuya Ehiro[1*] and Takeshi Onji[1]

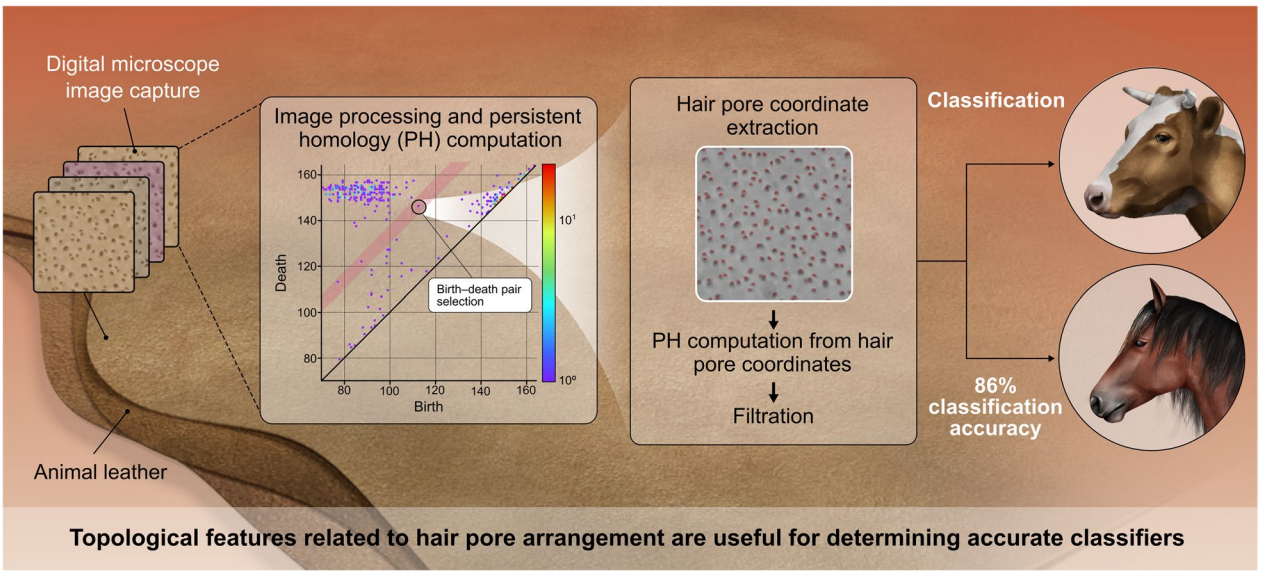
Abstract

This study presents a method for classifying cow and horse leather using a small number of digital microscope images and topological data analysis. In this method, hair pore coordinates in the images are used as essential information for classification. First, the coordinates were semiautomatically extracted using conventional image processing methods and persistent homology (PH) computation. Binary images with white pixels corresponding to the coordinates were generated, and their PHs were computed using filtration based on the Manhattan distance. In addition to the pairwise distance between the two pores, zeroth- and first-order lifetimes were used as explanatory variables to construct the classifier. Among the three explanatory variables, the zeroth-order lifetime resulted in the highest classification accuracy (86%) for the test data. Furthermore, we constructed logistic regression (LR) and random forest (RF) models using the zeroth-order lifetime computed from all images and conducted model interpretation. In both LR and RF, information on a zeroth-order lifetime of less than 10 was used as an important explanatory variable. Additionally, the inverse analysis of birth–death pairs suggested that the zeroth-order lifetime contains topological information distinct from the conventional pairwise distance. Our proposed method is designed to be robust in data-limited situations because it only uses hair pore coordinates as explanatory variables and does not require other information, such as hair pore density or pore size. This study demonstrates that accurate classifiers can be obtained using topological features related to hair pore arrangement.

Keywords Animal species classification, Machine learning, Topological data analysis, Persistent homology, Lifetime

*Correspondence:
Takuya Ehiro
ehirot@orist.jp

Graphical Abstract



1 Introduction

Leather, a versatile material derived from animal hides, plays a crucial role in various products, including footwear, accessories and automotive interiors. The intrinsic properties of leather, such as hardness, flexibility and durability, are largely determined by the animal species from which it originates. Consequently, accurate identification of the animal species used in leather products is of paramount importance for several reasons: it ensures transparency in raw material sourcing, enhances consumer trust, facilitates quality control, supports market research and competition analysis, and protects endangered species [1].

Traditionally, various analytical approaches have been employed for species identification in leather, including morphological studies, protein analysis and chemical analyses. However, these methods face limitations such as reduced effectiveness on processed leather, time-consuming procedures, potential inaccuracies due to sample degradation, and difficulties in achieving species-level resolution [2]. In traditional image analysis methods, leather experts have identified animal species by observing the arrangement of hair pore patterns, which vary among different animal species. While this identification method is practical, it relies heavily on expert knowledge and can be subjective, potentially introducing bias into the identification results.

In recent years, advanced image processing techniques have offered an effective approach to identifying leather species. Digital microscope and scanning electron microscope images of leather samples have been utilized to extract morphological, geometrical, textural and statistical features related to pore patterns. These extracted features serve as input for machine learning classification models, facilitating the automated identification of leather species [3–6]. Additionally, a neural network-based approach can be effective in automatically capturing more complex and nuanced features that are not easily defined by traditional image processing techniques [7]. Rapidly evolving deep learning models, typically convolutional neural networks (CNNs) [8], have been widely utilized in various fields. Owing to their combination of convolutional and pooling layers, CNNs can maintain their robustness in positioning and capturing important features for classification. Although an increase in the number of layers in a CNN can enhance its expressive power and classification accuracy [9], complex deep learning models can suffer from the risk of overfitting and require sufficient data for optimization.

To achieve robust classification, CNNs need both architectural complexity and large datasets to prevent overfitting. Although data augmentation and transfer learning techniques are effective approaches for overcoming the problem of limited data [10, 11], typical model-free data augmentation techniques, such as flipping, rotation and noise injection, do not contribute to essential data augmentation because the original data are the same. Moreover, model-based data augmentation

and transfer learning techniques in small-data situations require a model pretrained with sufficient data. Additionally, transfer learning generally requires careful selection of the source domain, which potentially correlates with the target domain, to prevent negative transfer. Therefore, these techniques require additional considerations and preparation steps compared with simply constructing a classification model.

These challenges have motivated the exploration of alternative methods that do not require large amounts of data or complex neural network models. In this study, we aimed to construct a leather image classification model using topological features extracted from digital microscope images to bypass the need for neural networks and large datasets. We employed topological data analysis (TDA), which applies algebraic topology to extract meaningful information from complex and high-dimensional data by studying the shape and structure rather than individual data points. In TDA, data are represented as a simplicial complex composed of simplices, enabling the computation of persistent homology (PH) [12, 13]. PH characterizes sets of holes with specific dimensions, providing information about topological features such as connected components, loops and voids in the data. It extends traditional homology by analyzing how these features persist across different scales, generating birth–death pairs that indicate when features emerge and disappear. These pairs are typically visualized in persistence diagrams (PD) as two-dimensional (2D) histograms. In our approach, we used the lifetime (persistence), which represents the birth-to-death time of a hole, as an explanatory variable for machine learning.

By utilizing birth–death pairs, it is possible to quantify the shapes of data that are conventionally difficult to extract and represent numerically. TDA is a unique method that can characterize the shape of data and is synergetic with machine learning [14]. TDA has been applied to various data analysis tasks, such as many-body atomic structures in glass [15], protein folding [16], nanoporous metal–organic frameworks [17], higher-order structures of polymers [18], molecular fingerprints [19], relapse risk prediction in patients with acute lymphoblastic leukemia [20], and brain network analysis [21].

In this study, we utilized the lifetime derived from PH as an explanatory variable reflecting the topological features extracted from the images. By using the zeroth-order lifetime, the classifier showed higher accuracy than when using pairwise distance, demonstrating the effectiveness of topological features in classifying leather images.

In some countries, such as Japan, legislation mandates the disclosure of animal species for leather products (e.g., Household Goods Quality Labeling Act of Japan). Mature cow and horse leather share similar characteristics, with subtle differences in pore size and distribution. These characteristics can vary significantly based on factors such as the animal's age, individual differences and the specific body part [22]. These similarities and variations make species classification challenging, particularly with limited datasets. Therefore, this study aims to classify these two types of leather using TDA techniques.

