applied sciences

MDPI

Article

Investigating Shape Variation Using Generalized Procrustes Analysis and Machine Learning

Wilfried Wöber [1,2,*], Lars Mehnen [3], Manuel Curto [1,4], Papius Dias Tibihika [1,5], Genanaw Tesfaye [6] and Harald Meimberg [1]

1 Department of Integrative Biology and Biodiversity Research, Institute of Integrative Conservation Research, University of Natural Resources and Life Sciences, Gregor Mendel Str. 33, 1080 Vienna, Austria; manuel.curto@boku.ac.at (M.C.); papiust@yahoo.com (P.D.T.); meimberg@boku.ac.at (H.M.)
2 Department Industrial Engineering, University of Applied Sciences Technikum Wien, 1200 Vienna, Austria
3 Department Computer Science, University of Applied Science Technikum Wien, 1200 Vienna, Austria; mehnen@technikum-wien.at
4 MARE, Marine and Environmental Sciences Centre, Faculty of Sciences, University of Lisbon, Campo Grande, 1749-016 Lisbon, Portugal
5 WorldFish, Lake Victoria Fisheries Organization, Jinja 1625, Uganda
6 EIAR—National Fisheries and Other Aquatic Life Research Center, Sebeta 100141, Ethiopia; bubuwiwi2008@gmail.com
* Correspondence: wilfried.woeber@technikum-wien.at; Tel.: +43-1-333-40-77-3157

Abstract: The biological investigation of a population's shape diversity using digital images is typically reliant on geometrical morphometrics, which is an approach based on user-defined landmarks. In contrast to this traditional approach, the progress in deep learning has led to numerous applications ranging from specimen identification to object detection. Typically, these models tend to become black boxes, which limits the usage of recent deep learning models for biological applications. However, the progress in explainable artificial intelligence tries to overcome this limitation. This study compares the explanatory power of unsupervised machine learning models to traditional landmark-based approaches for population structure investigation. We apply convolutional autoencoders as well as Gaussian process latent variable models to two Nile tilapia datasets to investigate the latent structure using consensus clustering. The explanatory factors of the machine learning models were extracted and compared to generalized Procrustes analysis. Hypotheses based on the Bayes factor are formulated to test the unambiguity of population diversity unveiled by the machine learning models. The findings show that it is possible to obtain biologically meaningful results relying on unsupervised machine learning. Furthermore we show that the machine learning models unveil latent structures close to the true population clusters. We found that 80% of the true population clusters relying on the convolutional autoencoder are significantly different to the remaining clusters. Similarly, 60% of the true population clusters relying on the Gaussian process latent variable model are significantly different. We conclude that the machine learning models outperform generalized Procrustes analysis, where 16% of the population cluster was found to be significantly different. However, the applied machine learning models still have limited biological explainability. We recommend further in-depth investigations to unveil the explanatory factors in the used model.

Keywords: generalized procrustes analysis; machine learning; convolutional autoencoder; Gaussian process latent variable models

Citation: Wöber, W.; Mehnen, L.; Curto, M.; Tibihika, P.D.; Tesfaye, G.; Meimberg, H. Investigating Shape Variation Using Generalized Procrustes Analysis and Machine Learning. *Appl. Sci.* **2022**, *12*, 3158. http://doi.org/10.3390/app12063158

Academic Editors: Miguel Ángel Maté-González and Julia Aramendi

Received: 18 February 2022
Accepted: 18 March 2022
Published: 20 March 2022

1. Introduction

The systematic visual inspection of specimen's morphological traits has a long history in biology, allowing divergent traits among species and populations of the same species to be defined and forming the field of morphometrics [1,2]. This inspection lately relies on digital images where landmarks are placed on diagnostic structures of the organism and

their relative position and distance are measured [3]. These manually placed landmarks provide some standardizations and automation of morphometrics analysis [4]. Their application to the field of fish biology has been especially successful because, due to their bilateral symmetry, it is possible to capture biologically significant traits in a two-dimensional image [5–8]. Nevertheless, placing landmarks in pictures can be quite labor-intensive and it requires prior biological knowledge (e.g., [4]). In this context, machine learning models can solve some of these limitations.

Since the publication of Alexnet [9] and its successors (e.g., VGG16 [10] or ResNet [11]), convolutional neuronal networks (CNNs) became a standard model for computer vision tasks. In contrast to landmark-based approaches where scientists carefully place the landmarks on images, CNNs learn to extract features in order to fulfil a user-defined task. CNNs are frequently implemented to solve object recognition or object detection applications. In fish biology, these models were successfully used for fish recognition [12–17]. Nevertheless, convolutional neuronal networks are black boxes and are hard to interpret [18]. Their behavior needs to be investigated after model training relying on prediction samples [19–22]. The latent explanatory factors learned by the CNN, which were previously discussed to be the key factor for reliable machine learning models [23], still cannot be unveiled. This means that biological factors contributing to the results of a neuronal network-based model cannot be extracted and analyzed.

This is a strong shortcoming in contrast to landmark-based analysis, where the landmarks can be interpreted in a biological manner. The authors of [24] compared landmark-based approaches to machine learning approaches for a Nile tilapia population classification task. The authors relied on supervised methods and found that the machine learning models, including CNNs, used image regions with no biological meaning that happened to be correlated with the specimen's population. This effect is known as clever-Hans predictors [18].

In contrast to the CNN-based approaches, the biological processing steps for visual diversity investigation differs. In geometric morphometry the landmarks are processed (e.g., generalized Procrustes analysis [25]) and the results are interpreted statistically. From a machine learning perspective, these landmarks represent manually defined features. These landmarks were defined without a priori information such as the specimen's location. However, the features learned by CNNs differ significantly and typically rely on a supervised training procedure. These learned features and the manually extracted landmarks cannot be discussed in the same way.

CNN features are typically learned using a priori information [9–11]. These features must not have biological meaning and may not be representative of a population's visual diversity [18]. In order to obtain reliable results, unsupervised machine learning was reported as an alternative to supervised methods [24,26,27]. Furthermore, to be able to compare machine learning-based visual diversity to landmark-based approaches, the feature extraction must be trained without a priori information in an unsupervised manner.

This study investigates the latent structure and visible diversity of populations in digital images unveiled by unsupervised machine learning models. We quantify the performance of the applied methods by measuring their capability to unveil known population clusters. These clusters were reported at the molecular genetic level [28–30] or are known due to geographical separation.

Since they breed among each other and tend to be exposed to similar environmental conditions, individuals of the same population are likely to share morphological features. In this study, we propose an unsupervised machine learning-based visual diversity investigation pipeline which is compared to landmark-based approaches. From two image datasets showing Nile tilapia specimens, landmarks are manually extracted and unsupervised machine learning models are trained to obtain features for each specimen. We hypothesize that:

Hypothesis 1. *The machine learning models learn biological meaningful features.*

Hypothesis 2. *These features have a higher relation to the actual population clusters in contrast to landmark-based approaches.*

The former hypothesis is evaluated by a visual inspection of the learned features. The latter hypothesis is evaluated in two ways. Initially, hypotheses tests are performed to investigate the relation between the features as well as landmarks to the known population clusters. Furthermore, we propose a novel non-parametric test to investigate the unambiguity of the known clusters in the learned feature space as well as in the extracted landmarks.

The contribution of this study is the investigation of the expressiveness of machine learning models in a biological context and the comparison of the results to landmark-based approaches. The hypotheses of this study are evaluated based on two Nile tilapia datasets that originate from specimens from Ethiopia and Uganda. These datasets were previously analyzed. In [24], the relation of populations from Ethiopia were investigated using supervised (deep) machine learning-based specimen classification. The authors were able to achieve a prediction accuracy above 90%. However, they showed that this accuracy was achieved using clever-Hans predictors, and the classifiers used biological uninformative parts of the image. Ref. [5] investigated Nile tilapia specimens from Uganda relying on landmark-based methods. The authors discussed population differences and showed overlapping population distributions.

To evaluate the hypotheses of this study, two unsupervised machine learning models are implemented. We use the Bayesian Gaussian process latent variable model [31] previously used for Nile tilapia images [24] and plant recognition [26]. Furthermore, we implement an unsupervised deep learning counterpart, namely a convolutional autoencoder [32]. These two models were chosen due to reported success in a biological context.

The remaining part of this study is structured as follows. Section 2 introduces the materials and methods used for the evaluation of the aforementioned hypotheses. Section 3 describes the results obtained by applying the proposed pipeline. Afterwards, the results are discussed before Section 5 summarizes this work.

2. Materials and Methods

To evaluate the research hypotheses of this study, two main strategies were implemented. Initially, the learned features were visualized and manually inspected. For this inspection, the existing biological knowledge represented by the proposed landmark positions was used. The biological explainability of all used models was investigated. For the purpose of this study, we defined the biological explainability as the ability of models to explain the reasoning process based on meaningful biological information. To this end, we used the explainability of generalized Procrustes analysis, relying on landmarks placed on specimens as the reference for biological explainability.

In addition to this visual inspection, the features and landmarks were investigated using hypothesis tests. Initially we used Spearman's rank correlation tests to investigate the relation between the features as well as landmarks to the known population clusters. However, these tests do not compare the capability of the machine learning models and landmark-based approaches to identify visible information which is useful to unveil the actual population clusters. In this study, we aimed to measure the unambiguity of known population clusters relying on processed landmarks or features obtained by machine learning models. The performance of the unsupervised machine learning models in contrast to the landmark-based approaches was evaluated by comparing the unveiled cluster unambiguity. To investigate this unambiguity of the known clusters in the learned feature space as well as in the extracted landmarks, we proposed a novel and non-parametric test. For this test, we created several population cluster hypotheses and used a novel Bayesian extension of consensus clustering [33] as well as the principle of self-similarity for the formulation of a Bayesian hypothesis test for population discriminability.

The technical core problem of this study was the quantification of a model's capability to unveil morphological structure. However, this quantification is a complex task due to

ambiguity. Different models and optimization strategies maybe result in biological meaningful morphological clusters using the same dataset [29]. Similarly, a model may unveil subpopulations, but fail to quantify obvious differences in water bodies and vice versa. This effect is visualized in Figure 1. On the left side, two dimensional representations of specimens (e.g., two principal coordinates of generalized Procrustes analysis features) are shown as blue points. The right side shows the visual representation of two models (red and green) with different global parameters ϕ and local parameters ψ. Using different optimization strategies, two valid cluster hypotheses, namely four clusters (red model) or two clusters (green model) in the two-dimensional space, maybe occur.

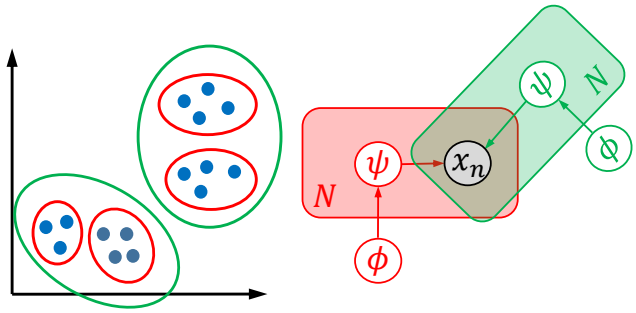


Figure 1. Two-dimensional representation of a specimen dataset and two valid clusters unveiling the latent structure (left side). Two models (right side, red and green model) with global parameter Φ and local parameter Ψ may result in two or four clusters.

Both cluster models unveil biological interpretable structures and may differ as a result of different mathematical formulations or optimization strategies. In order to be able to quantify methods, and inspired by infinite mixture models [34] as well as the idea of cluster ensemble [33,35], this study combined multiple population structure hypotheses unveiled in landmark and machine learning-based visual diversity data. These visual diversity data were generated relying on generalized Procrustes analysis and unsupervised machine learning models. We aimed to fuse the information of all population structure hypotheses. The combined results, as well as known clusters, were used to compare the landmark-based approaches with the machine learning approaches.

Our developed processing pipeline is shown in Figure 2.

The remaining part of this chapter introduces the used data, landmark processing methods, machine learning models and morphological diversity investigation. The Supplementary Materials (software as well as the used data including landmarks) is available under https://github.com/TW-Robotics/MorphoML (accessed on 17 February 2022).

2.1. Data Sources

This study relied on two image datasets, namely from Ethiopia (209 images, six populations) and Uganda [5] (462 images, 19 populations). All images were carefully gathered, prepared for digital processing and converted to grayscale images. The specimens in the images were cut out and resized to 224×96 pixels. A summary of the used image datasets is available in Table 1.

The population locations are visualized in Figure 3.

For the purpose of this study, the locations of the specimens were used to quantify the capability of the models to unveil meaningful structure. This approach was motivated by the previous work of [5,24], where the authors showed morphological differences for populations of Uganda and Ethiopia. However, if visible differences did not exist our approach would fail and no meaningful structure could be extracted.

