

# Uncovering Morphological Landmarks and Dataset Biases with a ViT–SVM Classification Pipeline

1<sup>st</sup> Julian Haring  
Robotics Engineering  
UAS Technikum Wien  
Vienna, Austria  
re23m009@technikum-wien.at

**Abstract**—With machine learning applications the results achieved are often more significant than the reasoning behind them. In many industrial applications accuracies above 95% suffice to justify the implementation into a product. Domains such as the medical field that are reliant on rational arguments to justify their decisions are still excluded from this revolution in our society and the way we work. Transformers, the newest form of artificial intelligence used in products such as ChatGPT, have proven especially hard to interpret in the way they achieve their decisions. Morphology, the field of biology responsible for categorizing animals based on their body shape, faces similar challenges. While their current machine learning implementations provide reliable and highly accurate results, their goal is not mainly correct prediction, but rather to understand the rules behind it. Therefore, in this paper we introduce a novel approach for a classification model that allows a high accuracy in categorizing fish into their respective lake of origin while also giving insight into its reasoning. This is achieved by combining the high accuracy and reliability of vision transformers in the feature extraction step with the more advanced model-wide explainability of the support vector machine in the classification step. The resulting pipeline achieves a classification accuracy of 97.6 %. The heatmaps predominantly agree with the current biological doctrine of relevant body portions responsible for correct classification, indicating the model’s partial understanding of the morphological identity of the fish. Points of interest previously underrepresented in literature are suggested, and deficiencies in the dataset resulting in Clever-Hans effects are discovered. In the end of the paper the results are discussed and an outlook into further development steps is given as well as possible implementations of our model in other industries.

**Index Terms**—morphology, XAI, ViT, SVM

## I. INTRODUCTION

Morphology is the biological discipline of describing organic shapes using visual features [1]. It plays a crucial role in objectively classifying different samples of animals and plants between genera [2], subspecies [3] and individual entities [4]. In the last decades, classic morphology has been performed by taking measurements or manually placing predefined landmarks on each individual image, then do statistical analysis. While this offers robust results, the effort needed especially for larger datasets and the limitation to known landmarks allows for improvements [1]. With numerous scientific breakthroughs in the last decade as well as advances in the processing power of CPUs and GPUs, data based machine learning (ML) is increasingly relevant to improve morphological classification tasks [5]. In comparison to landmark-based ML the amount of human interaction for classification tasks is reduced drastically. Data driven ML approaches in morphology are being used in different projects, such as the differentiation of cow breeds [6], the classification of insects [7] or the taxonomic sorting of primates using their mandible shape [8].

Although these methods often achieve classification accuracies above 90%, their explainability tends to decrease as model complexity increases [9]. As a result, ML models may differentiate classes based on artifacts from data acquisition rather than biologically meaningful traits [10]. Gaining insight into the model’s decision-making process enables the identification of such issues and ensures that the selected landmarks are morphologically relevant [11].

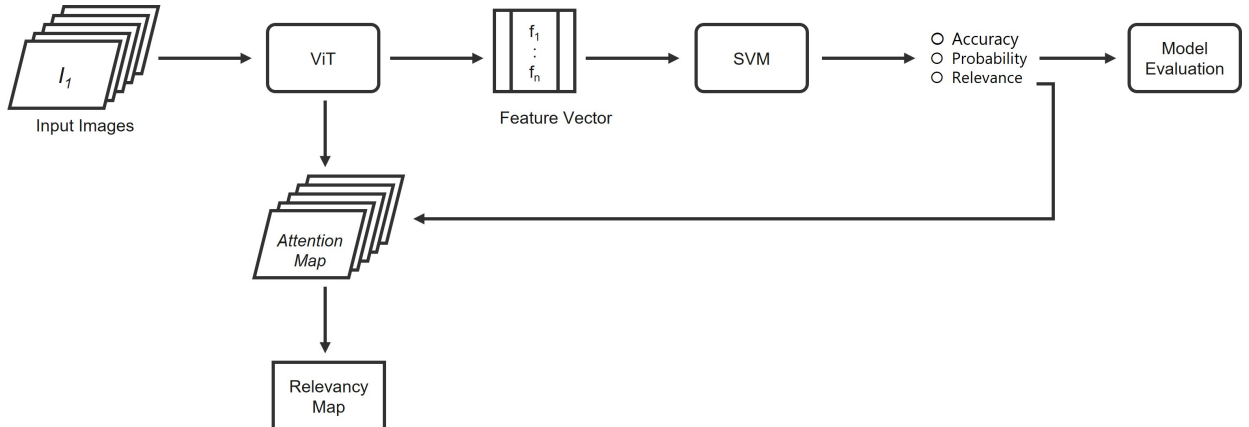


Fig. 1: ViT–SVM pipeline for classification and relevancy mapping.