2 Experimental section

2.1 Capturing and processing leather images

The classifier is constructed using scrap from commercially available tanned leather samples collected from cows and horses. These leather samples were procured from diverse leather suppliers, with a collection of over 20 small fragments. Fifty color leather images were captured using a digital microscope (KH-7700, Hirox Co., Ltd.) at a magnification of 40×for cow and horse leather, respectively. The obtained color images were converted to grayscale, cropped, and resized to half of their original size of 300×300 pixels. Subsequently, contrast-limited adaptive histogram equalization (CLAHE) [23] and 2D Fused Lasso [24, 25] were applied. The clipping limit and tile grid size for CLAHE were set to 0.2 and 4×4, respectively. OpenCV computer vision and a machine learning software library [26] were used for the image processing. A 2D Fused Lasso is a regularization technique that penalizes differences in brightness between adjacent pixels. The loss function $L(\mathbf{X}, \mathbf{W}; \alpha, \lambda)$ was minimized using the Nesterov accelerated gradient method [27].

$$L(\mathbf{X}, \mathbf{W}; \alpha, \lambda) = \frac{\alpha}{2} \cdot MSE(\mathbf{X}, \mathbf{W}) + \lambda \cdot TV(\mathbf{W}), \quad (1)$$

where $\mathbf{X}$ is input image data, $\mathbf{W}$ is the weight matrix corresponding to $\mathbf{X}$, λ is the regularization parameter for total variation (TV) L1 penalty, and α is a parameter to control the penalty to the mean squared error (MSE) between $\mathbf{X}$ and $\mathbf{W}$. In the TV term, the L1 penalty was applied to the differences between adjacent pixels. The weight α for the gradient derived from the MSE term was set to 0.01, and the regularization parameter λ controlling the subgradient from the TV term was set to 0.05, with 5000 iterations.

To determine hair pore coordinates in the grayscale images, an adaptive threshold for pore brightness was set for each image. Using EMPeaks [28, 29], the brightness of the grayscale images was fitted with a two-component Gaussian mixture distribution, and pixels with brightness above the 90th percentile of Gaussian distribution with the smaller mean were excluded from the extraction of pore coordinates. Subsequently, using the HomCloud library, level-set filtrations were constructed based on the brightness values of the grayscale images [30], and a

zeroth-order PD (PD0) was created. A birth–death pair far from the diagonal of the PD indicates a distinct difference in brightness between the background and the pores. However, birth–death pairs with small brightness differences due to noise are concentrated near the diagonal. Therefore, an appropriate birth–death pair serving as the boundary between them was selected, and its lifetime was used as the standard threshold. By investigating the coordinates in the original image that correspond to the birth–death pairs with lifetimes greater than or equal to the aforementioned lifetime and that do not exceed the brightness threshold, the pore coordinates were semi-automatically extracted. Although the above process often works effectively, the standard lifetime threshold is adjusted when the pore coordinate extraction does not work adequately. In addition, each image was resized to ensure that approximately 150 pores were included in the 300×300 images. To maintain the robustness of the classifier, slight discrepancies in the number of extracted pores were tolerated and the resulting binary images were utilized as training and test data.

2.2 Constructing classifiers

For the binary images, the zeroth- and first-order PHs were obtained by constructing white-pixel-based filtrations, using the Manhattan distance [31] to calculate the zeroth- and first-order lifetimes from the corresponding PHs. In addition, the pairwise distances between the two pores were calculated from the hair pore coordinates in the binary images. These were then used as explanatory variables by performing kernel density estimation (KDE) [32] with a Gaussian kernel. The bandwidth of the kernel function and the interval for discretizing the estimated probability density to prepare the explanatory variables were optimized along with the hyperparameters of the classification models using Optuna [33]. During optimization, the accuracy of the training data in fivefold cross-validation (CV) was maximized using the tree-structured Parzen estimator (TPE) [34]. The 100 image data were randomly split into training/test ratios of 0.20/0.80, 0.30/0.70, 0.50/0.50, 0.70/0.30 and 0.80/0.20, using stratified sampling to maintain consistent class proportions across all splits. Classification performance was evaluated using the following metrics: accuracy, precision, recall, F1 score and area under the receiver operating characteristic curve (AUC-ROC). The hyperparameter candidates for KDE and each model are listed in Table S1. Logistic regression (LR) [35], support vector machine (SVM) [36], random forest (RF) [37] and extreme gradient boosting (XGBoost) [38] were employed for the classification task. Linear and radial basis function (RBF) kernels were used for the SVM, and these SVM models were denoted as SVM (Linear) and SVM (RBF), respectively. Scikit-learn [39] and XGBoost [38] libraries were used to construct classification models.

For the comparison, CNN was implemented for binary classification using PyTorch [40], an open-source machine learning library. The network architecture comprises two convolutional layers, followed by two fully connected layers. Each convolutional layer is followed by ReLU activation and max pooling. Dropout is applied between the fully connected layers to mitigate overfitting. An on-the-fly data augmentation strategy was employed to enhance the model's generalization capability. Each training image was augmented by applying rotations of 0, 90, 180 and 270 degrees, quadrupling the size of the training dataset. The Adam optimizer and cross-entropy loss were used for training. The dataset was split into training (64%), validation (16%) and test (20%) sets using stratified sampling to maintain class balance across all sets. The hyperparameters listed in Table S1 were tuned using Optuna. The best-performing hyperparameters were then used to train the final model for 50 epochs.

The LR and RF were constructed using all images by tuning their hyperparameters in a fivefold CV. For the LR, standard regression coefficients were investigated for model interpretation. The constructed RF model was explained using SHapley Additive explanations (SHAP) [41], which is a widely used technique for model interpretation. In this study, TreeExplainer [42] was adopted to explain the RF output.

Variables with high correlation coefficients between explanatory variables and those with a significant number of repeated values were excluded. Prior to the analysis, each variable was autoscaled.

3 Results and discussion

3.1 Image processing for extracting hair pore coordinates

Figure 1 shows digital microscope images of cow and horse leather. As shown in Fig. 1, the color, surface roughness and pore size of the leather samples varied. Therefore, in data-limited situations, it is anticipated that classification between cows and horses is challenging because the variety of images can be noisy, which deteriorates model construction. Hence, in this study, we considered the pore arrangement pattern in the images, that is, the hair pore coordinates, as the most essential and crucial information and attempted to extract them for model construction.

Figure 2 illustrates the steps from image acquisition to model construction. First, the images were converted to grayscale and cropped to a specific size. Next, for the ease of image processing, we resized the images to half their original size and flattened the brightness distribution using CLAHE. Furthermore, we used a 2D Fused Lasso for image smoothing. A 2D Fused Lasso performs

Fig. 1 Digital microscope images of cowhides and horsehides

regularization to minimize the difference in brightness between adjacent pixels in both height and width directions. This not only removed noise but also prevented the extraction of multiple coordinates from a single hair pore area, ensuring that a representative position for each pore was extracted. In this study, mild CLAHE processing conditions were adopted as the images were to be binarized. The CLAHE parameters were preliminary determined through visual inspection of the resulting images. The selected parameters minimized the differences in brightness distribution among 144 subregions within each image. The distances between brightness distributions of the 144 subregions were evaluated using the mean earth mover's distance (Fig. S1).

Figure S2 shows the original grayscale image and images obtained after applying the 2D Fused Lasso with different L1 regularization penalties. Although the regularization parameter used in this study was not large, Fig. S2 shows that the smoothing effect on the image became more pronounced as the regularization parameter increased. Although a single regularization parameter was used in this study, it is also possible to apply different regularization parameters to each image. In general, a large regularization parameter strongly removes noise and reduces the difference in brightness within local hair pore areas. The 2D Fused Lasso loss function exhibits convexity and requires tolerable computational time due to the moderate dataset size. Based on these properties, the number of iterations was conservatively set to 5000, ensuring sufficient convergence. Figure S3 demonstrates that comparable smoothing results can be achieved with fewer iterations in the present dataset.