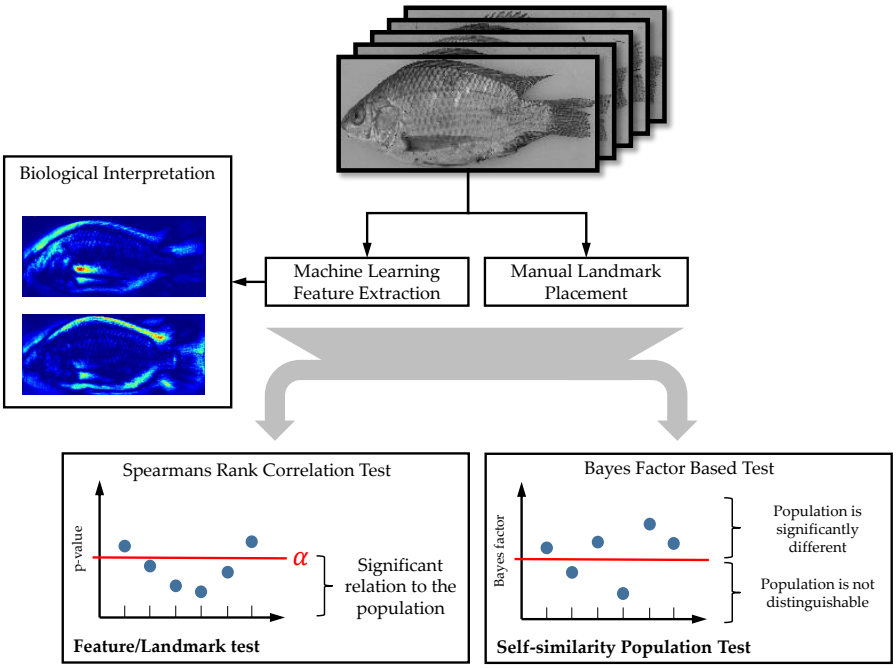


Figure 2. The pipeline used in this study to investigate the biological interpretation of the learned features as well as the statistical analysis of the discriminability of the known population clusters.

Table 1. Summary of image dataset from Ethiopia (209 specimens) and Uganda [5] (462 specimens). During this study, we used the location name as well as the abbreviation.

	Water Body	Abbr.	Nr. Spec.	Latitude	Longitude
Ethiopia	Chamo	Cham	36	5.83333	37.55
	Hawassa	Hawa	38	7.05	38.43333
	Koka	Koka	31	8.39197	39.07679
	Langano	Lang	26	7.61666	38.76666
	Tana	Tana	38	12.0166	37.29194
	Ziway	Ziwa	40	8.00083	38.82111
Uganda	Victoria Kakyanga	ViKak	28	−0.18079	32.29332
	Victoria Masese	ViM	28	0.4365	33.24081
	Victoria Gaba	ViG	23	0.25819	32.63727
	Victoria Sango Bay	ViSB	20	−0.86772	31.71332
	Victoria Kamuwunga	ViKam	16	−0.12747	31.93999
	Albert Ntoroko	AlN	22	1.05206	30.53464
	Albert Kyehooro	AlK	16	1.5099	30.9361
	George Hamukungu	Ge	34	−0.01739	30.08698
	Kazinga Channel Katungulu	KaC	30	−0.12541	30.04744
	Edward Kazinga	EdK	21	−0.20783	29.89252
	Edward Rwenshama	EdR	19	−0.40459	29.77283
	Kyoga Kibuye	KyK	32	1.40028	32.57949
	Kyoga Bukungu	KyB	3	1.43873	32.86809
	River Nile Kibuye	Ni	29	1.18734	32.96865
	Mulehe Musezero	Mu	27	−1.21345	29.72668
	Kayumbu Rugarambiro	Ka	28	−1.34679	29.78446
	Bangena Farm	BF	34	−1.25617	29.73622
	Sindi Farm	SF	22	−1.17578	30.06198
	Rwitabingi Farm	RF	30	0.97116	33.13924

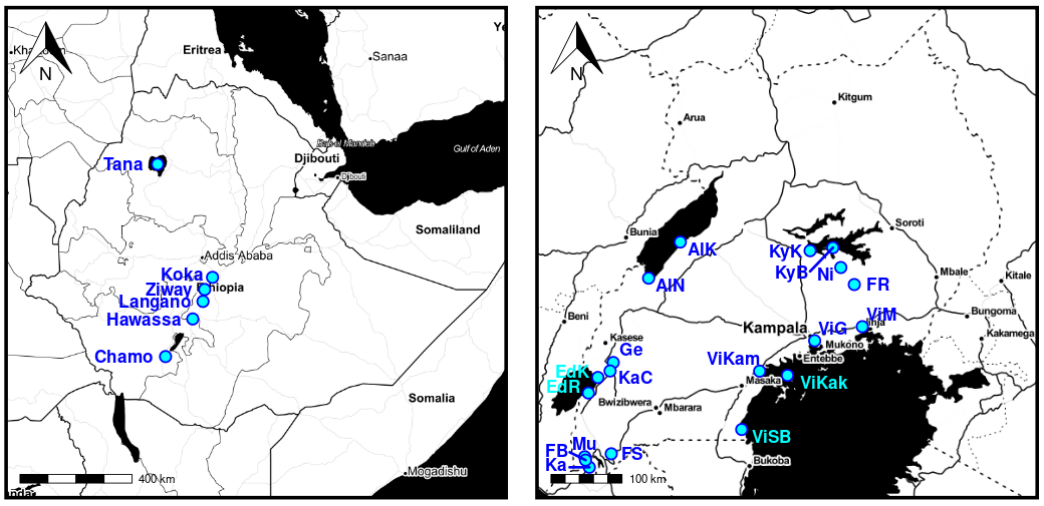


Figure 3. The locations in Ethiopia (**left**) and Uganda (**right**) used for this study. All images were also gathered in these locations. The cran R [36] package *rosm* [37] relying on the *OpenStreetMap* was used to generate the maps.

2.2. Visible Information Extraction

The image datasets from Uganda and Ethiopia were processed individually. We initially placed landmarks on the digital images and applied generalized Procrustes analysis (GPA) [25]. To obtain features from the machine learning models, we used the Bayesian Gaussian process latent variable model (B-GP-LVM) [31] previously used for Nile tilapia population classification [24]. Motivated by the success of deep learning, we used a convolutional autoencoder (cAE) [32] as a deep learning counterpart to the B-GP-LVM.

However, both machine learning models were based on image datasets $\{\mathcal{I}_1, \ldots, \mathcal{I}_n\}$ of n gray-scaled images. Each image $\mathcal{I}_j \in \mathbb{R}^{R \times C}$ consists of R rows and C columns. The B-GP-LVM as well as the cAE tackles the problem of estimating a latent representation $\vec{f}_j$ for the image $\mathcal{I}_j$. This latent representation is referred as *feature*. The models estimate the features for the specimens using different strategies and architectures. The features contain major information of the images and thus may represent visible characters representative of the populations.

2.2.1. Landmark Placement and Processing

The GPA relies on landmarks previously used for the Nile tilapia populations from Ethiopia [24] as well as Uganda [5]. 14 landmarks were used for Ethiopia and ten landmarks were used for Uganda. In order to investigate the impact of the number of landmarks, we used this different number of landmark positions for the images. The used landmark positions are shown in Figure 4 as well as Table 2.

OpenCV [38] was used to place the landmarks on the specimens. The landmarks were processed using the GPA implementation in the cran R [36] package *shapes* [39]. We used an F-test to investigate the relation between the Procrustes distances and the known specimen locations [40–43].

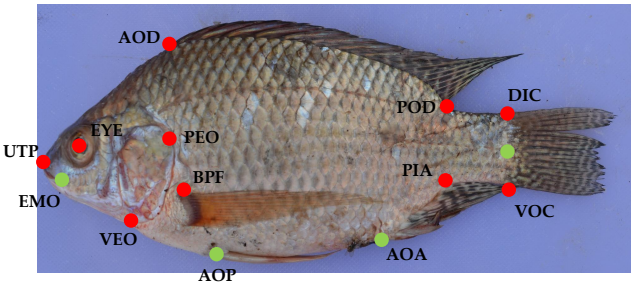


Figure 4. The used landmarks for GPA coordinate scaling. The green landmarks were not used for the images obtained in Uganda.

Table 2. Description of the landmarks used in this study. The asterix (*) indicates landmarks not used on Ugandan samples.

Landmark Name	Landmark Abbreviation
Upper tip of snout	UTP
Center of eye	EYE
Anterior insertion of dorsal fin	AOD
Posterior insertion of dorsal fin	POD
Dorsal insertion of caudal fin	DIC
Ventral insertion of caudal fin	VOC
Posterior insertion of anal fin	PIA
Dorsal base of pectoral fin	BPF
Most posterior edge of operculum	PEO
Ventral edge of operculum	VEO
Anterior insertion of anal fin *	AOA
Anterior insertion of pelvic fin *	AOP
Halfway between dorsal and ventral insertion of caudal fin *	HCF
Posterior end of mouth *	EMO

2.2.2. Bayesian Gaussian Process Latent Variable Model

A Gaussian process latent variable model (GP-LVM) [44,45] introduces a randomly initialized latent representation for each image sample and a set of approximated Gaussian processes [46] to recreate the images. In an optimization procedure, the parameters of the Gaussian processes as well as the latent image representations are adapted to the data. Using GP-LVM for images [47], each image $\mathcal{I}_j \in \mathbb{R}^{R \times C}$ is interpreted as a vector $\vec{y}_j \in \mathbb{R}^{R \cdot C \times 1}$ and the latent counterpart is a low-dimensional vector $\vec{x}_j$. The GP-LVM defines $p(\mathbf{Y}|\mathbf{X})$, which maps the features $p(\mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}(\vec{x}_n|\vec{0}, \mathbf{I})$ to $\mathbf{Y} = \begin{bmatrix} \vec{y}_1^T & \dots & \vec{y}_n^T \end{bmatrix}^T$. This mapping is performed using a set of sparse Gaussian processes [46].

Since this mapping is intractable, the authors of [31] proposed the Bayesian Gaussian process latent variable model. The authors used variational inference [48], where the prior $Q(\mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}(\vec{x}_n|\vec{\mu}_n, \Sigma_n)$ was used to optimize

$$\ln(p(\mathbf{Y})) \geq F(\mathbf{Q}) = \int_{\mathbf{X}} Q(\mathbf{X}) \ln\left(\frac{p(\mathbf{X})p(\mathbf{Y}|\mathbf{X})}{Q(\mathbf{X})}\right) d\mathbf{X}. \quad (1)$$

After optimization of Equation (1), we interpreted $\mathbb{E}(\vec{x}_j)$ as features for the image $\mathcal{I}_j$. $\mathbb{E}(\vec{x}_j)$ was obtained from the optimized Gaussian distribution in the latent space.

The number of features D as well as the number of auxiliary points used by the sparse Gaussian process are estimated using the log marginal likelihood $\ln(p(\mathbf{Y}))$ and the step-wise procedure:

1. Analyze $\mathbf{Y}$ using the principal component analysis [49]. Estimate the latent dimension D^*, which explains 75% of the data.
2. Keep D^* fixed and find the minimum number of auxiliary points reaching 95% of $\ln(p(\mathbf{Y}))$.

3. Keep the number of auxiliary points fixed and find the number of features D maximizing $\ln(p(\mathbf{Y}))$.

To visualize the visible characters learned by the B-GP-LVM, we applied the method of [24,47]. For this visualization of the D features, we generated n vectors $\vec{f}_{n,d} \in \mathbb{R}^{D\times 1}$ for each features. In this vector, we fixed all dimensions to the expectation except for the d-th dimension, which was set to the true specimen value. Following this procedure, n vectors for each feature could be generated. These vectors were projected to the image space. We calculated the pixel-wise variance and interpreted the resulting image as a visualization of the latent features space. For the interpretation of the images, we applied the pixel-wise hypothesis test for saliency maps (heatmaps) previously introduced in [24]. The visualization procedure is visualized in Figure 5.

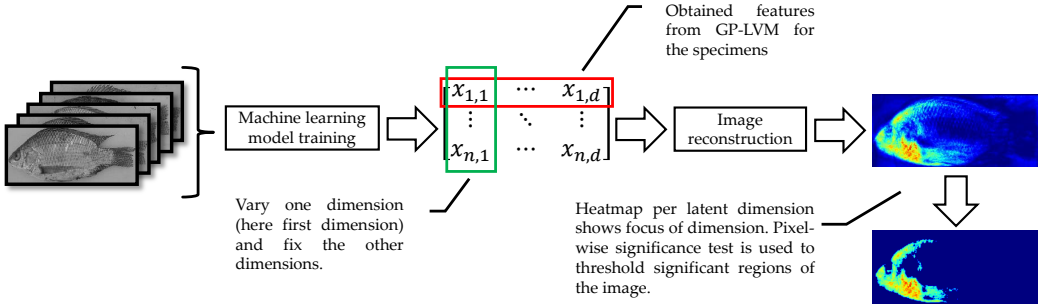


Figure 5. The procedure to visualize the latent dimensions. The image databases were used to obtain a latent space. For each dimension in the latent space, a heatmap was generated showing the variability generated by this dimension's variability.