To address these challenges, we propose a novel classification pipeline (Fig. 1). The dataset is first divided into an 80% training and 20% holdout set for unbiased evaluation. Features and attention maps are then extracted separately from both subsets using a pretrained Dino ViT-B/14 [12] distilled with registers. The training features are subjected to five-fold cross-validation using a Support Vector Machine (SVM) [13] as the classifier. In each fold, 80% of the training data are used to train the SVM and 20% for validation, with the roles of the samples rotating across folds. This procedure ensures that every training sample contributes to both model fitting and validation, providing a robust estimate of the SVM's performance. To calculate the relevancy maps, the mean attention map of all 12 attention heads is aggregated. The dot product of this average attention map with the learned SVM weights yields patch-wise relevancy scores. Those scores are transformed into heatmaps that highlight the image regions most influential for classification. This integration provides both strong predictive performance and interpretability by revealing morphological landmarks and potential dataset-specific biases that drive the model's decisions. In the end SVM weights are retrained on the entire training set and its performance is assessed on the holdout set to obtain an unbiased evaluation.

We hypothesize that applying this pipeline will not only improve the identification of relevant morphological landmarks for classification but also reveal dataset-specific biases present in the images.

The dataset [14] used for both the training and evaluation of the model comprises 209 images of Ethiopian Nile tilapia (*Oreochromis niloticus*) [15]. This species was selected due to its remarkable adaptability to varying environmental conditions, which has led to several mutations, including a reduction in body circumference in response to increased water salinity [15]. These genetic variations present a unique opportunity for machine learning models to identify such anomalies and classify the fish according to their respective lakes of origin.

In the following chapters we first go over the current state of the art of explainable AI and morphological classification in general. Afterwards, the exact functionality of the model is explained followed by an analysis on the technical results of the pipeline. Next, a more thorough evaluation of the biological outcome is presented. Finally, technical as well as biological findings are contextualized, and thoughts on future improvements of the pipeline are given.

## II. STATE OF THE ART

More rudimentary approaches to morphological classification using ML are employed by Duraiswami et al. [6]. While the pipeline consisting of a feature extractor and an SVM is similar to ours, the features relevant for classification are manually selected morphological traits (e.g. horns, ears, and coat patterns). This constrains the ability to investigate the current biological understanding of morphologically relevant body parts. Similarly Wöber et al. [14] aim to classify fish by their lake of origin using manually placed landmarks according

to known points of high variety. This is then compared to automatically placed landmarks using a convolutional autoencoder [16] and a Gaussian Process Latent Variable Model [17]. The authors conclude that while automatically placed landmarks achieve 10% higher accuracy than manual ones, the approach still lacks in explainability and recommends further research on the topic.

Moving beyond manually defined features, more recent studies employ deep learning models that automatically learn discriminative representations from raw data. Tsutsuami et al. [8] use a Variational Autoencoder [18] to classify different primate subspecies based on three-dimensional scans of their mandibles. While this model achieves high accuracy, the interpretability remains limited.

Because such deep learning pipelines prioritize predictive performance over transparency, subsequent research has focused on post-hoc methods to render these models more interpretable. Çifci and Kırbaş [19] compare ten different ML classifiers and analyze their reasoning using the universally applicable SHapley Additive exPlanation (SHAP) [20] and Individual Conditional Expectation (ICE) [21]. They conclude that the Logistic Regression offers the highest accuracy and are able to determine the most influential areas for classification.

Explainability is particularly crucial in fields such as radiology, where model transparency fosters trust and ethical standards. Hussain et al. [22] propose several pipelines for evaluating convolutional neural networks (CNN) [23]. By analysing GradCAM [24], LIME [25], t-SNE [26] and UMAP [27] the authors conclude that these methods can achieve a higher understanding of the relevant features, but agree that further research is necessary.