Sublevel filtrations based on brightness were constructed from the grayscale images processed using 2D Fused Lasso. Filtration enabled PH computation and generated the PD0s (Fig. 3). In this filtration, we tracked the changes in the topological features as a threshold for binarizing grayscale images, resulting in a chain of increasing complexes. Inclusion relationships of complexes in the filtration induced linear maps connecting homology groups at each threshold. Consequently, algebraic computations of PH are feasible and provide topological information on the numerical changes in homology groups in a specific dimension. Topological information can be represented as birth–death pairs in PD. Birth–death pairs in PD indicate

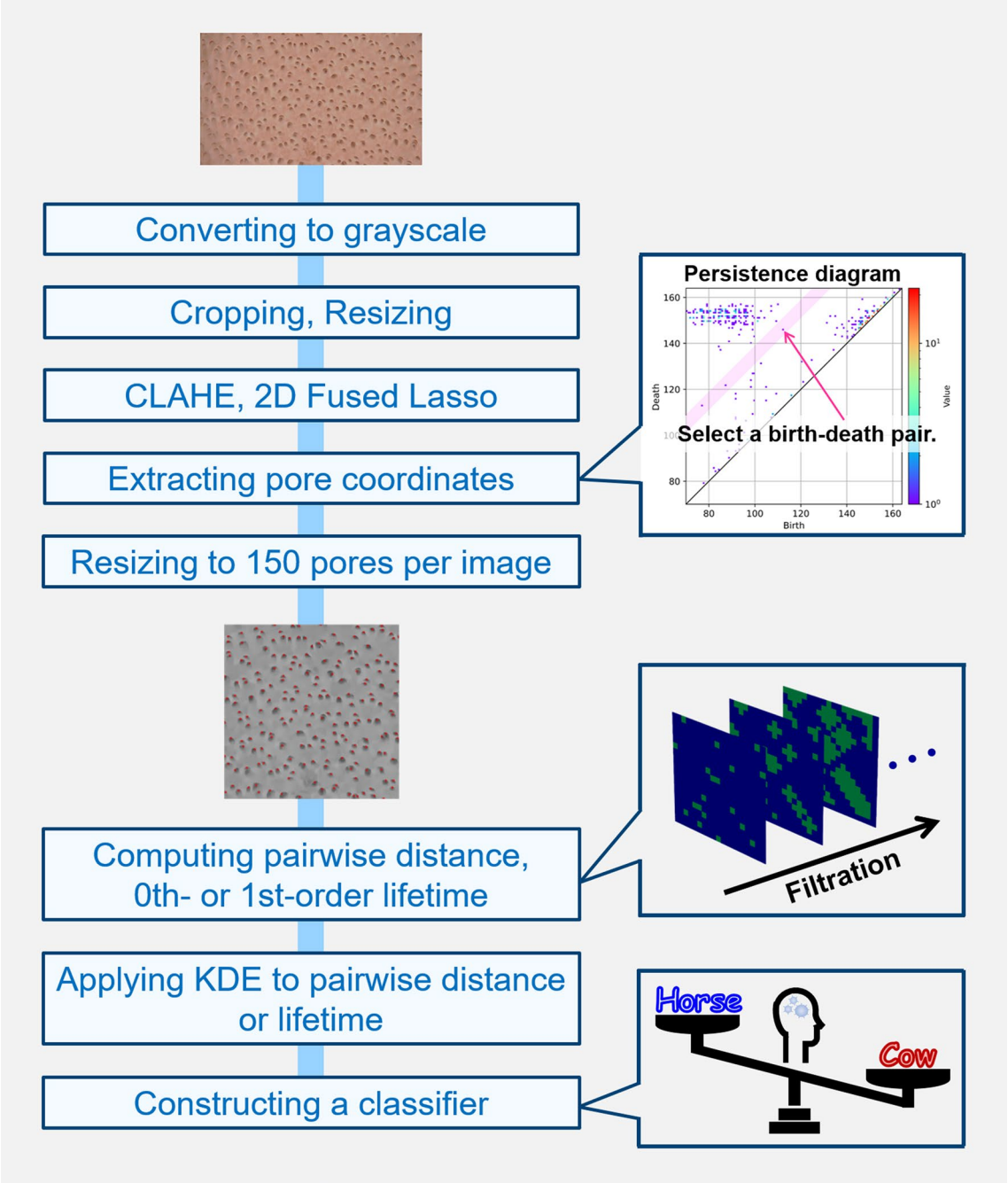


Fig. 2 The procedure to construct a classifier

the birth and death times of holes in a specific dimension. Birth–death pairs far from the diagonal of the PD corresponded to long-lived holes. In this analysis, the long-lived holes likely correspond to the dark hair pore areas. Furthermore, birth–death pairs near the diagonal correspond to short-lived holes, which often correspond to noise. Therefore, as shown in Fig. 3, a birth–death pair that appeared to be positioned near the boundary between short- and long-lived holes was selected, and its lifetime was used as the standard threshold for pore coordinate extraction. While some images exhibited a more gradual positioning of birth–death pairs than those in Fig. 3, the same process was applied to all images. The pore coordinates were

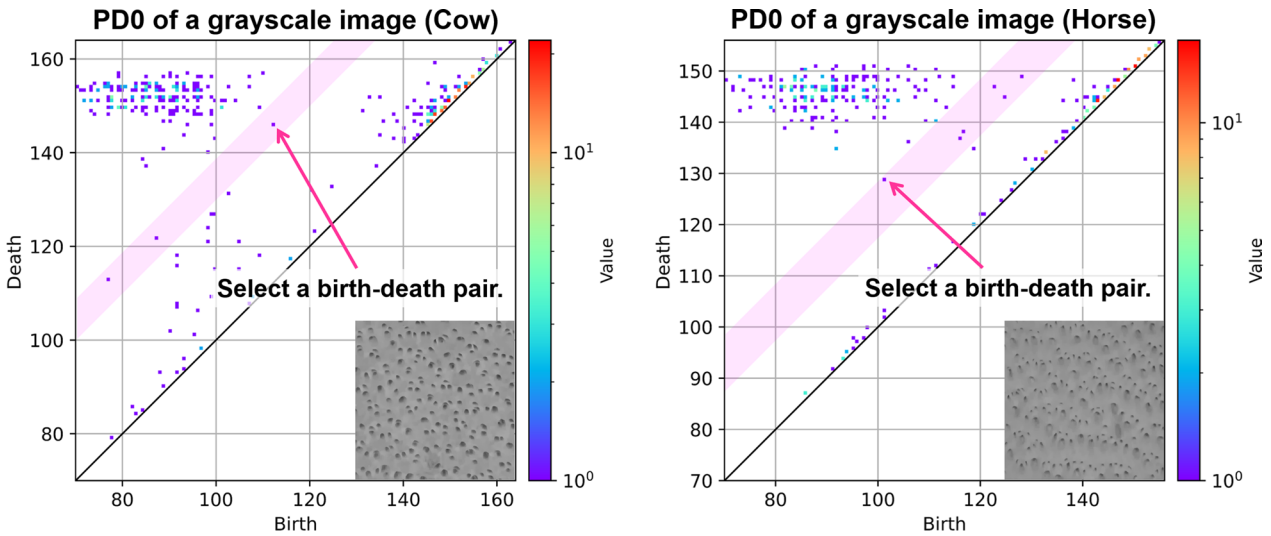


Fig. 3 PD0s of grayscale images

extracted by adaptively adjusting the standard threshold as required.

Because hair pore areas are generally darker than other areas, relatively bright areas should be eliminated as hair pore extraction candidates. Therefore, the brightness of the grayscale images was deconvoluted into two Gaussian distributions (Fig. S4). Deconvolution was performed using EMPeaks, in which an EM-algorithm-based approach was employed to process large amounts of data quickly and achieve stable and automatic calculations. As stated above, the purpose of deconvolution in this study was to prevent the extraction of areas that were too bright as pores. Thus, emphasis was placed on the deconvolution of bright and dark areas using a two-component Gaussian mixture distribution rather than focusing solely on the fitting accuracy. Pixels with brightness greater than the 90th percentile of the Gaussian distribution with a smaller mean (component 2) were excluded from the extraction targets. Hair pore coordinates were extracted using the above process (Fig. 4). Following this process, binary images were obtained by setting the pore coordinates to white, and all other areas to black. This process allows for semiautomation of the laborious task of manually extracting numerous pore positions.

