Zachary Hom

Shark Paleobiology

Dr. Jeffrey Agnew

December 13, 2025

A comparison of linear and non-linear dimension reduction techniques to identify taxonomic groups among modern shark teeth

Abstract

Elliptical Fourier analysis (EFA) and principal component analysis (PCA) are widely used statistical tools to identify and distinguish among shark taxa based on a set of indeterminate tooth images. EFA processes these images and outputs four Fourier descriptor coefficients for each ellipse contour that quantitatively describe the shape, size, and orientation of the object, which in our case is a single shark tooth sample. PCA is a linear dimension reduction technique that collapses multiple attribute channels to one, thereby allowing direct comparison and grouping across high-dimensional data. However, little investigation has been done to compare the performance between linear and non-linear dimension reduction algorithms for taxonomic classification, which is important to identify ancient shark species and better understand shark phylogeny. This study compares the clustering performance between PCA and t-Distributed Stochastic Neighbor Embedding (t-SNE), a non-linear dimension reduction model. From my results, t-SNE consistently performed better than PCA at clustering unlabeled images, highlighting its potential use for taxonomic classification. t-SNE shows promise as an unsupervised learning tool for grouping teeth from unknown or ancient shark taxa. However, t-SNE data lacks morphological significance, so researchers in the future should select which dimension reduction method best suits their intended purpose.

Introduction

Elliptical Fourier analysis (EFA) is a morphometric analysis tool that uses elliptical harmonic contours to describe the morphology of shark teeth quantitatively. EFA has largely replaced manual morphometric measurements and has become a standard within the field of taxonomy, botany, zoology, and microbiology (Marramà et al., 2017). This method outputs four descriptor coefficients per contour ellipse that describe the shape, size, and orientation of the tooth. To identify similarities and classify observations across multiple descriptor coefficients, a statistical method called dimension reduction can be employed. Principal component analysis (PCA) is the most widely used technique to investigate biological patterns and has shown success with taxa separation while avoiding over-splitting (Marramà et al., 2017). Another dimension reduction method is t-distributed stochastic neighbor embedding, or t-SNE. The major difference between these two approaches is how they manipulate and maintain the structure of a given dataset. PCA uses linear Euclidean distance to maintain the overall structure of the dataset post-reduction, while t-SNE employs non-linear methods and probabilistic distance to prioritize clustering similar observations (neighborhoods) instead of preserving global variance. Thus, non-linear dimension reduction techniques may identify groupings that linear reduction methods smooth out. t-SNE is effective at maintaining local relationships (clusters) among groups of data points, while PCA is more effective at maintaining global relationships across the dataset. For this reason, t-SNE may be equally, if not more, effective at grouping shark teeth based on variance across their Fourier descriptors. t-SNE has applications to identify ancient shark species and better understand shark phylogeny from unknown sets of teeth. I aim to test my hypothesis with three sets of teeth from modern sharks of the same genus

in order to quantify the performance of each reduction technique at grouping unlabeled teeth into distinct, known taxa.

Background

Early sharks first appeared in the fossil record during the Devonian Period, approximately 400 million years ago. Because sharks rapidly shed their teeth throughout their lifetime, their teeth are well-documented in the fossil record (Purdy, 1995). However, shark teeth morphology is extremely diverse, with examples of ontogenic, sexual, and dignathic heterodonts observed within modern species (Purdy, 1995). This fact makes species identification using tooth fossils alone difficult among modern shark species and even more so for ancient sharks. Today, paleontologists reconstruct sets of teeth by comparing them with modern species, eliminating species duplicates. This approach has revealed insights into shark speciation and evolutionary history since the Devonian and has increased our understanding of evolution through time of one of the oldest surviving species on Earth. Although these new approaches have proved useful to identify shark taxa, numerical methods now exist to quantify their shape. These methods can help modern taxonomists and paleontologists to cluster groups of similar teeth that can be used to identify different species among sets of fossils with unknown origin, deepening our understanding of shark evolutionary history.

Before elliptical Fourier analysis, taxonomists used morphometric measurements to distinguish among the teeth of shark species. A recent study has compared these shape descriptor methods with PCA as the unifying dimension reduction technique. The study concluded that EFA was the more objective morphometric method as it removed interobserver bias and was much less labor-intensive than manual measurement (Goodman, 2022). By increasing the number of harmonics, researchers can achieve a higher proportion of a shark

tooth's shape described by Fourier coefficients. Even as few as seven harmonics can describe 99.9% of the variation in shark tooth shape (Cullen & Marshall, 2019).