However methods such as SHAP, LIME or GradCAM are inherently limited, as they provide explanation only after the classification step rather than rendering the models intrinsically transparent. Therefore, Sun et al. [28] introduce the AS-XAI (Self-Supervised Automatic Semantic Interpretation) framework. It uses transparent embedding spaces and PCA to automatically extract robust, global semantic concepts from CNNs without needing human-labeled concepts. This method generates human-comprehensible explanations with no additional computational cost, directly addressing the reliability and usability issues of previous XAI techniques.

While these techniques advance explainability in CNN-based architectures, the emergence of Vision Transformers (ViTs) has opened new possibilities for combining accuracy with interpretability. Mzoughi et al. [29] directly compare CNNs with ViTs on the classification of brain tumors. The ViT achieves superior classification accuracy with 91.61% compared to the CNNs 83.37%. The reliability of the models is evaluated using GradCAM, LIME and SHAP. Building on this development, Zheng et al. [30] enhance ViT-based pipelines by introducing the Trilinear Attention Sampling Network. This framework employs a teacher-student model and dynamically adjusts image proportions according to attention scores, thereby improving both classification accuracy and model interpretability.

Despite these advancements, important gaps remain. Early approaches based on handcrafted features are limited in their biological interpretability and difficult to scale. Deep learning methods, while achieving high accuracies on complex datasets, struggle with transparency and typically depend on post-hoc explanations, which constrain their applicability in biological contexts. More recent attempts such as AS-XAI and TASN offer high accuracy with more transparent and biologically meaningful explanations, but come at the cost of substantial computational demands and architectural complexity. Thus, there is a need for a pipeline combining high predictive accuracy with strong explainability. Our proposed approach addresses this gap by combining the ability of ViTs to capture both fine-grained details and global structures with the relatively low computational cost and decision transparency of SVMs.

### III. METHODS

To overcome the tradeoff of morphological ML models either achieving high accuracy or explainability, we introduce a novel approach (Figure 1). For detecting landmarks a ViT is adopted. Introduced by Dosovitskiy et al. [31] in 2020, ViTs offer a high level of accuracy when compared to the former state of the art CNNs [32]. This is partly due to the image being split up into non-overlapping patches, each encoded with positional embeddings, allowing the model to retain spatial relationships within the data.

In our framework we employ a pre-trained DINOv2 B/14 ViT with registers [12] using pyTorch 2.5, reducing the reliance on large scale annotated training data. Initially, the dataset is partitioned into a train and validation set consisting of 80% and 20% respectively. From both subsets, we extract two feature matrices  $F$

$$\mathbf{F} = [f_1, f_2, \dots, f_N] \in \mathbb{R}^{N \times d},$$

consisting of the feature vectors  $f_j$  for each individual image. In addition, we extract the attention maps associated with the classification token (CLS) from the last transformer block of the model using a forward hook. The training features are used to train a SVM classifier (SciKit Learn 1.5.1) with 5-fold cross-validation, while performance is assessed on the validation set. The learned weight vector  $\mathbf{w}$  of the SVM enables the assessment of the relative importance of each feature for the classification result [33]. The SVM is trained with a regularization strength of 1.0, an optimization tolerance of  $1e-4$  and assumes a balanced set.

Nonetheless, the direct interpretation of the SVM weight vector  $\mathbf{w}$  remains opaque, as individual weights cannot be connected to specific visual traits. To overcome this limitation, the dot product of the class-specific weight vector  $\mathbf{w}_c$  with the corresponding feature vector  $f_i$  is calculated (1), resulting in patch-wise importance scores  $p_i$ .

$$p_i = \mathbf{w}_c^\top f_i, \quad (1)$$

These values are then combined with the median attention across all twelve heads for each patch, where  $A$  denotes the mean attention weight (2).

$$A = \frac{1}{H} \sum_{h=1}^H a_h \quad (2)$$

and  $a_h$  the attention score for head  $h$ . The final patch relevancy score is then computed as the product of the patch importance and its attention weight (3).

$$R_i = p_i \cdot A_i \quad (3)$$

This yields the relevancy map  $\mathbf{R} = [R_1, \dots, R_N]$ , which integrates both the classifier's discriminative features and the ViT's attention distribution, thereby highlighting patches most relevant to the prediction while reducing spurious activations.