Although the proposed method involves slight human intervention, it does not require specialized leather knowledge. In this method, there exists some arbitrariness in selecting points in a PD. Multiple point selections within the shaded area of Fig. 3 yielded consistent identification of the same pixels as pore locations (Fig. S5). This robustness suggests that the method maintains reliability despite the presence of a manual step. The visual inspection step for the binarized image serves as a final verification procedure to ensure the quality of the results.

3.2 Constructing a classification model using binary images

Hair pore density in an image can differ depending on factors, such as species and age. As shown in Fig. 1, the number of white pixels corresponding to the pores in the binary images obtained from the above process varies depending on the image. In situations involving diverse and abundant data, images with varying numbers of pores are important features. However, in data-limited situations, hair pore density can be noisy information. Therefore, we used only the pore arrangement patterns and excluded information on hair pore density and pore size. To eliminate the influence of hair pore density, the images were resized such that each image contained approximately 150 pores. To ensure applicability to diverse leather samples, including those with wider pore spacing, we fixed the magnification at 40x, which results in at least 150 pores per image for most samples.

The zeroth- and first-order PHs of the resized binary images were calculated to generate zeroth- and first-order lifetimes. As mentioned above, lifetimes represent the period from the birth to death of holes in a specific dimension. In addition to the lifetimes, the pairwise distance between the two hair pores was computed as an explanatory variable. Instead of using lifetimes and pairwise distances directly, KDE was performed to estimate the density distributions of these features. The estimated probability densities were discretized with specific grid

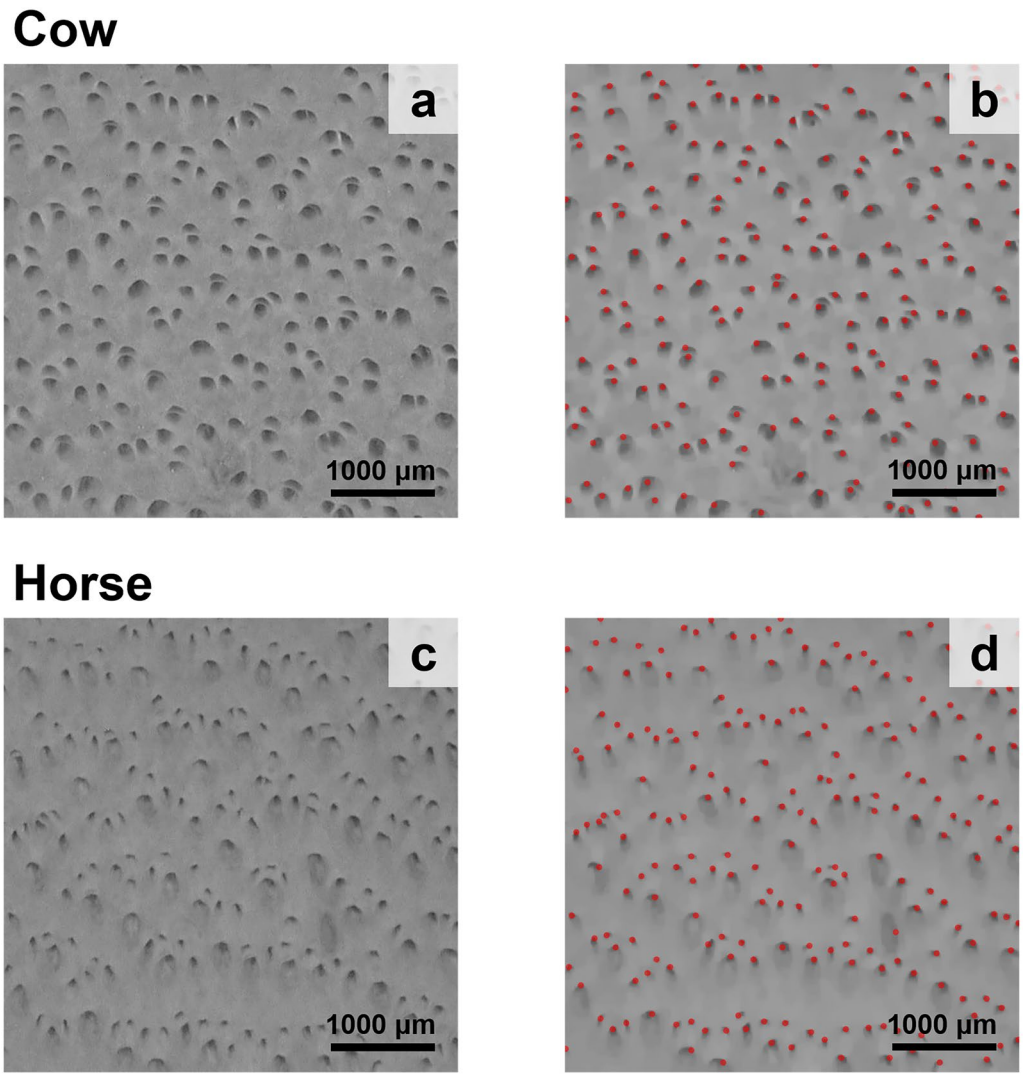


Fig. 4 The original grayscale images (**a**, **c**) and the results of pore coordinate extraction (**b**, **d**)

sizes, and the probability densities at each point were used as explanatory variables.

The results of the KDE with specific hyperparameters (Table S2) for the zeroth- and first-order lifetimes and pairwise distances are shown in Fig. 5. Each thin line represents the result of KDE for each image, whereas the thick lines represent the average probability density for each class. Figure 5 shows that the probability density of the zeroth-order lifetime most clearly highlights the differences between the two classes. Centering reduces the influence of dynamic ranges and distinguishes the average probability densities across all features.

Principal component analysis (PCA) was applied to the KDE results to reduce the dimensionality of each explanatory variable. In PCA, a new variable (PC1) was created as a linear combination of the original variables in the direction of the largest variance. A new variable (PC2) was subsequently created with the constraint of being orthogonal to PC1. Figure 6 presents the PCA score plots for the explanatory variables. The PCA visualizes the distribution of the two classes through linear mapping. The PCA results using the zeroth-order lifetime suggest that the distributions of the two classes are separated. Although a similar separation trend was observed for the pairwise distance, the boundary between the two classes was less distinct than that of the zeroth-order lifetime. However, the PCA result using the first-order lifetime indicates relatively similar distributions for both classes. Because PCA is a simple linear dimensionality reduction technique, it was suggested that complex information, that is, the shape of the data, in the binary images was extracted, especially from the zeroth-order lifetime.

In Figs. 5 and 6, the explanatory variables were generated with specific hyperparameters (Table S2); however,