We used the *GPy* [50] implementation of the B-GP-LVM relying on the radial base function (RBF) kernel including automatic relevance determination (ARD) [51]. The models were initialized using the principal component analysis. The optimization was based on *sklearns* L-BFGS-B optimizer [52]. We used the default parameters and analyzed the log marginal likelihood at each 10th step. If the log marginal likelihood had not increased at a minimum of 0.1%, we aborted the optimization.

2.2.3. Convolutional Autoencoder

The GP-LVM learns features by optimizing the probabilistic mapping $p(\mathbf{Y}|\mathbf{X})$, where $\mathbf{Y}$ represents flattened images. This methodology can be interpreted as the generalization of the principal component analysis [44]. An autoencoder (AE) [53] is a (deep) neuronal network-based counterpart to the GP-LVM. For visual problems, the AE is typically extended using convolution layers in order to obtain spatial relations. This model is referred as a convolutional autoencoder (cAE).

A cAE estimates features relying on an architecture of artificial neurons. This architecture is based on an encoder part which extracts features using $\vec{x}_j = g(\mathcal{I}_j)$ and an decoder part estimating and image reconstruction $\mathcal{R}_j = h(\vec{x}_j)$. During model training, a loss function $\mathcal{L}(\mathcal{I}, h(g(\mathcal{I})))$ is optimized. If this optimization converges, $h(g(\mathcal{I})) \approx \mathcal{I}$ and the encoder extracts useful features for image reconstruction. These features are not guaranteed to be independent.

Our implementation is based on Keras [54]. To this end, our encoder relies on the VGG-16 model [10] and the decoder is based on the inverted structure of the encoders architecture. We used an ImageNet [55] initialization of the encoder. The cAE was trained using the binary cross-entropy [53] loss. The loss was optimized using ADAM [56] implemented in Keras [57]. The default parameters were used. We trained models with $\{2, 5, 10, 25, 50, 75, 100, 125, 150, 200\}$ features relying on five iterations. The training was stopped after 1000 epochs.

The model selection was implemented by investigating the last 50 epochs of the training and tackled the identification of an appropriate number of features. We applied a

kernel density estimation relying on a Gaussian kernel. For each model, we extracted the maximum of the resulting density. We chose the model with the lowest maximum loss value obtained by the aforementioned procedure. Our model selection was implemented using the default kernel density estimation implementation in cran R [36] using the default parameters.

For visualization, we proposed an adaption of the same principle described in Section 2.2.2 for B-GP-LVM features. We fixed all elements in the feature vector to the mean value and varied the values for each dimension according to the obtained values. The vector was used to predict images using the decoder. We analyzed the pixel-wise variability of the obtained images.

2.2.4. Visual Interpretation of the Features

We obtained visual characters using the optimized B-GP-LVM as well as cAE. Nevertheless, both models reconstruct the image content using different strategies. This reconstruction includes background information (e.g., mounting frame) or location specific information (e.g., the mounting pin). As previously discussed by [18,24], machine learning algorithms may use biological uninformative image regions to obtain technical reasonable results. However, these image regions are not useful in the context of this study, where visible characters should be used to investigate the population's structure.

We followed the procedure proposed in [24] and carefully selected biological informative features manually. After model optimization and model selection, we visualized the learned features using the aforementioned methodology. All features containing non-biological information were rejected and not used for further investigation. The remaining features were analyzed using Spearman's rank correlation test [58], where for each feature we used j and population k $H_{0,j,k}$: *There is no relation between feature j and the population k*. For our visualization in the remaining part of this study, we showed $1 - p$ instead of the p-value for the machine learning-based models.

2.3. Multivariate Data Analysis

The data investigation methods and models discussed above mainly focus on the GPA coordinate and feature visualization as well as statistical tests. Nevertheless the aim of this study is the investigation of the latent structure of the used methodologies. The data complexity and population correlation relying on multivariate analysis of extracted GPA coordinates as well as machine learning features was investigated and visualized. For this analysis, the GPA coordinates as well as the cAE features were reduced to three dimensions relying on the principal component analysis (PCA) ([49], Chapter 12) as well as the fast independent component analysis (ICA) [59,60]. For the GP-LVM results, the first three independent dimension ranked relying on the relevance value of the kernel were used.

The visualization as well as correlation analysis was based on the *GGally* R package [61].

2.4. Investigation of Morphological Diversity

The application of the same models for structure investigation using the processed landmarks as well as the machine learning-based features may result in different population clusters. In genetics, this problem is tackled by seeking the optimal model using metrics such as cumulative ancestry contribution [62] or the log marginal likelihood [63,64]. However, these numerical values must not be useful for the biological question, and several models may contain useful information about the visible diversity of the data. Furthermore, the numerical optimization may differ using different data pre-processing methods such as GPA or machine learning.

To be able to compare the results obtained by the landmark-based methods and machine learning-based methods, we created several structure hypotheses and sought consensus in these models. We argue that this consensus contains the morphological structure unveiled by multiple hypotheses. The comparison was performed using the

fused hypotheses. In this study, we used Gaussian mixture model clustering [49] relying on different numbers of possible clusters to generate data structure hypothesis. The cluster models were fused by extending consensus clustering [33]. We empathize that any cluster model resulting in probability matrices may be used instead of Gaussian mixture models.

In the remaining subsections, the GPA scaled coordinates and machine learning features are referred to as "features".

2.4.1. Visible Features Clustering and Population Structure Investigation

The proposed clustering method is based on the Gaussian mixture model (GMM) [49]. The GMM is based on a set of k multivariate Gaussian distributions. Each specimen j represented by its features $\vec{f}_j$ is assigned to one of the k clusters. The model is based on

$$p(\mathbf{F}|\vec{\pi}, \mu, \Sigma) = \prod_{n=1}^{N} \sum_{j=1}^{k} \pi_j \mathcal{N}(\vec{f}_n|\vec{\mu}_j, \Sigma_j). \tag{2}$$

The affiliation of a specimen feature $\vec{f}_j$ to the cluster i is calculated by the variable $\vec{z} = (z_1, \ldots, z_k)$, which is a 1-of-K coded vector, and the conditional probability

$$p(z_i = 1|\vec{f}_j) = \frac{\pi_i \mathcal{N}(\vec{f}_j|\vec{\mu}_i, \Sigma_i)}{\sum_{m=1}^{k} \pi_m \mathcal{N}(\vec{f}_j|\vec{\mu}_m, \Sigma_m)}. \tag{3}$$

During optimization, the parameters $\{\Sigma_1, \ldots, \Sigma_k\}$, $\{\mu_1, \ldots, \mu_k\}$, $\vec{z}$ and $\vec{\pi}$ are optimized. After optimization, we use the class probability $p(z_i = m|\vec{f}_j)$ as an estimate for specimen j belongs to the cluster $m \in \{1, \ldots, k\}$.

We use the *mclust* [65] package relying on expectation maximization [49] for optimization. For the investigation of the morphological structure, k must be defined by the user. However, several strategies for k optimization exist, e.g., the Bayesian information criterion [66].

The optimization of the k parameter is a fundamental problem in genetics as well. In genetics the marginal log-likelihood [63], evidence lower bound [48] or biological motivated criterion [62] are used. Nevertheless, these optimization criteria are mainly focusing on technical parameters. Motivated by the facts,

1. That the used models may not be a good approximation for the unknown probability density functions,
2. The data are typically restricted as well as incomplete, and
3. Different k's may capture biological significant information on different scales [67].

We hypothesize that by finding consensus in a set of reasonable cluster models relevant pattern representative for the overall population can be obtained.

2.4.2. Consensus Clustering

Finding consensus in several cluster models or clustering ensembles [35] may be used to combine evidence unveiled by different models. The principle of co-association [33] is a model-free methodology, where the consensus of several cluster models is found by analyzing the pairwise occurrence of samples in the same cluster in several partitions. The cluster model i results in a partition $P_i = \{C_1^i, \ldots, C_k^i\}$, where C_m^n is the m-th cluster in the n-th partition. All models results in the set $\{P_1, \ldots, P_m\}$. The co-association (CA) of samples j and k relying on m cluster models is measured by

$$CA_{j,k} = \frac{1}{m} \sum_{t=1}^{m} \delta\big(P_t(\vec{x}_j), P_t(\vec{x}_k)\big). \tag{4}$$

The function $\delta(.)$ returns 1 if both samples happens to be in the same cluster in partition P_t. This methodology results in the co-association matrix, where the entries at j, k is $CA_{j,k}$.

After the investigation of all m clustering models, the co-association matrix contains a consensus about all models.

We propose a probabilistic extension of this method referred to as probabilistic co-association (pCA), where the cluster uncertainty is included. On one hand, this extension includes the probabilistic perspective of the structural investigation and allows on the other hand hypothesis testing. To investigate the pCA, the probabilistic formulation of Equation (4) is the probability of two samples m and n happens to be in the same cluster j of partition p. This probability is calculated using

$$p({}_pCA_{(m,n)}|\vec{f}_{m,n}) = p({}_pz_j^m = 1, {}_p z_j^n = 1|\vec{f}_m, \vec{f}_n) = p({}_pz_j^m = 1|\vec{f}_m)p({}_pz_j^n = 1|\vec{f}_n). \quad (5)$$

The variable ${}_pz_j^m$ indicates the affiliation of sample m to cluster j in partition p. Assuming that the p cluster models are independent given the features, we can estimate the co-association $CA_{m,n}$ between sample m and n by

$$p(CA_{m,n}|\vec{f}_{m,n}) = p({}_1CA_{m,n}, \ldots, {}_p CA_{m,n}|\vec{f}_m, \vec{f}_n) = \prod_{t=1}^{p} p({}_tCA_{m,n}|\vec{f}_{m,n}). \quad (6)$$

In this analysis, we assume that specimens belonging to the same population cluster happen to be in the same model cluster with a higher frequency as well as higher probability than specimens from different clusters. We visualize the conditional probability of $CA_{m,n}$ in a $n \times n$ matrix referred as the pCA matrix. Specimens that happen to be frequently in the same cluster appear bright in the entries of the pCA matrix.

Finally, the performance of the probabilistic consensus clustering for morphological data is evaluated. This evaluation is performed by analyzing the population intra co-association to the inter co-association. In a biological context, this analysis investigates the visual similarity of the specimens to the true population location and foreign population locations. We formulate a hypothesis test relying on the Bayes factor [68], where we evaluate the probabilistic consensus of inter-population specimens. The intra co-association is measured by analyzing the similarity of a specimen $i \in loc$ to all members of the location loc using the pCA. The inter location co-association is measured by analyzing the similarity $i \notin loc$ to all specimens belonging to another location. If the method used for visual information extraction results in useful information, the intra co-association will exceed the inter co-association. Note, that we use a priori knowledge in this test, namely the known specimen population location. The test is implemented using

$$\mathcal{B}_{loc} = \prod_{i=1}^{N} \frac{\sum_{n \in loc \setminus i} p(CA_{i,n}|\vec{f}_{i,n})}{\sum_{n \notin loc} p(CA_{i,n}|\vec{f}_{i,n})}. \quad (7)$$

The index $n \in loc \setminus i$ describes all members of the populations location loc without the actual specimen i of the population. Hence, we do not compare the visual similarity of the specimen to itself.

The evaluation of our hypothesis $H_{0,loc,model}$: *The locations morphological structure significantly differs from the other locations* for the location loc and given model (GPA, B-GP-LVM or cAE) relies on the Bayes factor. We found significant morphological differences (e.g., accept the hypothesis) between a location loc and the remaining specimens if $\mathcal{B}_{pop} > 10$ [68].

Our method was implemented in cran R [36]. For numerical stability, we analyzed the log of $p(CA_{m,n}|\vec{f}_{m,n})$ as well as $\log(\mathcal{B}_{loc})$ instead of raw probabilities. We analyzed $\{2, 3, \ldots, 45\}$ cluster partitions for Ethiopia and Uganda. To avoid numerical instabilities, we added a uniform distributed jitter using $\mathcal{U}(0.005, 0.01)$, which is the probability of 0.5% to 1.0% that the specimens are in the same population.

3. Experimental Results

This section initially provides insight into information extracted relying on the landmark-based approach and machine learning models. Afterwards, the structure investigation as well as the result of the hypothesis test are presented.

3.1. Visible Diversity Relying on Landmarks

The landmarks were manually placed on the digital images. Analysis for specimens from Ethiopia was based on 14 landmarks, while for specimens from Uganda it was performed with ten landmarks. The result of the GPA feature scaling is illustrated in Figure 6. For the purpose of visibility and readability, the water bodies of Victoria, Albert, Edward and Kyoga were visualized together.