For further analysis, the resulting relevancy map is overlaid on the corresponding input image using matplotlib 3.8.4 and interpolated to provide a visual representation of the contributing biological traits. To aggregate across samples, class-level median relevancy maps  $\tilde{R}_c$  are computed by aggregating over all images  $j$  in class-specific subsets  $\mathcal{D}_c$  (4).

$$\tilde{R}_c = \text{median}\{R^{(j)} \mid j \in \mathcal{D}_c\} \quad (4)$$

To examine how the model differentiates between classes, we implement a subtractive relevancy pipeline [33]. For two classes  $c_1$  and  $c_2$ , the differential relevancy map is defined as (5):

$$\Delta R_{c_1, c_2} = \tilde{R}_{c_1} - \tilde{R}_{c_2} \quad (5)$$

More generally, the unique discriminative relevance of class  $c$  is obtained by subtracting the average relevancy of all other classes  $c' \in \mathcal{C} \setminus \{c\}$  (6):

$$\Delta R_c = \tilde{R}_c - \frac{1}{|\mathcal{C}| - 1} \sum_{c' \in \mathcal{C} \setminus \{c\}} \tilde{R}_{c'} \quad (6)$$

This differential analysis reveals image regions that are specifically discriminative between classes. Additionally, subtracting the median relevancy of all other classes isolates the features most unique to the class of interest.

### IV. TECHNICAL DISCUSSION

When discussing the technical performance of the pipeline, it is important to note the small dataset size of only 209 images across six classes. This limitation reduces the reliability of reported evaluation metrics, as they may not fully capture the systems behaviors. To mitigate this limitation, evaluation was carried out in two stages: first, using accuracies averaged over k-fold cross-validation and second, by training the SVM on the complete training set and validating on an independent 20% hold-out set that had not been used during training. This approach provides a more robust measure of generalization. In 5-fold cross validation the SVM achieved and overall average accuracy of 96% (Figure 2). All classes reached precision,

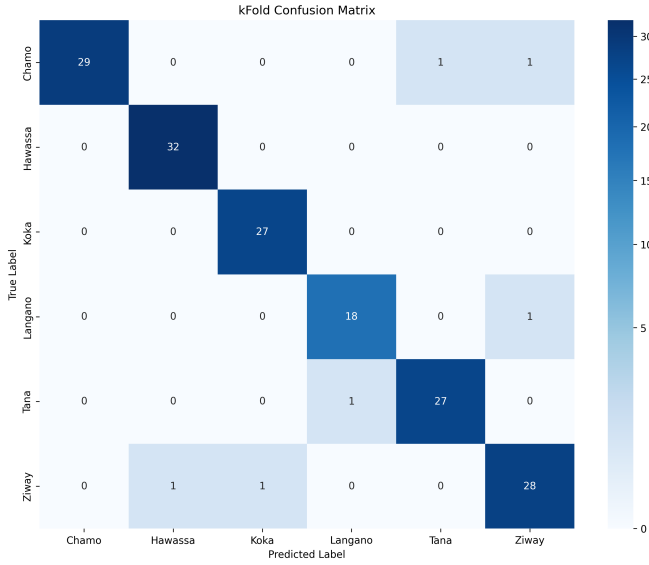


Fig. 2: Combined confusion matrix of all folds

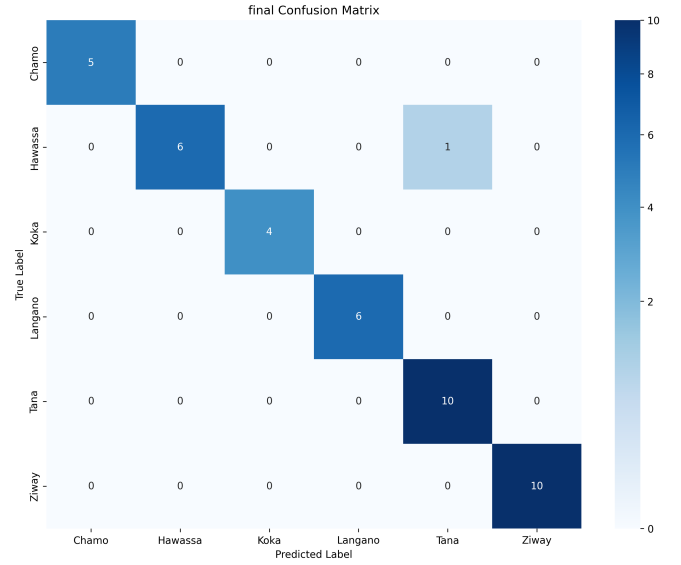


Fig. 3: Confusion matrix of holdover dataset

recall and F1-scores of at least 93%, suggesting that no major misclassification is present. The majority of misclassifications included fishes from lake Ziway, with a total of four out of six errors. On the hold-out validation set the pipeline achieved an accuracy of 97.6% (Figure 3), which indicates a robustness of the system against overfitting. While Ziway achieved perfect precision in this set, the low amount of samples ( $n=10$ ) makes this metric unreliable. In summary, both cross-validation and hold-out validation confirm that the pipeline achieves consistently high accuracy. Although some classes appear more prone to confusion, no clear outlier emerges. However, due to the limited dataset size, deeper conclusions about inter-class difficulties cannot be drawn.