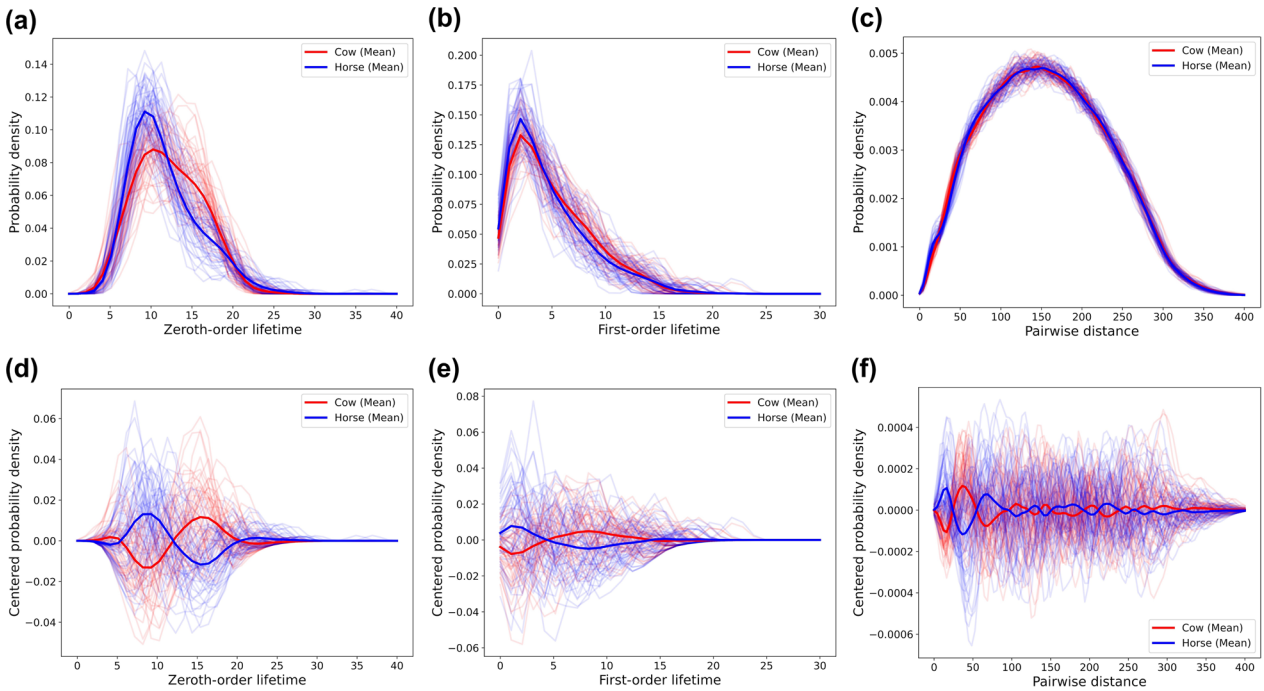


Fig. 5 Density distributions of zeroth-order lifetime (**a**, **d**), first-order lifetime (**b**, **e**) and pairwise distances obtained via KDE (**c**, **f**) (top row), along with their centered density distributions (bottom row)

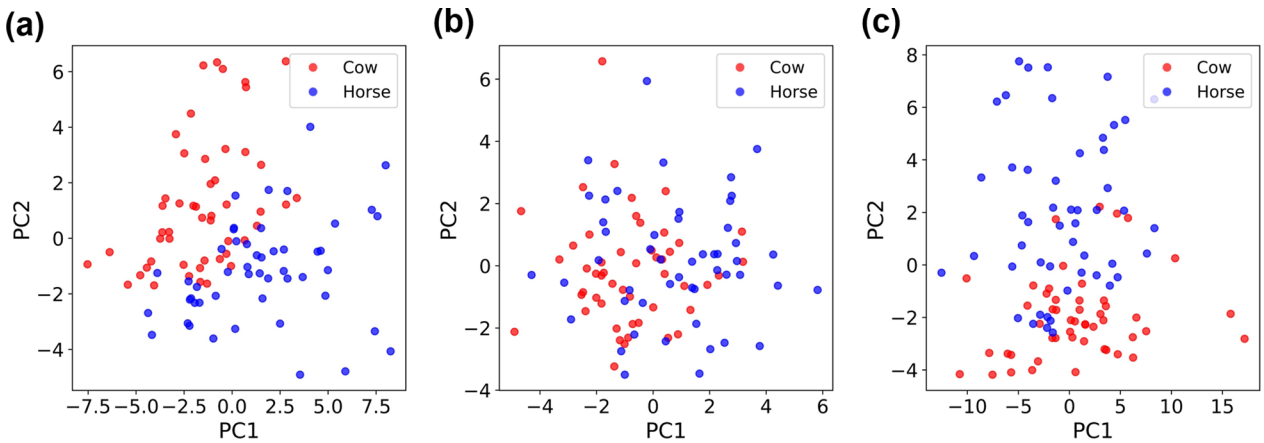


Fig. 6 The PCA score plots of density distributions of zeroth-order lifetime (**a**), first-order lifetime (**b**) and pairwise distances (**c**)

these may not be optimal for classification. Therefore, we performed Bayesian optimization (BO) to adjust the hyperparameters of KDE along with those of the classifier (Table S1). BO was performed using a TPE sampler [34].

Figure 7 illustrates the classification performance metrics of the models evaluated using a 20% test split. The detailed classification results in the form of confusion matrices are presented in Fig. S6. Figure S7 demonstrates the classification results for various test sizes (30%, 50%, 70% and 80%). Overall, the trends in the classification performance were generally consistent. LR and SVM (Linear) using the zeroth-order lifetime showed the highest classification accuracy. These classification accuracies were higher than those of nonlinear classification models such as SVM (RBF), RF and XGBoost. As shown in Figs. 7, S6 and S7, the classification performances on the training data exceeded those on the test data, with this trend being particularly pronounced in tree-based models such as RF and XGBoost. These nonlinear classifiers

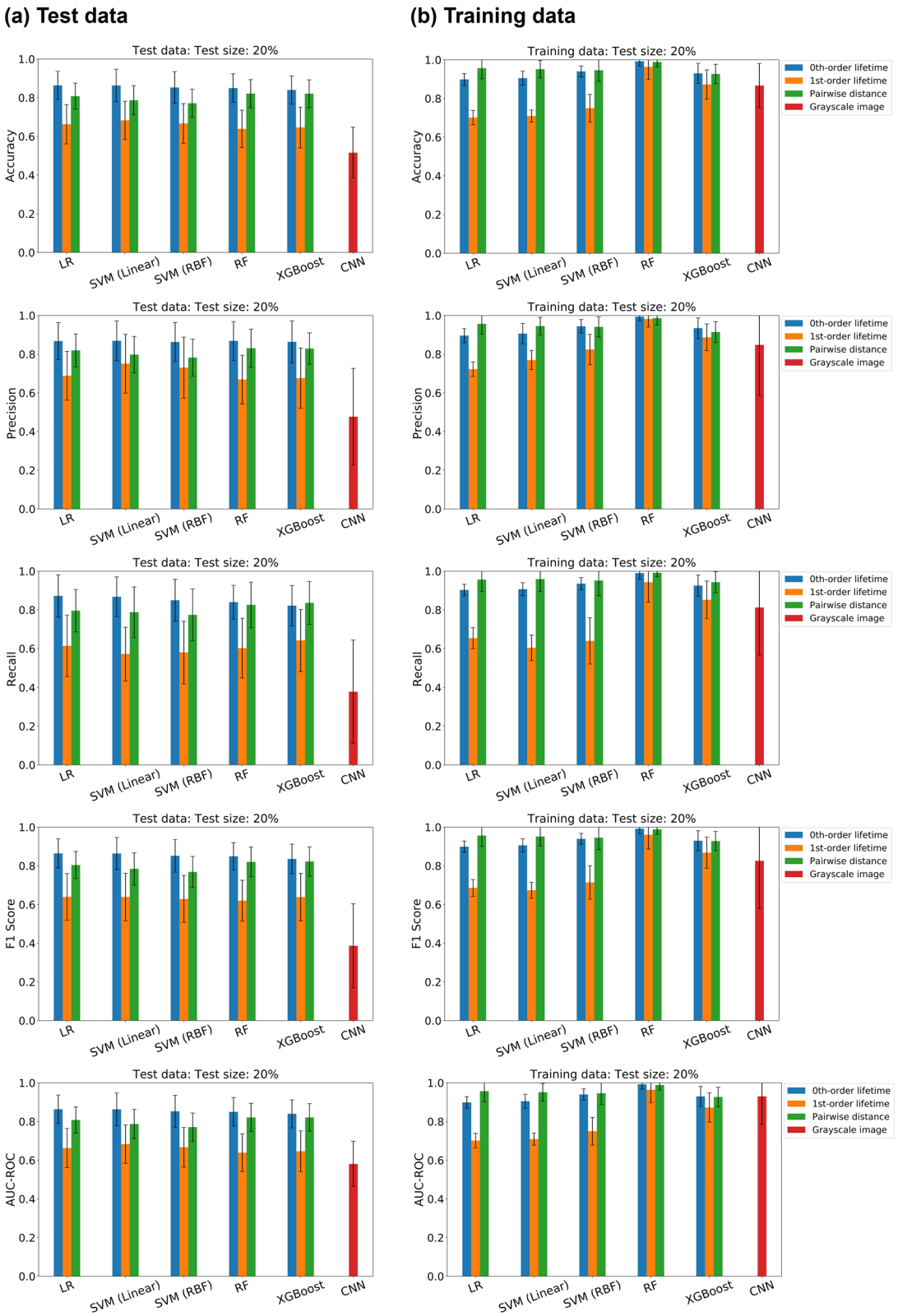


Fig. 7 Classification performance metrics (accuracy, precision, recall, F1 score and AUC-ROC) of different models for each explanatory variable on test (**a**) and training (**b**) datasets with 20% test size