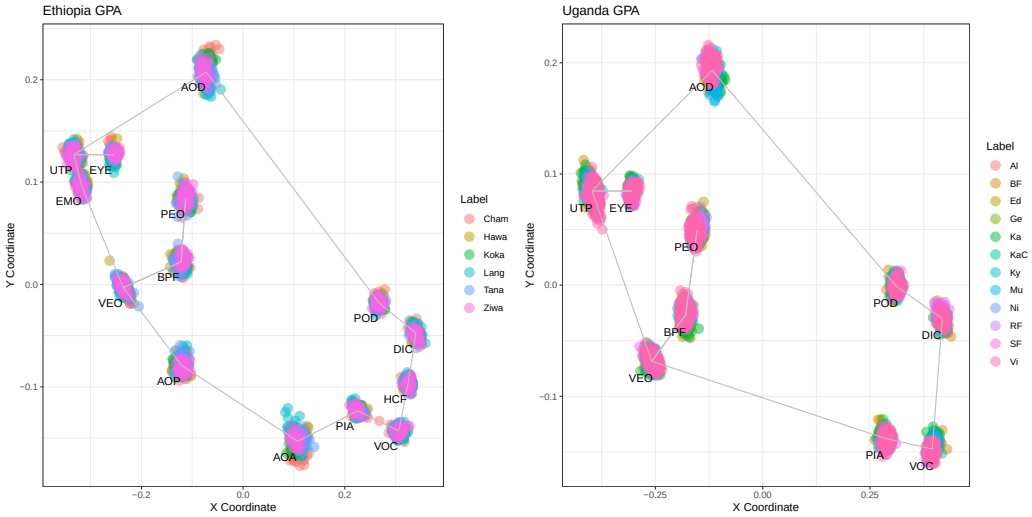
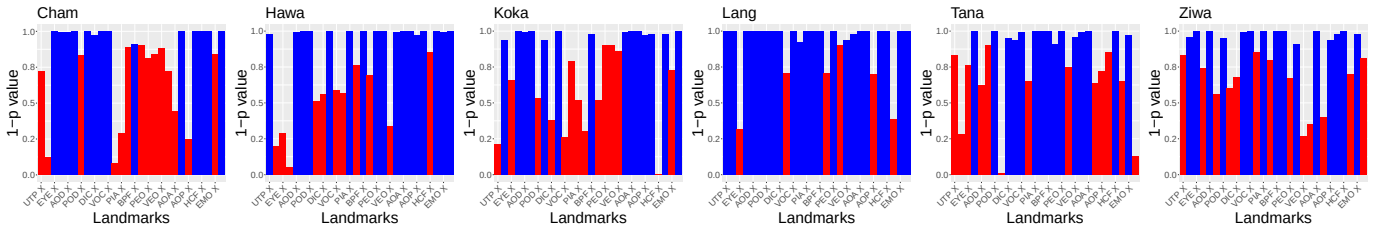


Figure 6. Visualization of the GPA coordinates for the population of Ethiopia (**left**) and Uganda (**right**).

An F-test for the investigation of the relation of the population to the coordinates relying on the Procrustes distance [40–43] showed significant relations between the locations and the Procrustes distance with a p-value below 0.01 for both datasets.

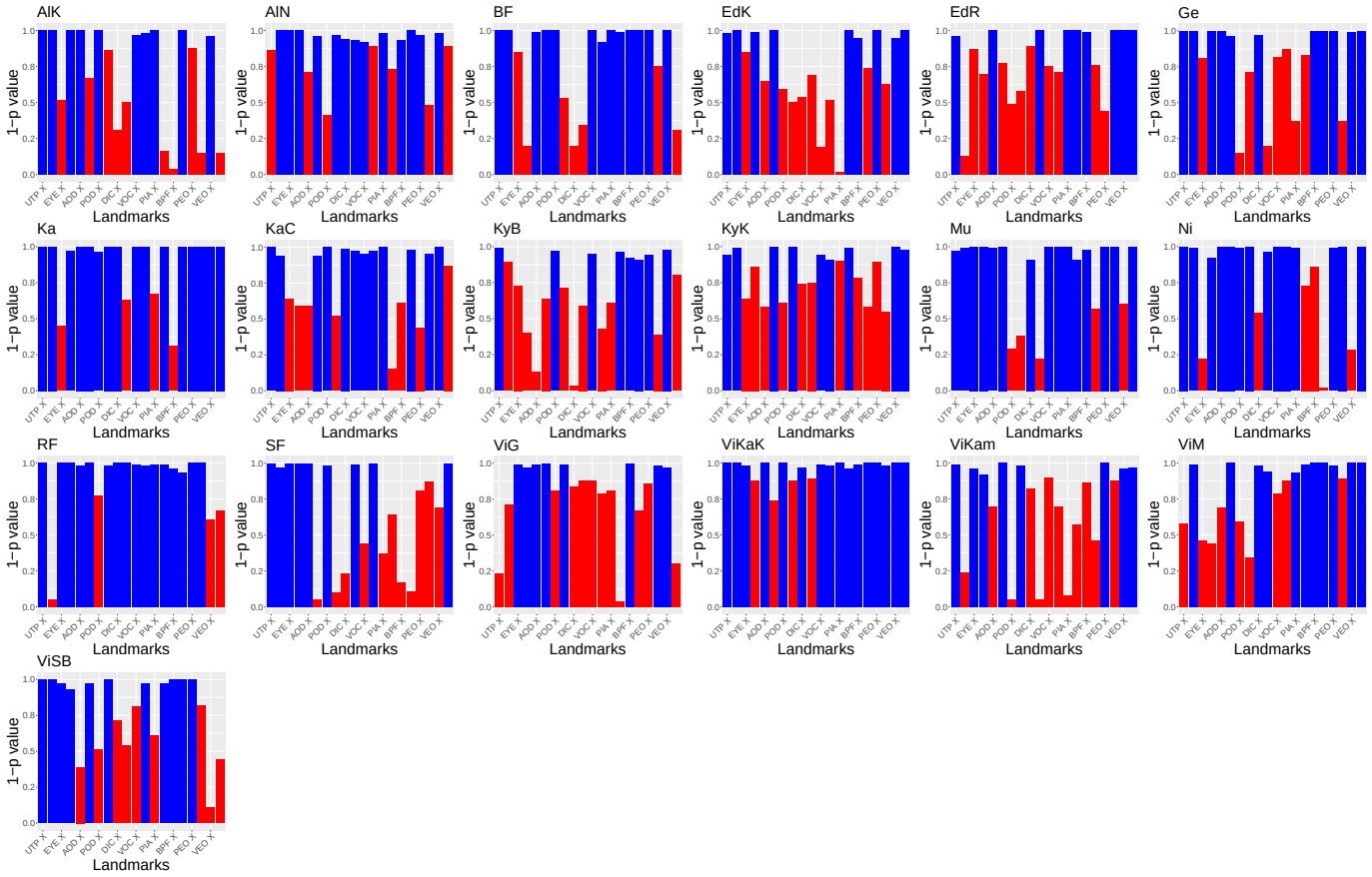
Furthermore, the relation of the X and Y coordinates of the landmarks to the population locations was tested. The p-values of Spearman's rank correlation test are visible in Figure 7. For the purpose of visibility, $1 - p$ is shown. Landmarks above $\alpha = 0.1$ are visualized in red.

The hypothesis test results indicate that there are significant relations between the X and Y GPA coordinates and the locations of the populations. No differences between the usage of ten or 14 landmarks were found. However, these tests did not unveil the differences between the populations relying on the GPA data.



(**a**)

Figure 7. *Cont.*



(**b**)

Figure 7. Visualization of Spearman's rank correlation test for the relation between GPA coordinates and the specimens population. $1 - p$ is shown instead of the p-value. A p-value below $\alpha = 0.1$ indicates a significant correlation of the GPA landmarks coordinate to the population. GPA coordinates with a p-value above $\alpha = 0.1$ are shown as red bars. (**a**) Result of Spearman's rank correlation test for the relation between the GPA coordinate and population locations in Ethiopia. (**b**) Result of Spearman's rank correlation test for the relation between the GPA coordinate and population locations in Uganda.

3.2. Visible Diversity Relying on Machine Learning Models

This section presents the results of the learning procedures as well as the visualization and biological interpretation of the feature vectors.

3.2.1. Gaussian Process Latent Variable Model

The GP-LVM was optimized using the aforementioned procedure. This optimization approach resulted in 125 features as well as 200 auxiliary points for Ethiopia and 50 features as well as 125 auxiliary points for Uganda. Afterwards, noisy features and all features focusing on background information such as mounting pins or specimen fixtures were removed using a visual inspection of the features. After this manual procedure, 26 features for Ethiopia and 22 for Uganda remained in the feature set.

The relation of the remaining features to the population's locations were tested using Spearman's rank correlation test. The results of these tests as well as the visualization of the used features are shown in Figures 8 and 9. This visualization shows the eight GP-LVM features with the highest ARD value of the optimized kernel. For better visualization $1 - p$ is visualized instead of the p-value. The features above $\alpha = 0.1$ are indicated with red bars.

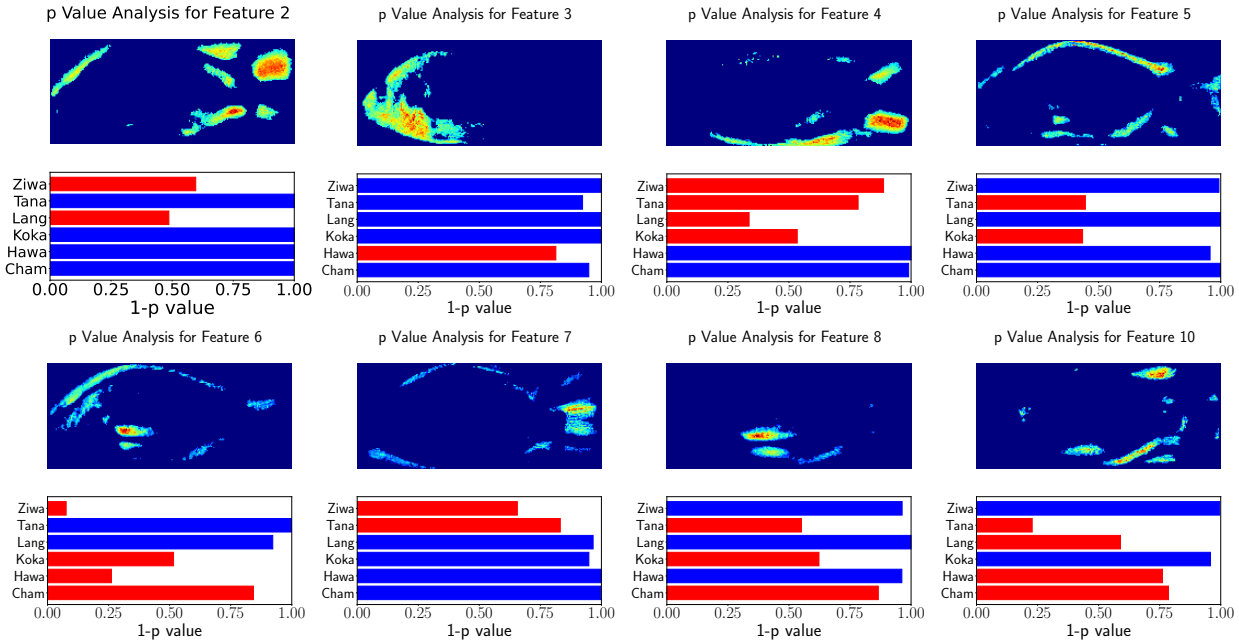


Figure 8. Result of Spearman's rank correlation test for the relation between the manually selected GP-LVM features and population locations in Ethiopia. The features with highest ARD values are shown. Features with a p-value above $\alpha = 0.1$ are shown as red bars.