## V. BIOLOGICAL DISCUSSION

For evaluation of the pipeline and the morphological strategies learned by the model, it is necessary to identify meaningful regions of the fish that are relevant in traditional ML as well as ML approaches. Wöber et al. [34] applied a set of 14 landmarks on the body of a Nile tilapia originally derived from Ndiwa et al [15]. For comparability the landmark definition as well as nomenclature of Wöber et al. are retained (Figure 4), (Table I).

A simple method for analyzing the strategy employed by a Vision Transformer is via the attention applied to each individual image patch by the cls patch of each head. Since Dino V2 is trained unsupervised, the classification heads are not trained for inter-class discrimination. Nevertheless, attention maps provide insight in which image regions later become features, allowing the comparison to regions traditionally used for morphological differentiation. Analysis of the median values of each head allow us to pair each head to a region its mainly focused on. The exact type of information extracted by the ViT remains uncertain, as it is unclear whether the model

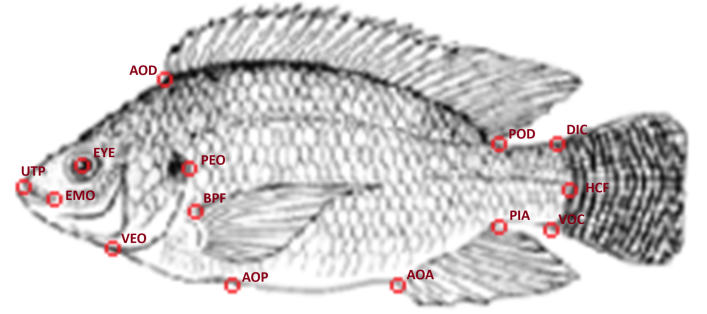


Fig. 4: Landmarks placed by Wöber et al. on the Nile Tilapia indicating locations of great morphological variance. Abbreviations according to Table I

Code	Description
UTP	Upper tip of snout
EYE	Center of eye
AOD	Anterior insertion of dorsal fin
POD	Posterior insertion of dorsal fin
DIC	Dorsal insertion of caudal fin
VOC	Ventral insertion of caudal fin
PIA	Posterior insertion of anal fin
BPF	Dorsal base of pectoral fin
PEO	Posterior edge of operculum
VEO	Ventral edge of operculum
AOA	Anterior insertion of anal fin
AOP	Anterior insertion of pelvic fin
HCF	Midpoint between dorsal and ventral insertions of caudal fin
EMO	Posterior end of mouth

TABLE I: Morphological landmarks and their abbreviations.

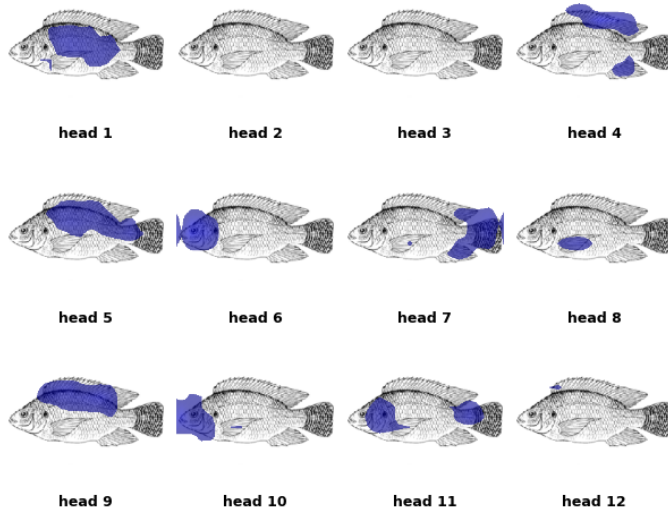


Fig. 5: Mean attention applied from each head to the images.