are more complex than LR and SVM (Linear), likely resulting in higher risk of overfitting and consequently lower classification performance on the test data. On the other hand, using pairwise distance as an explanatory variable showed a trend of relatively high classification accuracy with RF, but the performance was comparable to or lower than that of the classification models using the zeroth-order lifetime. As shown in Fig. S7, although the classification accuracies on test data tended to decrease as the amount of training data was reduced, the models maintained superior performances compared to CNN with 80% test size. The CNN model exhibited lower and more unstable classification performance even after data augmentation, suggesting both the inherent difficulties in implementing CNN models for small datasets and the potential effectiveness of zeroth-order lifetime as a discriminative feature in this task.

As shown in Figs. 5 and 6, the differences in the distributions of the two classes are more prominent with the zeroth-order lifetime compared with the other explanatory variables. Therefore, relatively high classification accuracy was achieved using simple linear classification models. However, when using the pairwise distance as an explanatory variable, the differences between the two classes become slightly ambiguous, and the explanatory variable is higher-dimensional. This may have resulted in complex nonlinear classification models with a higher classification accuracy. Among all metrics, LR or SVM (Linear) tended to show the highest classification accuracy using the zeroth-order lifetime. On the other hand, the use of first-order lifetime did not result in high classification accuracy. One possible reason for this is the smaller number of one-dimensional holes compared to zero-dimensional holes. Although there are approximately 150 zero-dimensional holes in all binary images, the number of one-dimensional holes varies among binary images and is often less than 50. Therefore, the scarcity of one-dimensional holes may have led to variability in the first-order lifetime and decreased classification performance.

As part of an ablation study, classification models were constructed and evaluated without applying 2D Fused Lasso (Fig. S8). Although the overall model accuracy remained largely unaffected by the omission of 2D Fused Lasso in this dataset, linear models such as LR and SVM (Linear) tended to decrease classification performance on the test data when using pairwise distance as the explanatory variable.

3.3 Important topological features in classification

Using the zeroth-order lifetime as an explanatory variable resulted in LR with an accuracy of 86% and RF of 85% when splitting all data with a training/test ratio of 0.80/0.20. As relatively high classification performances were confirmed for LR and RF in the test set evaluation, these models were constructed with all images using a fivefold CV. The accuracies of LR and RF for the validation data in fivefold CV were 89% and 90%, respectively.

Figures S9 and S10 show the centered density distributions of zeroth-order lifetime and their PCA score plots for samples which LR misclassified. These figures indicate that the misclassified samples exhibit slightly different KDE patterns from the correctly classified samples, resulting in their shifted distribution in the design space. These results suggest that samples located near classification boundaries were likely to be misclassified.

Figure 8 shows the standard regression coefficients of LR for the zeroth-order lifetime. In this LR model, L1 regularization was selected as the hyperparameter "*penalty*," and the number of explanatory variables was reduced to four. This suggests that informative features for classification can be extracted from images using TDA. As shown in Fig. 8, the zeroth-order lifetime of approximately 0 and 7.59 significantly contributes to the classification. Moreover, when the probability density of the zeroth-order lifetime corresponds to these ranges, the predicted probabilities for cows and horses increase. However, information from the zeroth-order lifetime of approximately 24.3 and 27.34, which included more global patterns of pore arrangements, weakly contributed to the prediction. The standard regression coefficients suggest that the LR model uses zeroth-order lifetime of less than 10 as an important feature. This lifetime range provided pronounced differences between the two classes in the KDE results (Fig. 5).

RF resulted in a classification model using 13 explanatory variables. Figure 9 presents the results of the RF

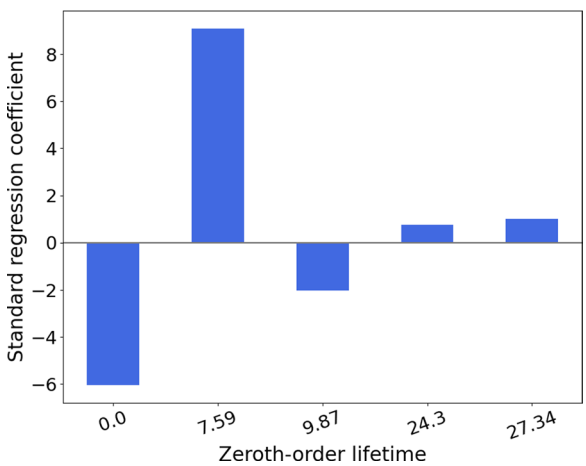


Fig. 8 Standard regression coefficients of the LR model for predictive probabilities of horse classes

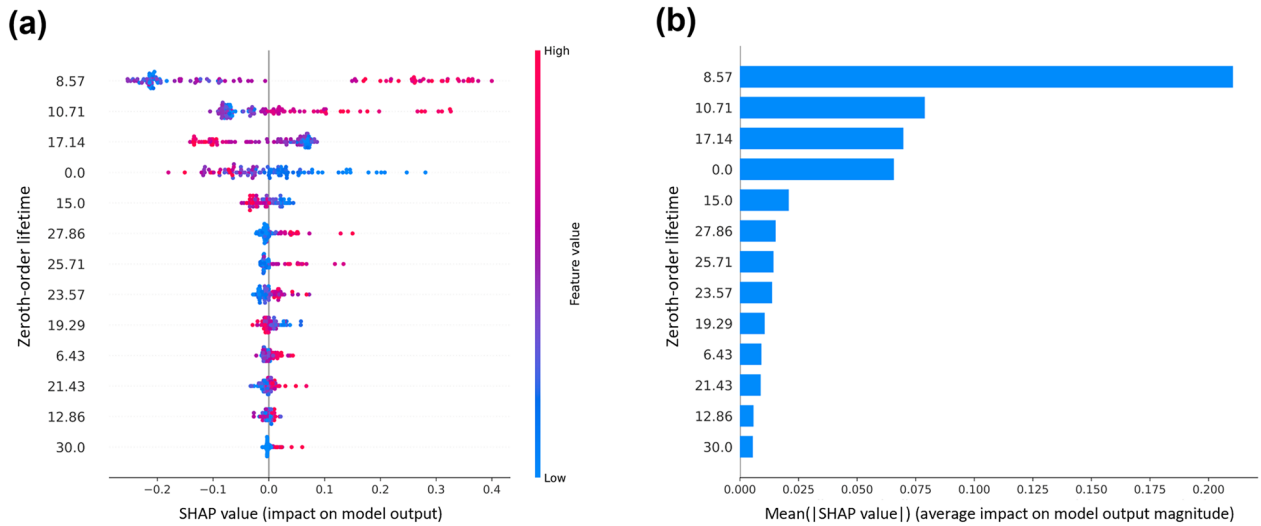


Fig. 9 The SHAP analysis results for the predictive probabilities of horse classes using the RF model: **a** Impact of explanatory variable on model output and **b** average impact of explanatory variable on model output magnitude

model interpretation using SHAP. SHAP is a model interpretation technique based on cooperative game theory that is applicable to any classification model. In this study, the SHAP values were computed using TreeExplainer [42], an explainer suitable for decision tree-based models. Figure 9a presents a summary plot of the contributions to the predictive probability of the horse class. The color represents the original feature value of each explanatory variable, whereas the horizontal axis indicates the contribution to the predicted probability of the horse class. The important explanatory variables for RF correspond to the areas of the zeroth-order lifetime where differences were observed between the two classes, as shown in Fig. 5a. These results suggest that RF has a reasonable basis for classification. In both LR and RF, information from the zeroth-order lifetime area less than 10 was estimated as an important explanatory variable. This suggests a statistically significant difference in the number and arrangement of hair pores in close proximity.