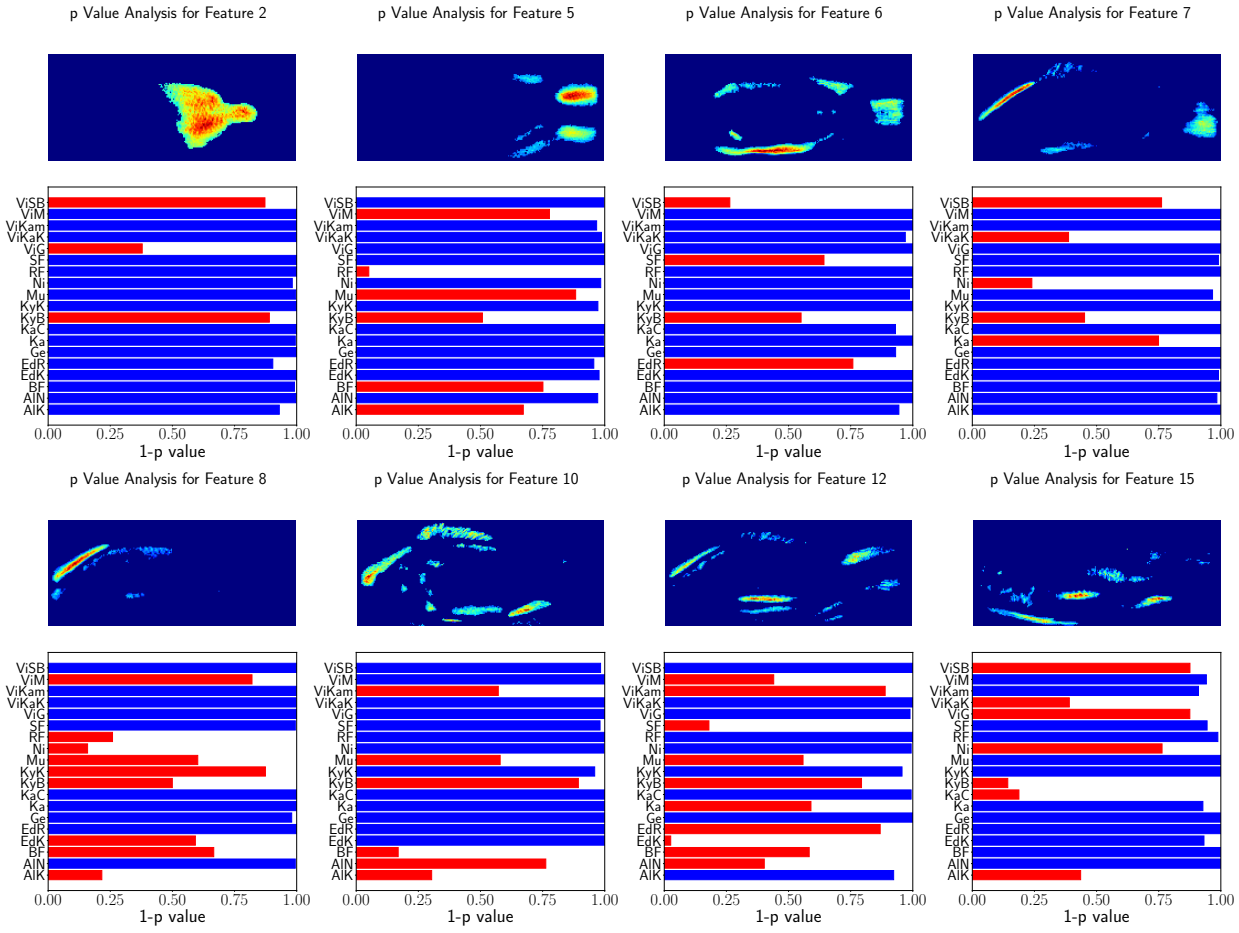


Figure 9. Result of Spearman's rank correlation test for the relation between the manually selected GP-LVM features and population locations in Uganda. The features with highest ARD values are shown. Features with a p-value above $\alpha = 0.1$ are shown as red bars.

The analysis of both datasets results in similar image regions. These image regions are in a similar position as the landmarks used for GPA. However, in contrast to the GPA coordinates, the heatmaps show the variability of image regions. While the GPA relies on the variability of discrete points, the GP-LVM analysis results in image regions in which the variability was tested to have a high relation to the population locations. The head and caudal fin region can clearly be seen in the visualization with a significant relation to the population's location.

3.2.2. Convolutional Autoencoder

The optimization of the number of features used in the cAE was performed using the investigation of the loss after optimization discussed above. This optimization results in 25 features for Ethiopia and 100 features for Uganda. All extracted features were visualized relying on the GP-LVM feature visualization technique adapted for cAE [26]. These features were manually investigated in a similar manner to the GP-LVM features. The selection procedure resulted in 23 features for Ethiopia and 46 features for Uganda. The number of features for Ethiopia was similar to the GP-LVM results. However, the number of features for Uganda exceeded the GP-LVM model selection.

Furthermore, the feature relation to the population location was tested using Spearman's rank correlation test. This procedure is visualized in Figures 10 and 11 for randomly selected features. Similarly, for better visualization $1 - p$ was is visualized instead of the p value of Spearman's rank correlation test, and the features above $\alpha = 0.1$ are indicated with red bars.

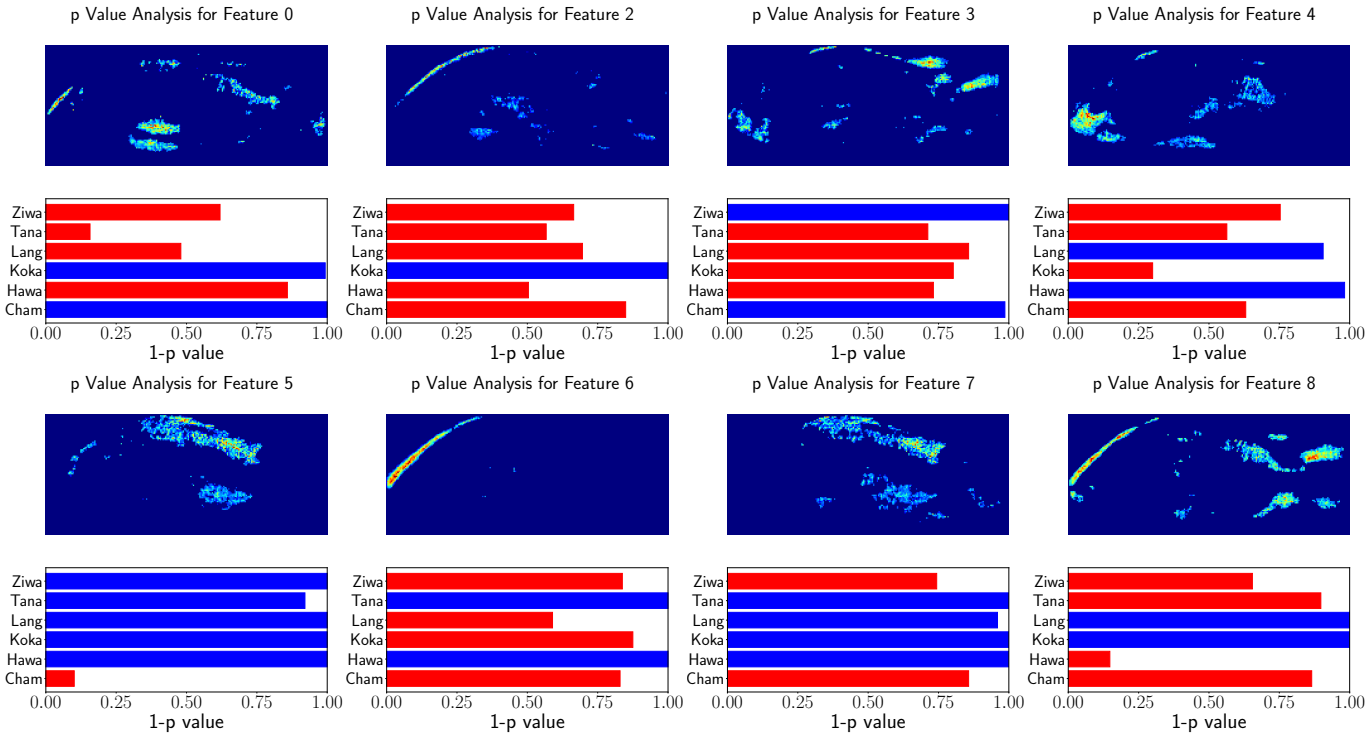


Figure 10. Result of Spearman's rank correlation test for the relation between the manually selected cAE features and population locations in Ethiopia. Randomly selected features are visualized. Features with a p-value above $\alpha = 0.1$ are shown as red bars.

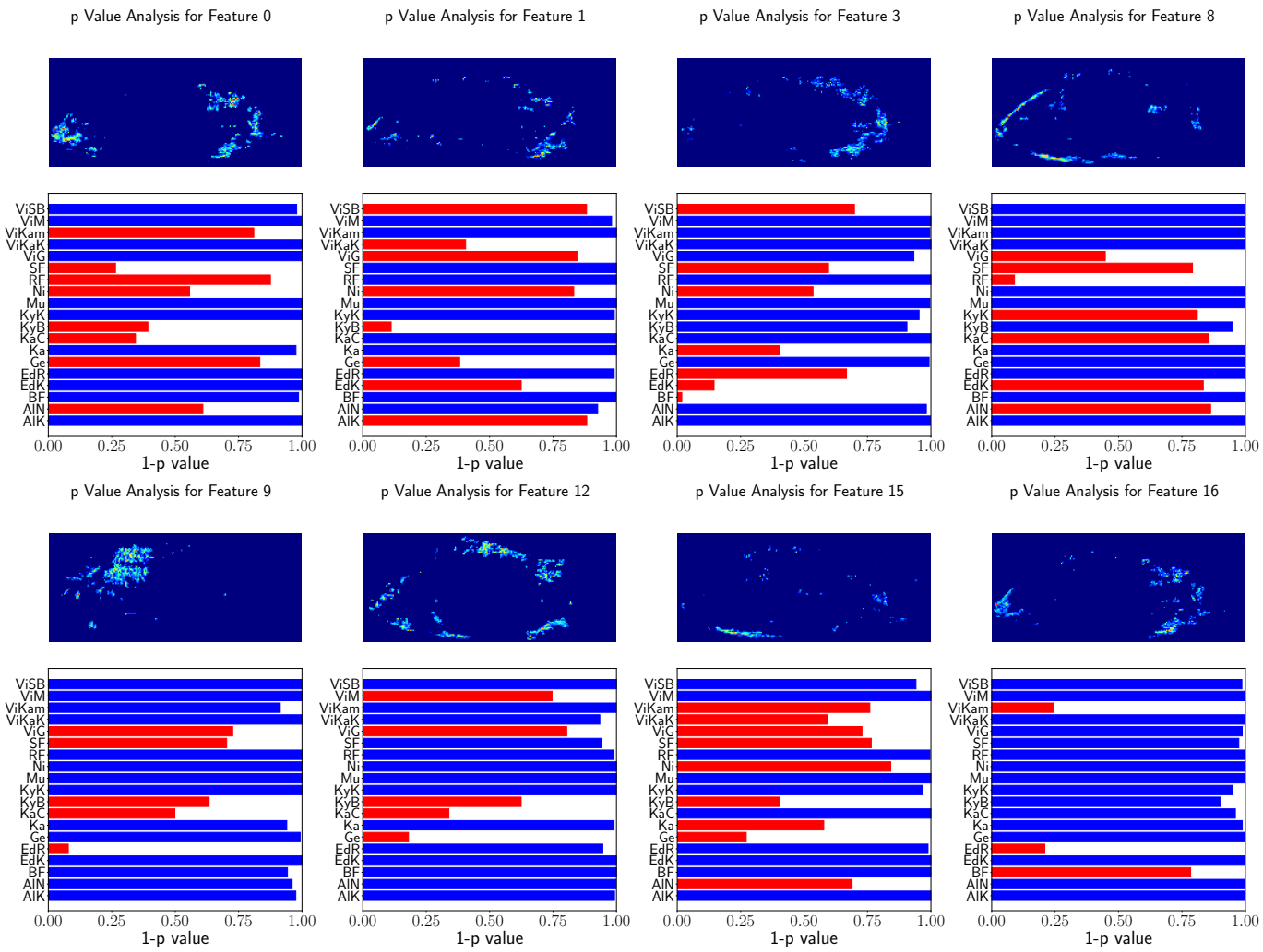


Figure 11. Results of Spearman's rank correlation test for the relation between the manually selected AE features and population locations in Uganda. Randomly selected features are visualized. Features with a p-value above $\alpha = 0.1$ are shown as red bars.

Again, the heatmaps result in similar image regions used for GPA. However, the heatmaps are noisy and not as clear as the GP-LVM heatmaps. Nevertheless, the features show significant relation to the population locations.

3.3. Multivariate Data Analysis

The multivariate analysis is shown in Figure 12. All applied methodologies suffer from overlapping population locations. Nevertheless, several populations such as Tana, Langano or Chamo in Ethiopia do show different densities. However, relying on the presented low-dimensional data, no population location is separable.