Head No.	regions of focus
1	Lateral belly, scales/texture
2	mounting wires
3	mounting wires
4	Dorsal, anal and pelvic fin
5	Dorsal scales/texture
6	Head with operculum
7	Caudal and anal fin
8	Pectoral fin / sometimes operculum
9	Dorsal scale and fin
10	Head without operculum
11	Operculum and insertion of caudal fin
12	Distal end of dorsal fin, often empty

TABLE II: Areas focused on by different attention heads.

primarily encodes shape, texture, global structures, or local features.

The majority of heads focus on regions previously noted as morphologically relevant (Figure 5). Notably, the extractor applies substantial focus on the dorsal surface (3 out of 12 heads), where Wöber et al [34] only placed 2 out of 14 landmarks in the same region. Conversely, head two and three appear not to be used in the median summary. When analyzing individual attention maps head No. 2 as well as head No. 3 only apply a high level of attention to the image if a mounting wire is visible right below the snout (Figure 6). These artifacts reflect potential spurious correlations and Clever Hans effects, which may reduce generalization.

To increase the information of classifying strategies deployed, relevancy maps, combining the attention maps from the feature extractor with the weights used by the SVM are created. Compared to the previous attention maps, relevancy maps offer a more direct insight into regions used for clas-

sification. The majority of high level regions overlap with landmark based landmarks used in previous works such as Ndiwa et al [15].

When aggregating relevancy maps (Figure 7), several deviations from traditional approaches become evident. Among the previously used landmarks only EYE and EMO consistently prove to be of high relevance for classification. For the pectoral fin, the ventral surface appears to be more discriminative than the dorsal base BPF. The dorsal fin shape and edge structure, largely overlooked in studies, emerge as highly relevant features. The region directly beneath UTP shows high relevance as well, especially in the classification of Langanu and Ziway. The area matches with the position of the fixation wire and the attention applied by heads 2 and 3.

Applying subtractive relevancy analysis the suspicion is further increased. The differentiation of Langanu and Ziway to Hawassa and Koka appear to be mainly focused on the existence of the wire (Figure 8) confirming the suspicion raised by attention head 2 and 3 of a clever Hans phenomena. Nonetheless compared to other classes the SVM found valid strategies mainly focusing on the head and dorsal body for Langanu and the operculum and dorsal body for Ziway. Biologically feasible strategies can be noted for Hawassa and Koka with the former focusing on the ventral surface of the pectoral fin and the later on the dorsal fin and dorsal scales. No apparent tactic is apparent for Tana. The classifier mainly focuses on seemingly random spots in the background which may either imply the existence of undetectable background correlations or a lack of clear morphological discriminators.

As hypothesized, the pipeline demonstrates a high degree of interpretability, uncovering strategies that align with biologically meaningful features while also confirming parts of the established morphological knowledge base. At the same time, it exposes cases of Clever Hans effects, where the model exploits spurious background or fixation artifacts rather than intrinsic morphological traits. This dual outcome underscores both the potential and the limitations of unsupervised ViTs for biological image analysis: while they capture valid taxonomic features, they remain vulnerable to dataset-specific biases that could mislead classification on unseen material.

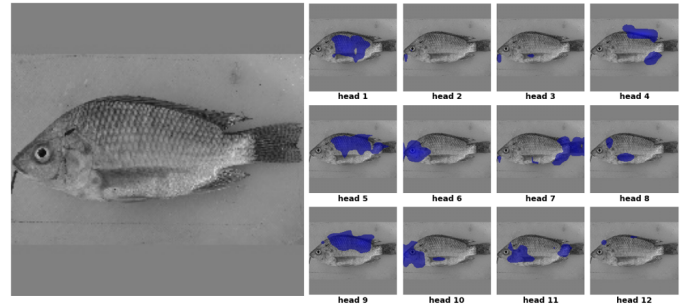


Fig. 6: Attention applied to Ziway image. Head No. 2 and 3 focus on fixation wire.