Finally, we visualized the geometric structures corresponding to a zeroth-order lifetime of less than 20. To select representative samples from each class, we clustered the KDE results (Fig. 5) using k-means ($k=3$) for each animal species, and extracted samples close to the centroids (Fig. 10). Considering the results of the elbow method, the cluster numbers were set to $k=3$ (Fig. S11).

Figure 11 shows the results of the inverse analysis of the binary images. An inverse analysis was performed in

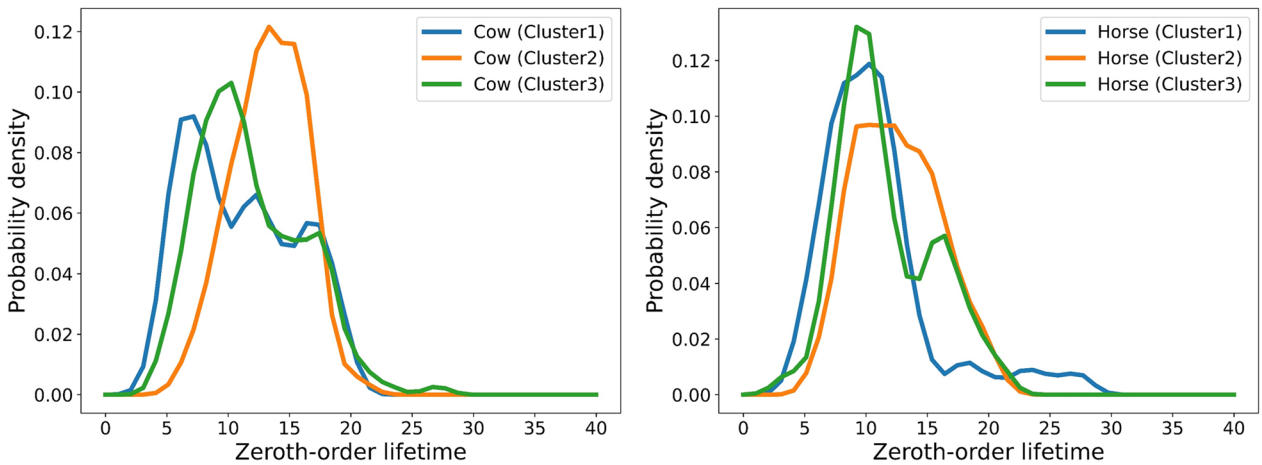


Fig. 10 The representative samples obtained using k-means clustering ($k=3$) for each class

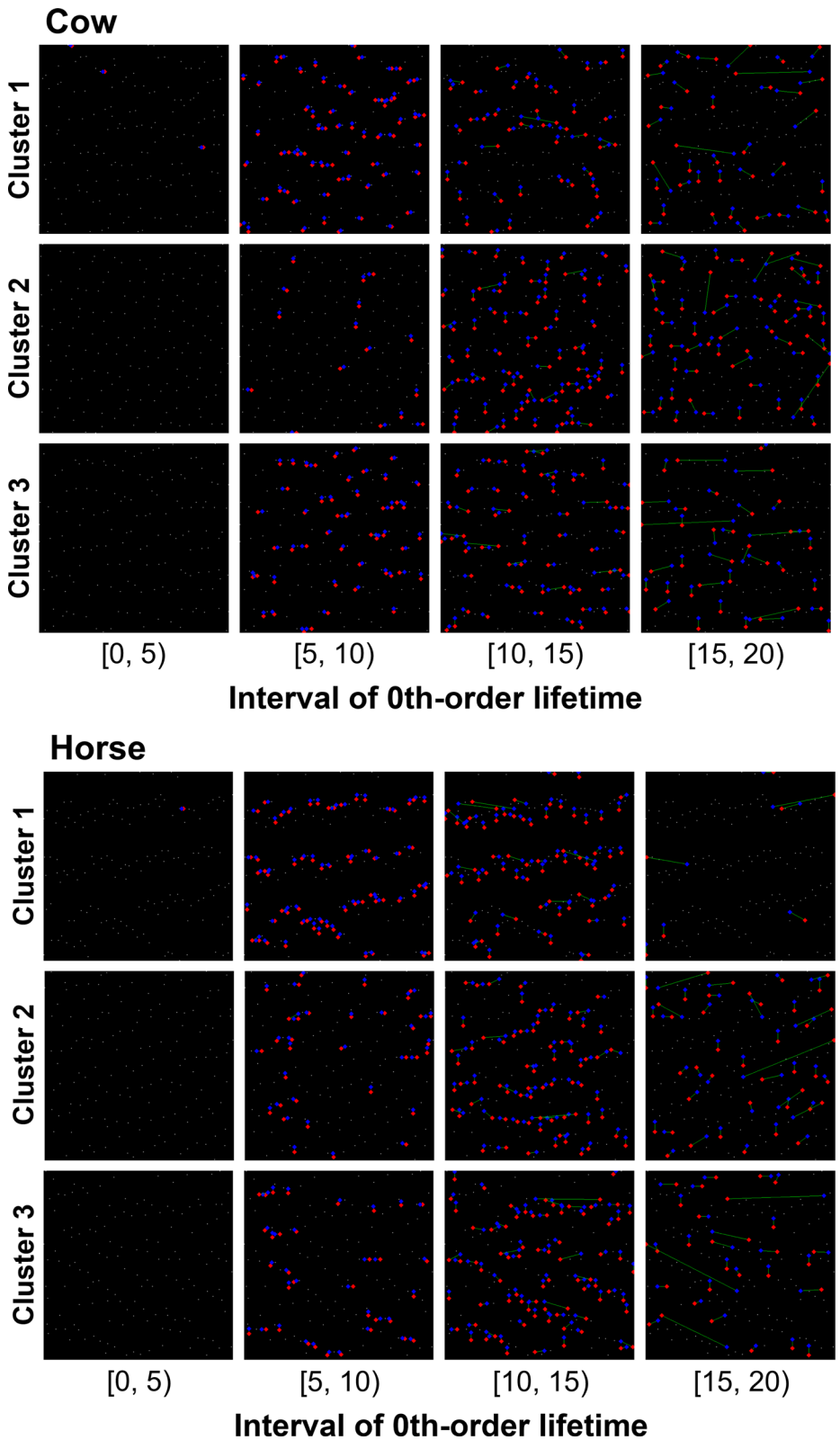


Fig. 11 The results of the inverse analysis for the representative samples of the clusters (red: birth pixel, blue: death pixel)

the lifetime range of [0, 20). The red points correspond to the birth positions of the zeroth-order homology, whereas the blue points correspond to their death positions. The birth and death pixels are connected by green lines. As expected, in the [0, 5) and [5, 10) lifetime ranges, birth–death pairs are generally formed by adjacent pixels, but some pairs are not necessarily formed by the closest pairs of pixels. Lifetimes in these ranges were important explanatory variables in both LR and RF models. By contrast, the variability in the distance between the birth and death pixels increased in the [10, 15) and [15, 20) ranges. This implies that the TDA extracted information on the global arrangement patterns of the binary images. In the RF model, lifetimes in these ranges were used as relatively important explanatory variables. However, these features were not used in the LR, but lifetimes of approximately 24 and 27 were used, with relatively low importance. The results of the model interpretations suggest that the probability densities of the lifetimes in the range of [0, 20), particularly in [0, 10.71] contain useful information for classification. Zeroth-order lifetime features not only extract local information but also global information in a specific area, providing different information from a simple pairwise distance. This study reveals that the zeroth-order lifetime can work effectively as a valuable descriptor.

By extracting information about hair pore arrangement patterns as zeroth-order lifetime, it was found that relatively high classification performance can be achieved under data-limited conditions. The proposed method assumes model construction using a small number of images and does not rely on information regarding hair pore density or pore size. Nevertheless, the topological features of hair pore arrangement enabled the construction of an accurate classification model.