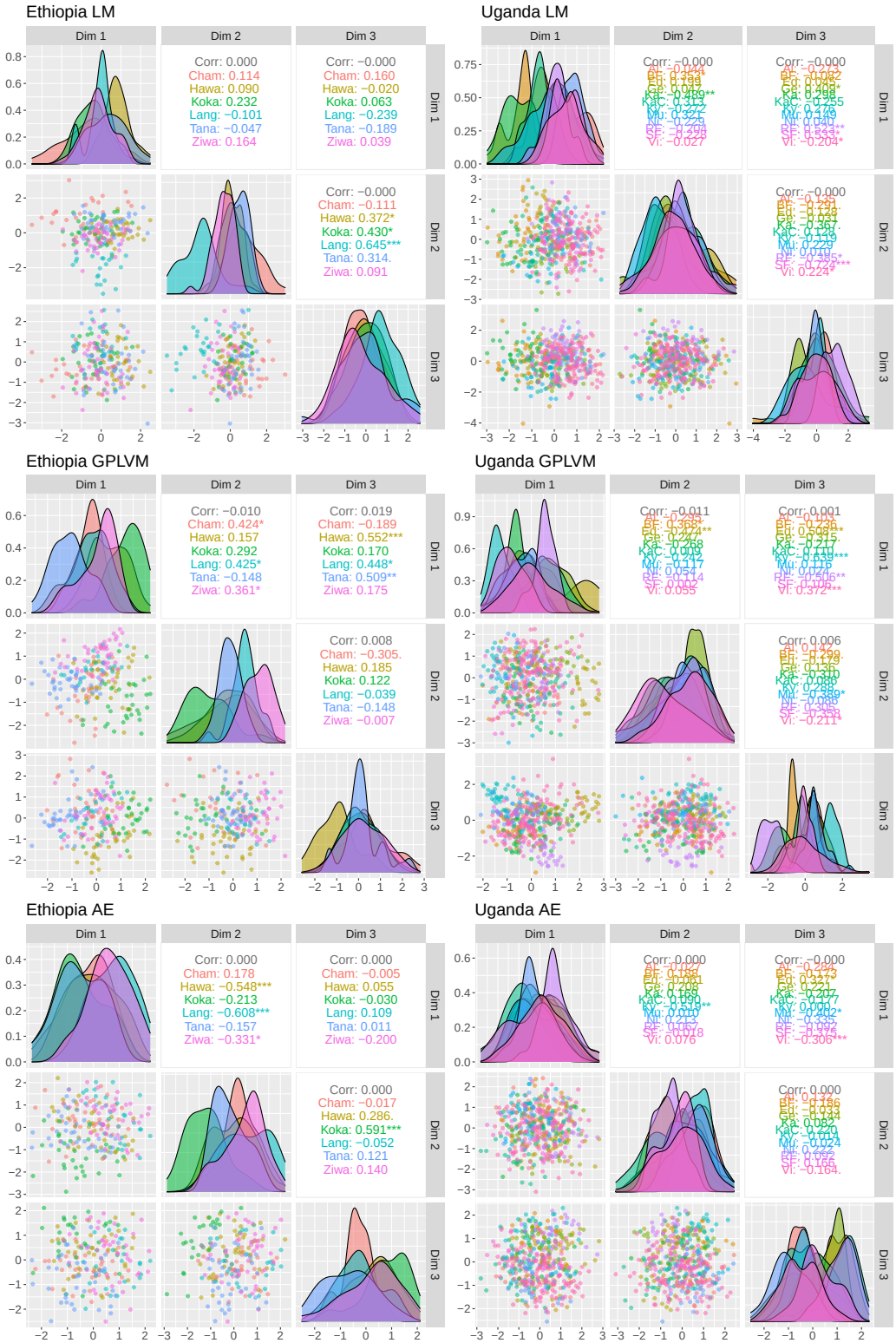


Figure 12. Visualization of the multivariate analysis of Ethiopia (**left column**) and Uganda (**right column**). The figures show the PCA/ICA reduced landmarks (**top**), GP-LVM features (**middle**) as well as reduced cAE features (**bottom**). The symbols '***', '**', '*' as well as '.' next to the numeric correlation values indicates significant levels below 0.001, 0.01, 0.05 and 0.1. If no symbol is given, the significance level was obtained to be larger than 0.1.

3.4. Latent Structure Investigation Relying on pCA

The pCA result for GPA for both datasets is summarized in in the pCA matrix as well as the Bayes factor plot in Figure 13.

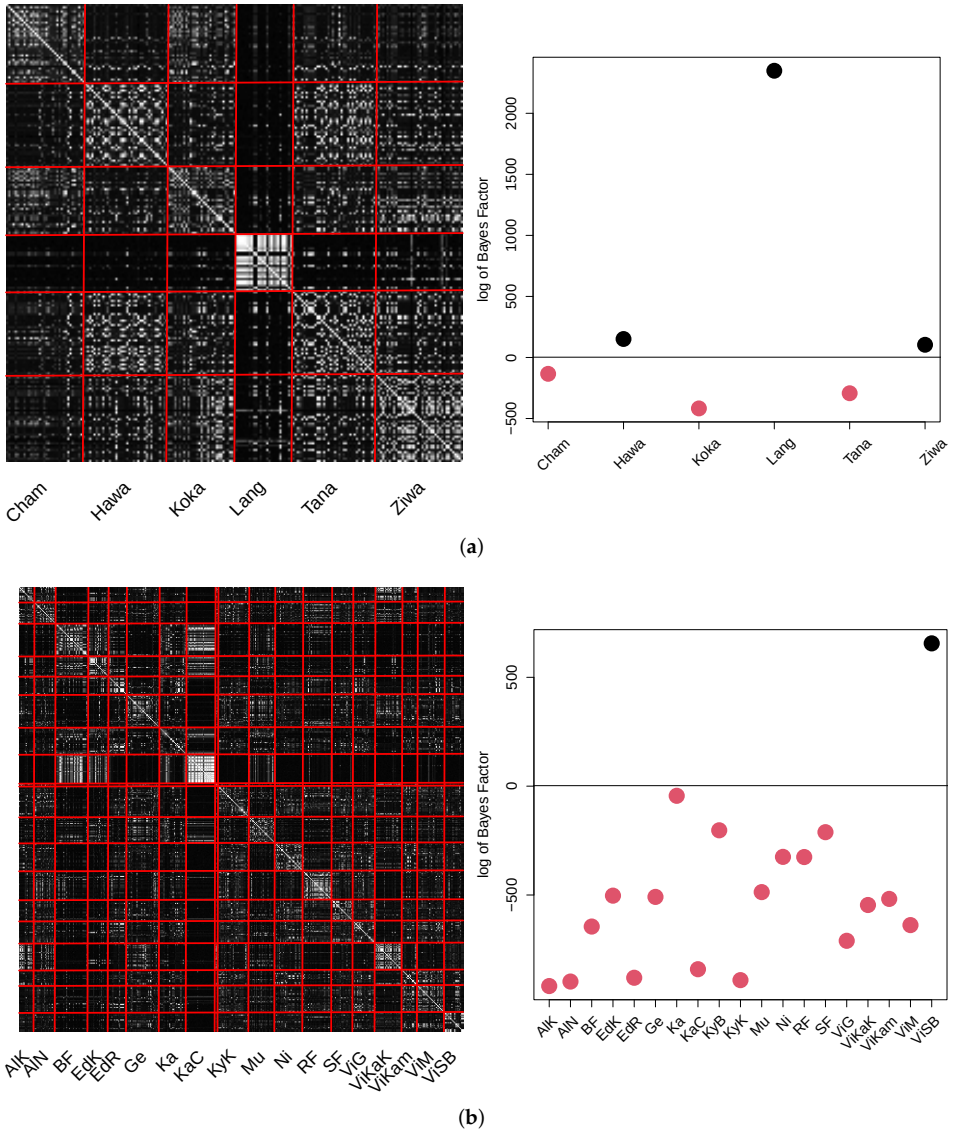


Figure 13. Visualization of the results obtained by pCA and the Bayes factor hypothesis test relying on GPA. Both pCA matrices show minor visible structure. Three of six locations of Ethiopia were found to be significantly different to the remaining populations. Similarly, one out of nineteen locations were identified to be significantly different in Uganda's locations. (**a**) pCA and Bayes factor results for Ethiopia relying on GPA scaling. (**b**) pCA and Bayes factor results for Uganda relying on GPA scaling. The KyB population label in the pCA matrix was removed due to readability.

The results show that minor individual population location structure was identified to be different. Three out of six locations in Ethiopia (Hawassa, Langano and Ziway) and one out of nineteen locations in Uganda (Victoria Sango Bay) were identified to be significantly different. We emphasize that the pCA matrix visualization may lead to incomprehensible results in the Bayes factor analysis due to the brightness. The Bayes factor investigation is based on the comparison of the intra-location similarity. Thus this value decreases, even if minor obvious relations were measured in the remaining locations.

The results obtained by the B-GP-LVM are visualized in Figure 14.

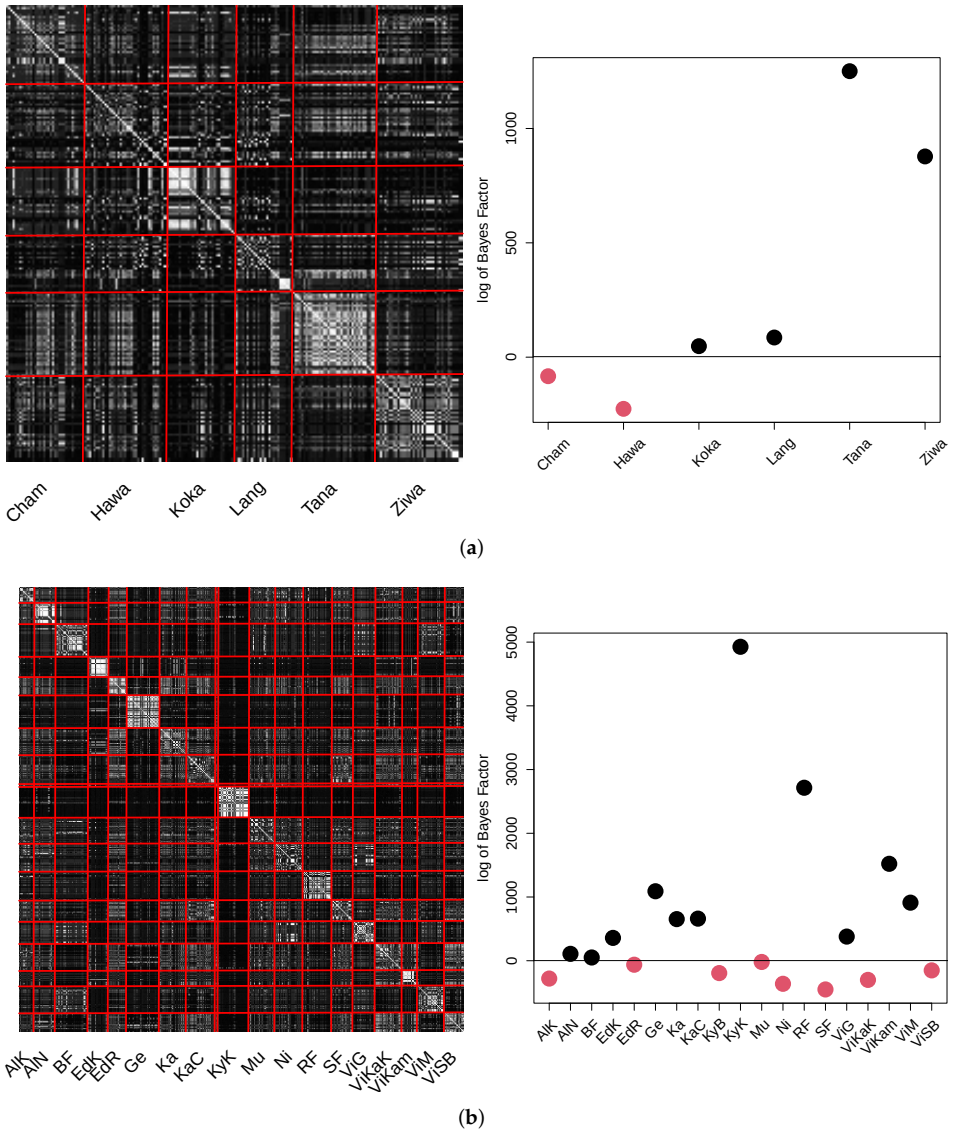


Figure 14. Visualization of the results obtained by pCA and the Bayes factor hypothesis test relying on GP-LVM. Both pCA matrices show visible structure. Four of six locations in Ethiopia were found to be significantly different to the remaining locations. Eleven out of nineteen locations were identified to be significantly different in Uganda's locations. (**a**) pCA and Bayes factor results for Ethiopia relying on GP-LVM. (**b**) pCA and Bayes factor results for Uganda relying on GP-LVM. The KyB population label in the pCA matrix was removed due to readability.

The pCA matrix for Ethiopia appears to be noisy. However, the individual probability values of the intra population locations exceed the probability values of inter population locations. This led to four out of six populations in Ethiopia which were significantly different to the other populations. Different to the GPA, lake Tana appears to be significantly different to the remaining locations. This significant difference and distinctiveness of the Lake Tana population has also been reported at molecular genetic level [28,29]. The pCA matrix for Uganda shows visible structure for the individual locations. The different locations in the same water bodies (e.g., the lake Victoria locations) are visible. However, similarities of these water bodies (e.g., Albert or Victoria) are visible as well. Eleven out of nineteen of Uganda's population locations significantly differed from the remaining locations.

Finally, the results for cAE are summarized in the visualizations in the Figure 15.