Fig. 7: Median relevancy of all six classes (Top 10%)

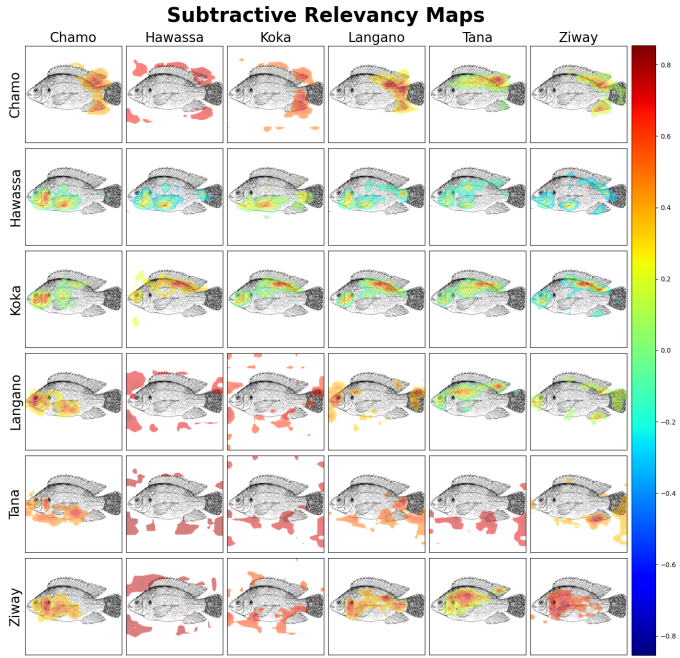


Fig. 8: Subtractive relevancy analysis. Each image shows the difference of classification of the entries on the x axis compared to y. The diagonal values display the difference to all other classes. Only the top 10% are visible.

## VI. CONCLUSIO

The proposed pipeline combining a Dino V2 feature extractor as well as SVM for classification, demonstrates a good compromise of achieving high classification accuracy while making the model interpretable. Despite the small set of 209 images across six classes the model reached an accuracy of 97,6%. Importantly, the combination of ViT attention scores and the weights accumulated by the SVM yielded a relevancy

map granting unusually detailed insight into the regions most relevant for classification, extending beyond the typical “black box” behavior of transformer-based systems.

From a biological perspective, the model not only confirmed the importance of landmarks used in traditional morphology, but enhanced the current understanding by suggesting an increased focus on the dorsal and pectoral fin. At the same time the model discovered Clever Hans effects, where spurious artifacts, such as fixation wires the the anterior end or background artifacts lead to wrongfully accurate classifications. This dual outcome underscores the potential of the method both to validate and refine established morphological knowledge while also exposing the vulnerabilities of data-driven models.

Looking forward, future work should be invested on applying the pipeline to larger, more diverse datasets, enabling broader assessment of morphological strategies while reducing dataset specific biases. A valuable improvement could be the implementation of more sophisticated aggregation methods of model wide behavior. Beyond morphology, the pipeline offers promising results for other industries that may benefit from the implementation of ML but require transparent decision making such as medical diagnostics.

## REFERENCES

- [1] M. Webster and H. D. Sheets, “A practical introduction to landmark-based geometric morphometrics,” pp. 163–188, 2010.
- [2] L. A. Courtenay, J. Yravedra, R. Hugueta, J. Aramendi, M. Ángel Maté-González, D. González-Aguilera, and M. C. Arriaza, “Combining machine learning algorithms and geometric morphometrics: A study of carnivore tooth marks,” *Palaeogeography, Palaeoclimatology, Palaeoecology*, vol. 522, pp. 28–39, 2019.
- [3] R. Meier, J. Smith, and E. Jackson, “Distinct subspecies or phenotypic plasticity? genetic and morphological differentiation of mountain honey bees in east africa,” *Ecology and Evolution*, vol. 7, no. 5, pp. 1119–1131, 2017.
- [4] L. Karczmarski, S. C. Y. Chan, D. I. Rubenstein, S. Y. S. Chui, and E. Z. Cameron, “Individual identification and photographic techniques in mammalian ecological and behavioural research—part 1: Methods and concepts,” *Mammalian Biology*, vol. 102, no. 3, pp. 545–549, June 2022.
- [5] Y. He, J. M. Mulqueeney, E. C. Watt, A. Salili-James, N. S. Barber, M. Camaiti, E. S. E. Hunt, O. Kippax-Chui, A. Knapp, A. Lanzetti, G. Rangel-de Lázaro, J. K. McMinn, J. Minus, A. V. Mohan, L. E. Roberts, D. Adhami, E. Grisan, Q. Gu, V. Herridge, S. T. S. Poon, T. West, and A. Goswami, “Opportunities and challenges in applying ai to evolutionary morphology,” *Integrative Organismal Biology*, vol. 6, no. 1, p. obae036, 09 2024.
- [6] N. R. Duraiswami, S. Bhalerao, A. Watni, and C. N. Aher, “Cattle breed detection and categorization using image processing and machine learning,” in *2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*, 2022, pp. 1–6.
- [7] L. Karczmarski, S. C. Y. Chan, D. I. Rubenstein, S. Y. S. Chui, and E. Z. Cameron, “Classification of aedes mosquito larva using convolutional neural networks and extreme learning machine,” in *Proceedings of the 2023 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 2023, pp. 1–5.
- [8] M. Tsutsumi, N. Saito, D. Koyabu, and C. Furusawa, “A deep learning approach for morphological feature extraction based on variational auto-encoder: an application to mandible shape,” *npj Systems Biology and Applications*, vol. 9, no. 1, p. 30, 2023.
- [9] V. Kamakshi and N. C. Krishnan, “Explainable image classification: The journey so far and the road ahead,” *Free Online Library*, 2023.
- [10] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, “Unmasking clever hans predictors and assessing what machines really learn,” *Nature Communications*, vol. 10, no. 1, p. 1096, 2019.