4 Conclusions

In this study, a novel method for classifying cow and horse leather using digital microscope images was developed, specifically designed to handle small datasets effectively. The preprocessing pipeline included grayscale conversion, CLAHE, 2D Fused Lasso and brightness deconvolution, followed by PH computation. Hair pore coordinates were semiautomatically extracted using the results of PH computation, and binary images were generated from the coordinates. Based on the binary images, zeroth- and first-order lifetimes and pairwise distances were calculated. With a 20% test set allocation, linear classification models utilizing zeroth-order lifetime features achieved 86% accuracy on test data, outperforming both the pairwise distance approach and CNN with grayscale images. Model interpretation techniques revealed that lifetimes in the [0, 20) range were particularly important for classification, suggesting that topological features capture more complex pore arrangement patterns than simple pairwise distances. The proposed method demonstrates that PH enables effective semiautomatic feature extraction and robust classification, particularly effective when only limited training data are available.

Abbreviations

2D	Two-dimensional
AUC-ROC	Area under the receiver operating characteristic curve
BO	Bayesian optimization
CLAHE	Contrast limited adaptive histogram equalization
CNN	Convolutional neural network
CV	Cross-validation
KDE	Kernel density estimation
LR	Logistic regression
MSE	Mean squared error
PCA	Principal component analysis
PD	Persistence diagram
PD0	Zeroth-order persistence diagram
PH	Persistent homology
RBF	Radial basis function
RF	Random forest
SHAP	SHapley Additive explanations
SVM	Support vector machine
TDA	Topological data analysis
TPE	Tree-structured Parzen estimator
TV	Total variation
XGBoost	Extreme gradient boosting

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s42825-024-00187-1.

Additional file 1.

Acknowledgements
Not applicable.

Author contributions
TE was involved in conceptualization, investigation, methodology, data acquisition, software, writing of the paper. TO helped in conceptualization, investigation, methodology, data acquisition.

Funding
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Availability of data and materials
The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Research Division of Polymer Functional Materials, Osaka Research Institute of Industrial Science and Technology, Izumi, Osaka, Japan.

Received: 9 May 2024 Revised: 16 December 2024 Accepted: 19 Decem-

ber 2024
Published online: 14 February 2025

References

1. Merheb M, Vaiedelich S, Maniguet T, Hanni C. DNA for species identification in leather: fraud detection and endangered species protection. Res J Biotechnol. 2015;10(9):65–8.
2. Izuchi Y, Takashima T, Hatano N. Rapid and accurate identification of animal species in natural leather goods by liquid chromatography/mass spectrometry. Mass Spectrom (Tokyo). 2016;5(1):A0046.
3. Varghese A, Jain S, Prince AA, Jawahar M. Digital microscopic image sensing and processing for leather species identification. IEEE Sens J. 2020;20(17):10045–56.
4. Varghese A, Jawahar M, Prince AA. Learning species-definite features from digital microscopic leather images. Expert Syst Appl. 2023;220: 119971.
5. Jawahar M, Vani K, Babu NC. Leather species identification based on surface morphological characteristics using image analysis technique. J Am Leather Chem Assoc. 2016;111(8):308.
6. Varghese A, Jawahar M, Prince AA, Gandomi AH. Texture analysis on digital microscopic leather images for species identification. In Proceedings of the 9th International Conference on Soft Computing & Machine Intelligence (ISCMI). 2022;Toronto, ON, Canada:223–7.
7. Varghese A, Jawahar M, Prince AA. A study on deep learning models for automatic species identification from novel leather images. In Proceedings of the 2023 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT). 2023;BALI, Indonesia:25–30.
8. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
9. Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. Proc Adv Neural Inf Process Syst. 2012;25:1090–8.
10. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. 2019;6:60.
11. Xu M, Yoon S, Fuentes A, Park DS. A comprehensive survey of image augmentation techniques for deep learning. Pattern Recognit. 2023;137: 109347.
12. Edelsbrunner H, Harer J. Persistent homology-a survey. Contemp Math. 2008;453:257–82.
13. Chazal F, Michel B. An introduction to topological data analysis: fundamental and practical aspects for data scientists. Front Artif Intell. 2021;4: 667963.
14. Pun CS, Lee SX, Xia K. Persistent-homology-based machine learning: a survey and a comparative study. Artif Intell Rev. 2022;55:5169–213.
15. Nakamura T, Hiraoka Y, Hirata A, Escolar EG, Nishiura Y. Persistent homology and many-body atomic structure for medium-range order in the glass. Nanotechnology. 2015;26: 304001.
16. Ichinomiya T, Obayashi I, Hiraoka Y. Protein-folding analysis using features obtained by persistent homology. Biophys J. 2020;118:2926–37.
17. Krishnapriyan AS, Montoya J, Haranczyk M, Hummelshø J, Morozov D. Machine learning with persistent homology and chemical word embeddings improves prediction accuracy and interpretability in metal-organic frameworks. Sci Rep. 2021;11:8888.
18. Shimizu Y, Kurokawa T, Arai H, Washizu H. Higher-order structure of polymer melt described by persistent homology. Sci Rep. 2021;11:2274.
19. Ehiro T. Descriptor generation from Morgan fingerprint using persistent homology. SAR QSAR Environ Res. 2024;35:31–51.
20. Chulián S, Stolz BJ, Martínez-Rubio Á, Blázquez Goñi C, Rodríguez Gutiérrez JF, Caballero Velázquez T, Molinos Quintana Á, Ramírez Orellana M, Castillo Robleda A, Fuster Soler JL, Minguela Puras A, Martínez Sánchez MV, Rosa M, Pérez-García VM, Byrne HM. The shape of cancer relapse: Topological data analysis predicts recurrence in paediatric acute lymphoblastic leukaemia. PLoS Comput Biol. 2023;19: e1011329.
21. El-Yaagoubi AB, Chung MK, Ombao H. Topological data analysis for multivariate time series data. Entropy. 2023;25:1509.
22. Japan Leather Technology Association Leather Handbook Editorial Committee. Leather handbook. Tokyo: Jukai Shobo; 2005.
23. Stimper V, Bauer S, Ernstorfer R, Schölkopf B, Xian RP. Multidimensional contrast limited adaptive histogram equalization. IEEE Access. 2019;7:165437–47.
24. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. J R Stat Soc B. 2005;67:91–108.
25. Tibshirani RJ, Taylor J. The solution path of the generalized lasso. Ann Stat. 2011;39:1335–71.
26. Bradski G. The OpenCV library. Dr Dobb's J Softw Tools. 2000;120:122–5.
27. Su W, Boyd S, Candès EJ. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. J Mach Learn Res. 2016;17:1–43.
28. Matsumura T, Nagamura N, Akaho S, Nagata K, Ando Y. Spectrum adapted expectation-maximization algorithm for high-throughput peak shift analysis. Sci Tech Adv Mater. 2019;20:733–45.
29. Matsumura T, Nagamura N, Akaho S, Nagata K, Ando Y. Spectrum adapted expectation-conditional maximization algorithm for extending high–throughput peak separation method in XPS analysis. Sci Tech Adv Mater Method. 2021;1:45–55.
30. Obayashi I, Nakamura T, Hiraoka Y. Persistent homology analysis for materials research and persistent homology software: HomCloud. J Phys Soc Jpn. 2022;91: 091013.
31. Obayashi I, Hiraoka Y, Kimura M. Persistence diagrams with linear machine learning models. J Appl Comput Topol. 2018;1:421–49.
32. Węglarczyk S. Kernel density estimation and its application. ITM Web Conf. 2018;23:00037.
33. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2019. 2019;2623–31.
34. Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. In Advances in Neural Information Processing Systems. 2011;2546–54.
35. Stoltzfus JC. Logistic regression: a brief primer. Acad Emerg Med. 2011;18:1099–104.
36. Pisner DA, Schnyer DM. Chapter 6 - Support vector machine. In: Mechelli A, Vieira S, editors. Machine Learning: Methods and applications to brain disorders. Academic Press. 2020;101–21.
37. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
38. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining . 2016;New York, NY, USA: ACM:785–94.
39. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
40. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S. PyTorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst. 2019;32:8024–35.
41. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765–74.
42. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2:56–67.

Publisher's Note