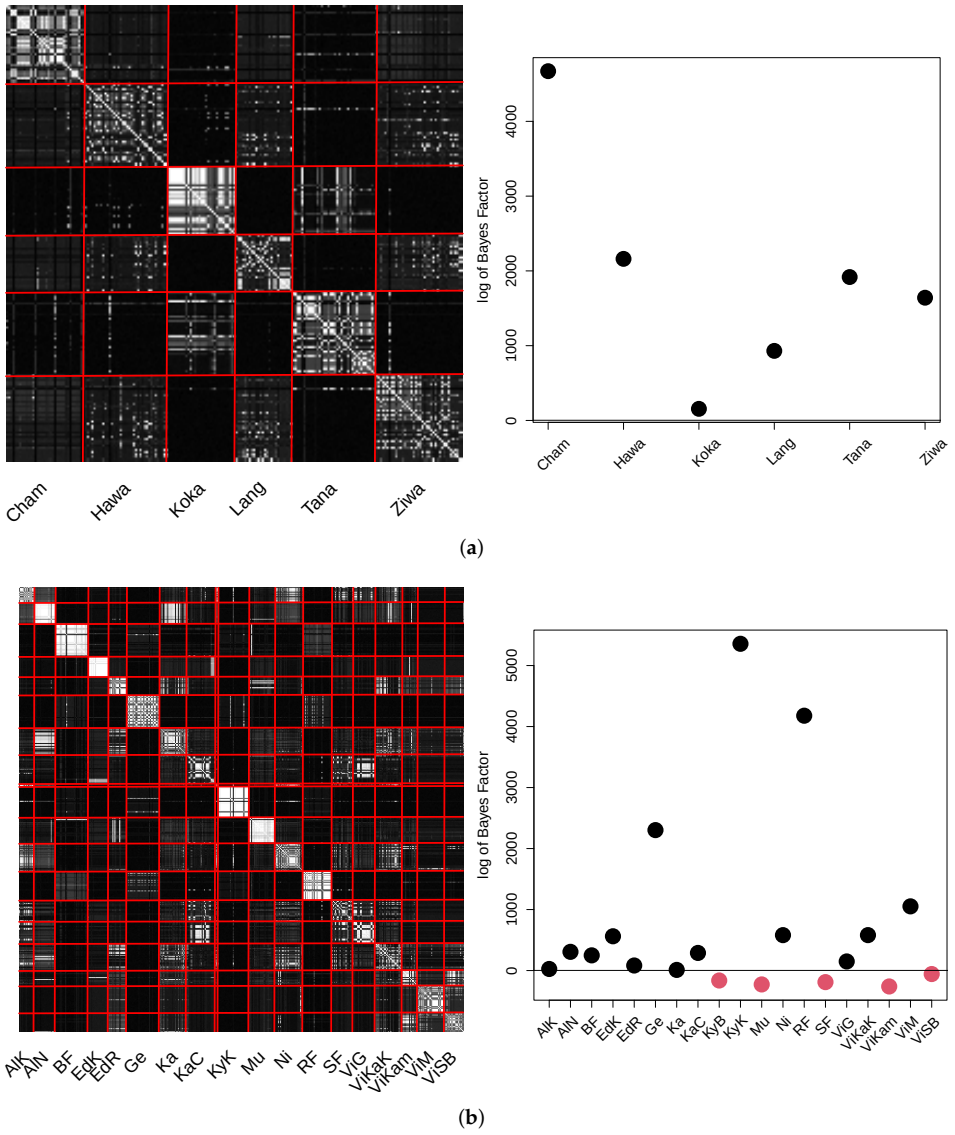


Figure 15. Visualization of the results obtained by pCA and the Bayes factor hypothesis test relying on cAE. Both pCA matrices show visible structure. All locations of Ethiopia were found to be significantly different. Fourteen out of nineteen locations were identified to be significantly different in Uganda's locations. (**a**) pCA and Bayes factor results for Ethiopia relying on cAE. (**b**) pCA and Bayes factor results for Uganda relying on cAE. The KyB population label in the pCA matrix was removed due to readability.

In both pCA matrices, the majority of locations can be clearly distinguished. All locations of Ethiopia were found to be significantly different to the remaining locations. Fourteen of nineteen locations in Uganda differed significantly. However, again the locations of Lake Viktoria did show similarities.

We summarize our findings of the proposed approach in Table 3, where the population's locations (which were observed to be significantly different) are marked with a cross (×).

We observed that for all applied methods the locations Langano and Ziway significantly differed from the remaining populations in Ethiopia. Furthermore, the location Victoria Sango Bay in Uganda was observed to be significantly different, only relying on GPA scaled landmarks. The locations Kyoga Bukungu, Mulehe Musezero as well as the Sindi Farm in Uganda were never observed to be significantly different to the remaining locations.

Table 3. Summary of results obtained with generalized procrustes analysis (GPA), Gaussian process latent variable models (GP-LVM) as well as deep convolutional autoencoder (cAE). The significantly different locations are indicated with a cross (×).

		Abbr.	GPA	GP-LVM	cAE
Ethiopia	Chamo	Cham			×
	Hawassa	Hawa	×		×
	Koka	Koka		×	×
	Langano	Lang	×	×	×
	Tana	Tana		×	×
	Ziway	Ziwa	×	×	×
Uganda	Victoria Kakyanga	ViKak			×
	Victoria Masese	ViM		×	×
	Victoria Gaba	ViG		×	×
	Victoria Sango Bay	ViSB	×		
	Victoria Kamuwunga	ViKam		×	
	Albert Ntoroko	AlN		×	×
	Albert Kyehooro	AlK			×
	George Hamukungu	Ge		×	×
	Kazinga Channel Katungulu	KaC		×	×
	Edward Kazinga	EdK		×	×
	Edward Rwenshama	EdR			×
	Kyoga Kibuye	KyK		×	×
	Kyoga Bukungu	KyB			
	River Nile Kibuye	Ni			×
	Mulehe Musezero	Mu			
	Kayumbu Rugarambiro	Ka		×	×
	Bangena Farm	BF		×	×
	Sindi Farm	SF			
	Rwitabingi Farm	RF		×	×

4. Discussion

This study investigated the quality of extracted GPA scaled coordinates in contrast to machine learning-based features. The quality was measured by the differentiability of the known population locations. To obtain comparable latent structures, the consensus clustering method was extended by a probabilistic interpretation as well as a hypothesis test.

Manually placed landmarks were extracted from image datasets obtained in Ethiopia and Uganda. These landmarks were processed using GPA. The results were statistically investigated. We observed significant relation of the GPA scaled coordinates to the population locations. Furthermore, a significant relation of the locations to the Procrustes distance was obtained. Relying on the visualization of the GPA coordinates (see Figure 6), as well as the Spearman's rank correlation tests, we concluded that the landmark-based approach results in a interpretable reduction of the image data with significant correlation to the populations location. Nevertheless, the GPA-scaled coordinate visualizations and hypotheses tests do not quantify the discriminability of the populations unveiled by the GPA approach. Furthermorer, the multivariate analysis showed overlapping population distribution in the PCA and ICA reduced GPA coordinates.

The proposed pCA was able to unveil significant differences for three out of six locations in Ethiopia. One of nineteen locations was found to be significantly different from the other locations in Uganda. Ref. [24] obtained similar results for supervised Nile tilapia location classification using biologically interpretable GP-LVM features as well as GPA-scaled coordinates. Their results, enhanced by the latent structure investigation of this study show, that classic landmark based-approaches are limited in terms of information discovery. Regardless, the landmark-based approaches outperform the autonomous deep learning counterparts in terms of biological explainability. The machine learning-based approaches were investigated with visualization methods and major focus on biological uninformative image regions was removed.

In contrast to the manually placed landmarks of the GPA, the GP-LVM learns a latent representation of the specimen images. After training, the model reproduces the image content including background information such as specimen mounting material. We manually removed uninformative biological features using the variance-based feature visualization technique. The automation of this procedure is still an open problem in machine learning [69].

The features which were visualized and shown to focus on biological meaningful information were in similar specimen locations to the GPA results. These features were tested using Spearman's rank correlation test. This test indicates a significant relation of the chosen features to the population's locations. Similar to the GPA-scaled landmarks, the multivariate analysis showed overlapping population location distributions. However, the pCA method was able to obtain population structures significantly different to other locations. Four of six locations of Ethiopia were shown to be significantly different. Eleven out of nineteen locations in Uganda were shown to be significantly different. This already shows an improvement compared with previous analyses based on classical morphometrics, where just a few populations were clearly separated [5]. However, the GP-LVM still has limitations. The GP-LVM learns the latent representation of the image datasets using a set of Gaussian processes. On the one hand, the learning procedure is limited in terms of statistical black-box modeling [31]. Furthermore the optimization procedure does not include biological knowledge. The learning procedure could be enhanced using prior biological knowledge in the variational approximation of the model [46].

Similar to the GP-LVM, the cAE learns a latent representation using a learning procedure. Instead of a set of Gaussian processes, the cAE relies on an architecture of artificial neurons including convolutional layers. The latent features were processed in the same manner as GP-LVM features, including manual feature selection focusing on biologically meaningful image regions. All locations in Ethiopia and fourteen of the nineteen locations in Uganda were obtained to be significantly different relying on the pCA. However, the visualization of the selected features show noisy image regions. The reasons for this noisy visualization may be related to the dependent feature space learned by the cAE. The independent variation procedure for visualization may not be applicable for the learned feature space. Furthermore, the limitations of the GP-LVM are the same for the cAE.

We conclude this section by emphasizing the biological explainability of the manually placed landmarks. However, minor population location differences were obtained relying on GPA-scaled coordinates. The machine learning-based methods resulted in major population location differences relying on the pCA. We emphasize that the machine learning models were trained without knowledge of the known population clusters.

Thus, we fail to reject both hypotheses of this study and conclude that the machine learning models can learn biological meaningful features and that these features have a higher relation to the true population clusters than landmark-based features. We conclude that these features do contain information useful for population location discriminability and that the machine learning features exceed the explanation power of the used landmark-based method. Furthermore, our results indicates that larger parts of the specimens (e.g., the head in its entirety or the caudal fin region) are related to population locations. The visualizations of the learned features show that the machine learning models focus on areas with no landmarks. We recommend the investigation of these areas using GPA with additional landmarks. Furthermore, we recommend the GP-LVM for further investigation, including the integration of biological knowledge in the model as well as additional explanation investigation of the deep convolutional autoencoder due to the results of the pCA.

5. Summary

This study unveiled the explanatory power of image processing methodologies for visible diversity investigation. Generalized Procrustes analysis was compared to Gaussian process latent variable models as well as a convolutional autoencoder. Relying on two

image databases, GPA-scaled coordinates as well as GP-LVM- and cAE-based features were extracted. The biological explanatory power of all applied methods was investigated. Furthermore, Spearman's rank correlation test was used to investigate the relation of the obtained features to the population's locations. However, a multivariate analysis of the aforementioned features showed that the population distributions overlaps. In order to overcome this problem and unveil the latent structure available in the image representations, a probabilistic consensus analysis was proposed.

Relying on the pCA, several GMMs were combined and the overall latent structure was visualized using the pCA matrix. Based on the model consensus, a Bayesian hypothesis test was formulated. The machine learning models outperformed the landmark based method. However, restricted explainability limits the biological usage of these models. The GP-LVM resulted in explainable image regions. Regardless, the behavior of the model is not fully explained. On the other hand, the performance of the cAE relies on very noisy image regions. However, the visualization technique used was discussed to be limited for cAE applications and the explanatory factors for the cAE may be still hidden in the model.

We conclude this study by emphasizing the performance of the machine learning models in terms of unsupervised features extraction. We recommend further research to investigate the explanatory methodologies in order to fully unveil the explanatory factors of the models. Furthermore, we recommend including existing biological knowledge in order to convert the black box models to fully explainable statistical tools.

Supplementary Materials: The used data sets as well as the implemented software are available at https://github.com/TW-Robotics/MorphoML (accessed on 17 February 2022).

Author Contributions: Conceptualization, W.W., L.M., M.C. and H.M.; methodology, W.W., L.M., M.C., P.D.T. and H.M.; software, W.W.; validation, W.W., L.M., M.C. and H.M.; formal analysis, W.W.; investigation, W.W., M.C. and H.M.; resources, W.W., P.D.T. and G.T.; data curation, W.W., P.D.T. and G.T.; writing—original draft preparation, W.W. and L.M.; writing—review and editing, W.W., L.M., M.C., P.D.T., G.T. and H.M.; visualization, W.W. and L.M.; supervision, H.M.; project administration, H.M.; funding acquisition, W.W. and H.M. All authors have read and agreed to the published version of the manuscript.

Funding: Open Access Funding by the University of Applied Sciences Technikum Wien.

Institutional Review Board Statement: All specimens used in this study were collected as part of commercial fishing activities performed in collaboration with the National Agricultural Research Organization of Uganda and did not require any permission.

Data Availability Statement: The data are available at https://github.com/TW-Robotics/MorphoML (accessed on 17 February 2022).