- [11] E. Akyol and C. Öztürk, "Explainable artificial intelligence for differentiating honey bee genotypes using morphometrics and ssr markers," *Apidologie*, 2024.
- [12] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2024.
- [13] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [14] W. Wöber, L. Mehnen, M. Curto, P. D. Tibihika, G. Tesfaye, and H. Meimberg, "Investigating shape variation using generalized procrustes analysis and machine learning," *Applied Sciences*, vol. 12, no. 6, p. 3158, 2022.
- [15] T. C. Ndiwa, D. W. Nyngi, J. Claude, and J.-F. Agnèse, "Morphological variations of wild populations of nile tilapia (*oreochromis niloticus*) living in extreme environmental conditions in the kenyan rift-valley," *Environmental Biology of Fishes*, vol. 99, pp. 473–485, 2016.
- [16] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Proceedings of the Artificial Neural Networks and Machine Learning – ICANN 2011*. Springer, 2011, pp. 52–59.
- [17] N. D. Lawrence, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," *Journal of Machine Learning Research*, vol. 6, pp. 1783–1816, 2005.
- [18] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [19] A. Çifci and I. Kirbas, "Fusion of machine learning and explainable ai for enhanced rice classification: A case study on cameo and osmancik species," *European Food Research and Technology*, November 2024, advance online publication.
- [20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [21] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: A simple method for ICE plots," *Proceedings of the 2015 International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 859–867, 2015.
- [22] S. M. Hussain, D. Buongiorno, N. Altini, F. Berloco, B. Prencipe, M. Moschetta, V. Bevilacqua, and A. Brunetti, "Shape-based breast lesion classification using digital tomosynthesis images: The role of explainable artificial intelligence," *Applied Sciences*, vol. 12, no. 12, p. 6230, 2022.
- [23] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.
- [24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2017, pp. 618–626.
- [25] M. T. Ribeiro, S. Singh, and C. Guestrin, "'why should i trust you?': Explaining the predictions of any classifier," 2016.
- [26] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.
- [27] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," in *Proceedings of the 2018 31st International Conference on Neural Information Processing Systems (NeurIPS) Workshop on Algorithmic Foundations of Data Science*, 2018.
- [28] C. Sun, H. Xu, Y. Chen, and D. Zhang, "As-xai: Self-supervised automatic semantic interpretation for cnn," *arXiv preprint arXiv:2312.14935*, 2023.
- [29] H. Mzoughi, I. Njeh, M. BenSlima, N. Farhat, and C. Mhiri, "Vision transformers (vit) and deep convolutional neural network (d-cnn)-based models for mri brain primary tumors images multi-classification supported by explainable artificial intelligence (xai)," *The Visual Computer*, June 2024.
- [30] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," pp. 5007–5016, 2019.
- [31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020.
- [32] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *CoRR*, vol. abs/2108.08810, 2021.
- [33] B. Gaonkar, R. T. Shinohara, and C. Davatzikos, "Interpreting support vector machine models for multivariate group-wise analysis in neuroimaging," *Medical Image Analysis*, vol. 24, no. 1, pp. 190–204, 2015.
- [34] W. Wöber, M. Curto, P. D. Tibihika, P. Meulenbroek, E. Alemayehu, L. Mehnen, and H. Meimberg, "Identifying geographically differentiated features of ethiopian nile tilapia (*oreochromis niloticus*) morphology with machine learning," *PLOS ONE*, vol. 16, no. 4, p. e0249593, 2021.