Acknowledgments: We thank the UAS Technikum Wien library for funding this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AE	Autoencoder
ARD	Automatic relevance determination
B-GP-LVM	Bayesian Gaussian process latent variable model
CA	Co-association
cAE	Convolutional autoencoder
CNN	Convolutional neuronal network
GMM	Gaussian mixture model
GPA	Generalized Procrustes Analysis
GP-LVM	Gaussian process latent variable model
ICA	Independent component analysis

MSE	Mean squared error
pCA	Probabilistic co-association
PCA	Principal component analysis
RBF	Radial base function

References

1. Thompson, D.W. *On Growth and Form*; Cambridge University Press: London, UK, 1945.
2. Abzhanov, A. The old and new faces of morphology: The legacy of D'Arcy Thompson's 'theory of transformations' and 'laws of growth'. *Development* **2017**, *144*, 4284–4297. [CrossRef] [PubMed]
3. Webster, M.; Sheets, D.H. A Practical Introduction to Landmark-Based Geometric Morphometrics. *Paleontol. Soc. Pap.* **2010**, *16*, 163–188. [CrossRef]
4. Strauss, R.; Bond, C. Taxonomic Methods: Morphology. *Methods Fish Biol.* **1990**, 109–140.
5. Tibihika, P.D.; Waidbacher, H.; Masembe, C.; Curto, M.; Sabatino, S.; Negash, E.; Meulenbroek, P.; Akoll, P.; Meimberg, H. Anthropogenic impacts on the contextual morphological diversification and adaptation of Nile tilapia (*Oreochromis niloticus*, L. 1758) in East Africa. *Environ. Biol. Fishes* **2018**, *101*, 363–381. [CrossRef]
6. Kerschbaumer, M.; Bauer, C.; Herler, J.; Postl, L.; Makasa, L.; Sturmbauer, C. Assessment of traditional versus geometric morphometrics for discriminating populations of the *Tropheus moorii* species complex (Teleostei: Cichlidae), a Lake Tanganyika model for allopatric speciation. *J. Zool. Syst. Evol. Res.* **2008**, *46*, 153–161. [CrossRef]
7. Kerschbaumer, M.; Sturmbauer, C. The Utility of Geometric Morphometrics to Elucidate Pathways of Cichlid Fish Evolution. *Int. J. Evol. Biol.* **2011**, *2011*, 290245. [CrossRef]
8. Rüber, L.; Adams, D. Evolutionary convergence of body shape and trophic morphology in cichlids from Lake Tanganyika. *J. Evol. Biol.* **2001**, *14*, 325 – 332. [CrossRef]
9. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2012; Volume 25.
10. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
11. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778. [CrossRef]
12. Salman, A.; Siddiqui, S.A.; Shafait, F.; Mian, A.; Shortis, M.R.; Khurshid, K.; Ulges, A.; Schwanecke, U. Automatic fish detection in underwater videos by a deep neural network-based hybrid motion learning system. *ICES J. Mar. Sci.* **2020**, *77*, 1295–1307. [CrossRef]
13. Qin, H.; Li, X.; Liang, J.; Peng, Y.; Zhang, C. DeepFish: Accurate Underwater Live Fish Recognition with a Deep Architecture. *Neurocomputing* **2016**, *187*, 49–58. [CrossRef]
14. Villon, S.; Mouillot, D.; Chaumont, M.; Darling, E.S.; Subsol, G.; Claverie, T.; Villeger, S. A Deep Learning Method for Accurate and Fast Identification of Coral Reef Fishes in Underwater Images. *Ecol. Inform.* **2018**, *48*, 238–244. [CrossRef]
15. Cui, S.; Zhou, Y.; Wang, Y.; Zhai, L. Fish Detection Using Deep Learning. *Appl. Comput. Intell. Soft Comput.* **2020**, *2020*, 3738108. [CrossRef]
16. Allken, V.; Handegard, N.O.; Rosen, S.; Schreyeck, T.; Mahiout, T.; Malde, K. Fish species identification using a convolutional neural network trained on synthetic data. *ICES J. Mar. Sci.* **2018**, *76*, 342–349. [CrossRef]
17. Marini, S.; Fanelli, E.; Sbragaglia, V.; Azzurro, E.; del Rio, J.; Aguzzi, J. Tracking Fish Abundance by Underwater Image Recognition. *Sci. Rep.* **2018**, *8*, 13748. [CrossRef]
18. Lapuschkin, S.; Wäldchen, S.; Binder, A.; Montavon, G.; Samek, W.; Müller, K.R. Unmasking Clever Hans predictors and assessing what machines really learn. *Nat. Commun.* **2019**, *10*, 1096. [CrossRef] [PubMed]
19. Samek, W.; Wiegand, T.; Müller, K.R. Explainable Artificial Intelligence: Understanding, Visualizing, and Interpreting Deep Learning Models. *ITU J. ICT Discov.* **2018**, *1*, 49–58.
20. Samek, W.; Montavon, G.; Lapuschkin, S.; Anders, C.J.; Müller, K.R. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proc. IEEE* **2021**, *109*, 247–278. [CrossRef]
21. Samek, W.; Müller, K.R. Towards Explainable Artificial Intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 5–22._1. [CrossRef]
22. Montavon, G.; Samek, W.; Müller, K.R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* **2018**, *73*, 1–15. [CrossRef]
23. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef]
24. Wöber, W.; Curto, M.; Tibihika, P.; Meulenbroek, P.; Alemayehu, E.; Mehnen, L.; Meimberg, H.; Sykacek, P. Identifying geographically differentiated features of Ethopian Nile tilapia (*Oreochromis niloticus*) morphology with machine learning. *PLoS ONE* **2021**, *16*, e0249593. [CrossRef]
25. Gower, J.C. Generalized procrustes analysis. *Psychometrika* **1975**, *40*, 33–51. [CrossRef]
26. Wöber, W.; Mehnen, L.; Sykacek, P.; Meimberg, H. Investigating Explanatory Factors of Machine Learning Models for Plant Classification. *Plants* **2021**, *10*, 2674. [CrossRef] [PubMed]

27. Marcus, G. Deep Learning: A Critical Appraisal. *arXiv* **2018**, arXiv:1801.00631.
28. Tibihika, P.D.; Curto, M.; Negash, E.; Waidbacher, H.; Masembe, C.; Akoll, P.; Meimberg, H. Molecular genetic diversity and differentiation of Nile tilapia (*Oreochromis niloticus*, L. 1758) in East African natural and stocked populations. *BMC Evol. Biol.* **2020**, *20*, 16. [CrossRef] [PubMed]
29. Tesfaye, G.; Curto, M.; Meulenbroek, P.; Englmaier, G.K.; Tibihika, P.D.; Negash, E.; Getahun, A.; Meimberg, H. Genetic diversity of Nile tilapia (*Oreochromis niloticus*) populations in Ethiopia: Insights from nuclear DNA microsatellites and implications for conservation. *BMC Ecol.* **2021**, *21*, 113. [CrossRef]
30. Kariuki, J.; Tibihika, P.D.; Curto, M.; Alemayehu, E.; Winkler, G.; Meimberg, H. Application of microsatellite genotyping by amplicon sequencing for delimitation of African tilapiine species relevant for aquaculture. *Aquaculture* **2021**, *537*, 736501. [CrossRef]
31. Titsias, M.K.; Lawrence, N.D. Bayesian Gaussian Process Latent Variable Model. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Sardinia, Italy, 13–15 May 2010; pp. 844–851.
32. Dong, G.; Liao, G.; Liu, H.; Kuang, G. A Review of the Autoencoder and Its Variants: A Comparative Perspective from Target Recognition in Synthetic-Aperture Radar Images. *IEEE Geosci. Remote Sens. Mag.* **2018**, *6*, 44–68. [CrossRef]
33. Fred, A.L.; Jain, A. Combining Multiple Clusterings Using Evidence Accumulation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 835–50. [CrossRef]
34. Rasmussen, C. The Infinite Gaussian Mixture Model. In *Advances in Neural Information Processing Systems*; Solla, S., Leen, T., Müller, K., Eds.; MIT Press: Cambridge, MA, USA, 2000; Volume 12.
35. Vega-Pons, S.; Ruiz-Shulcloper, J. A Survey of Clustering Ensemble Algorithms. *Int. J. Pattern Recognit. Artif. Intell.* **2011**, *25*, 337–372. [CrossRef]
36. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021.
37. Dunnington, D. rosm: Plot Raster Map Tiles from Open Street Map and Other Sources; R Package Version 0.2.5. 2019. Available online: https://rdrr.io/cran/rosm/ (accessed on 17 February 2022).
38. Bradski, G. The OpenCV Library. *Dr. Dobb's J. Softw. Tools* **2000**, *120*, 122–125.
39. Dryden, I.L. shapes: Statistical Shape Analysis; R Package Version 1.2.6. 2021. Available online: ttps://cran.r-project.org/web/packages/shapes/shapes.pdf (accessed on 17 February 2022).
40. Baken, E.; Collyer, M.; Kaliontzopoulou, A.; Adams, D. geomorph v4.0 and gmShiny: Enhanced analytics and a new graphical interface for a comprehensive morphometric experience. *Methods Ecol. Evol.* **2021**, *12*, 2355–2363. [CrossRef]
41. Adams, D.C.; Otárola-Castillo, E. geomorph: An r package for the collection and analysis of geometric morphometric shape data. *Methods Ecol. Evol.* **2013**, *4*, 393–399. [CrossRef]
42. Collyer, M.L. RRPP: Linear Model Evaluation with Randomized Residuals in a Permutation Procedure. 2019. Available online: https://cran.r-project.org/package=RRPP (accessed on 17 February 2022).
43. Collyer, M.L.; Adams, D.C. RRPP: An r package for fitting linear models to high-dimensional data using residual randomization. *Methods Ecol. Evol.* **2018**, *9*, 1772–1779. [CrossRef]
44. Lawrence, N.D. Gaussian Process Latent Variable Models for Visualisation of High Dimensional Data. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2004; pp. 329–33
45. Li, P.; Chen, S. A Review on Gaussian Process Latent Variable Models. *CAAI Trans. Intell. Technol.* **2016**, *1*, 366–376. [CrossRef]
46. Titsias, M. Variational Learning of Inducing Variables in Sparse Gaussian Processes. In Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, Clearwater Beach, FL, USA, 16–18 April 2009; Volume 5, pp. 567–574.
47. Wöber, W.; Aburaia, M.; Olaverri-Monreal, C. Classification of Streetsigns Using Gaussian Process Latent Variable Models. In Proceedings of the 2019 IEEE International Conference on Connected Vehicles and Expo, ICCVE 2019, Graz, Austria, 4–8 November 2019; pp. 1–6. [CrossRef]
48. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational Inference: A Review for Statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. [CrossRef]
49. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer Science+Business Media: Berlin/Heidelberg, Germany, 2006.
50. GPy. GPy: A Gaussian Process Framework in Python. 2012. Available online: http://github.com/SheffieldML/GPy (accessed on 17 February 2022).
51. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*; The MIT Press: Cambridge, MA, USA, 2005.
52. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
53. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; The MIT Press: Cambridge, MA, USA, 2016.
54. Chollet, F. Keras. 2015. Available online: https://keras.io (accessed on 17 February 2022).
55. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
56. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
57. Chollet, F. Adam. 2015. Available online: https://keras.io/api/optimizers/adam/ (accessed on 17 February 2022).

58. Pett, M.A. *Nonparametric Statistics for Health Care Research: Statistics for Small Samples and Unusual Distributions*; SAGE Publications: Thousand Oaks, CA, USA, 2015.
59. Marchini, J.; Heaton, C.; Ripley, B.D. fastICA: FastICA Algorithms to Perform ICA and Projection Pursuit; R Package Version 1.2-2. 2019. Available online: https://cran.r-project.org/web/packages/fastICA/fastICA.pdf (accessed on 17 February 2022).
60. Hyvärinen, A.; Oja, E. Independent component analysis: Algorithms and applications. *Neural Netw.* **2000**, *13*, 411–430. [CrossRef]
61. Schloerke, B.; Cook, D.; Larmarange, J.; Briatte, F.; Marbach, M.; Thoen, E.; Elberg, A.; Crowley, J. GGally: Extension to 'ggplot2'; R Package Version 2.1.2. 2021. Available online: https://cran.r-project.org/web/packages/GGally/index.html (accessed on 17 February 2022).
62. Raj, A.; Stephens, M.; Pritchard, J.K. fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data sets. *Genetics* **2014**, *197*, 573–589. [CrossRef]
63. Earl, D.A.; vonHoldt, B.M. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **2011**, *4*, 359–361. [CrossRef]
64. Evann, G.; Regnaut, S.; Goudet, J. Detecting the number of clusters of individuals using the software structure: A simulation study. *Mol. Ecol.* **2005**, *14*, 2611–2620. [CrossRef]
65. Scrucca, L.; Fop, M.; Murphy, T.B.; Raftery, A.E. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J.* **2016**, *8*, 205–233. [CrossRef]
66. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
67. Blöschl, G.; Sivapalan, M. Scale issues in hydrological modelling: A review. *Hydrol. Process.* **1995**, *9*, 251–290. [CrossRef]
68. Kass, R.E.; Raftery, A.E. Bayes Factors. *J. Am. Stat. Assoc.* **1995**, *90*, 773–795. [CrossRef]
69. Kauffmann, J.R.; Ruff, L.; Montavon, G.; Müller, K. The Clever Hans Effect in Anomaly Detection. *arXiv* **2020**, arXiv:2006.10609.