Although there seems to be a consensus on morphological assessment, there does not seem to be a unifying dimension reduction technique to process EFA morphometric data. However, some comparisons between dimension reduction techniques have been conducted in the past. A study comparing discriminant analysis (DA) and PCA for taxonomic identification and phylogenetic signal detection was conducted, highlighting strengths and weaknesses for both linear reduction methods (Marramà et al., 2017). The study concluded that both dimension reduction techniques were useful to distinguish between different taxa given a set of shark tooth images, with DA being particularly useful to classify indeterminate teeth into known taxa. However, few studies have compared the performance between linear and non-linear dimension reduction techniques to identify taxa from unlabeled sets of shark teeth.

Methods

I began my investigation by downloading a dataset of 117 high-resolution images of *Carcharhinus leucas* (Bull shark), *Carcharhinus amboinensis* (Java shark), and *Carcharhinus obscurus* (Dusky shark) teeth obtained from Dr. Jeffrey Agnew's Box repository (38 Bull, 35 Java, and 44 Dusky shark teeth). These taxa were selected to provide an appropriate test case to evaluate cluster performance across different dimension reduction techniques. Since my goal is to compare model-derived groupings against known biological taxa, it is essential to use teeth from modern species with identified origins. *C. leucas*, *C. amboinensis, and C. obscurus* provide this ideal comparison; they are genetically and morphologically similar enough that the distinction between species can be ambiguous to the human eye, yet they retain slight structural differences that statistical methods like elliptical Fourier analysis can detect.

A key component of elliptical Fourier analysis is image pre-processing to map the color scale from RGB to binary grayscale (background to black and tooth to white). This supports effective contouring during EFA by preserving the tooth's silhouette while minimizing graphical noise. By leveraging OpenCV, an open-source computer vision Python package, we can set a binary color threshold that maps any RGB values below it to the minimum value in the colorspace (black) and any value above it to the maximum (white), as seen in Figure 1.

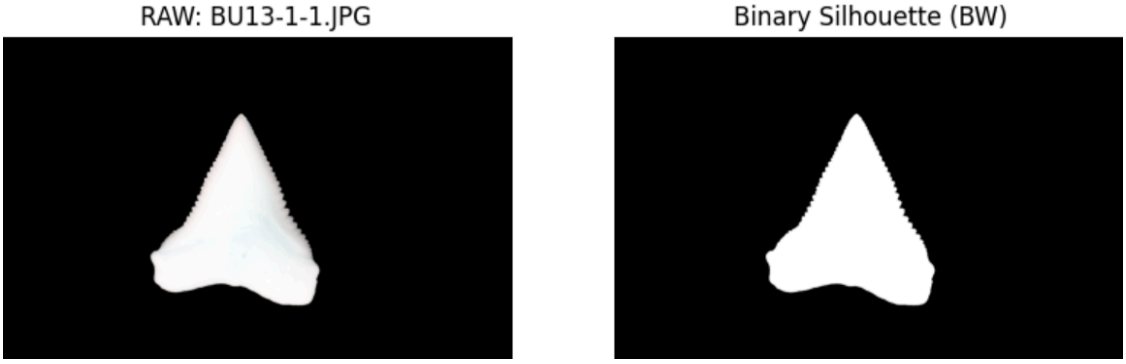


Figure 1: Raw, high-resolution Bull shark tooth image (left) and the digitized version produced by OpenCV threshold image processing (right). Graphical noise on the tooth face (shading, indentation, nested structure) is removed to render a binary silhouette suited for elliptical Fourier analysis.

I chose a threshold of 127, which is the RGB value for 'medium gray' and commonly used for binary black and white thresholding (Image thresholding, OpenCV). Then, I used OpenCV to draw external contours around the silhouette of the teeth, as seen in Figure 2. Each contour is then fed into the elliptical Fourier analysis function from the PyEFD Python package, which supports EFA with various input parameters, such as 'order' (equivalent to harmonics). I used 15 harmonics during my analysis. Four coefficients per contour are returned, and I take an average across all contours for each coefficient channel to get a single set of coefficients per tooth.

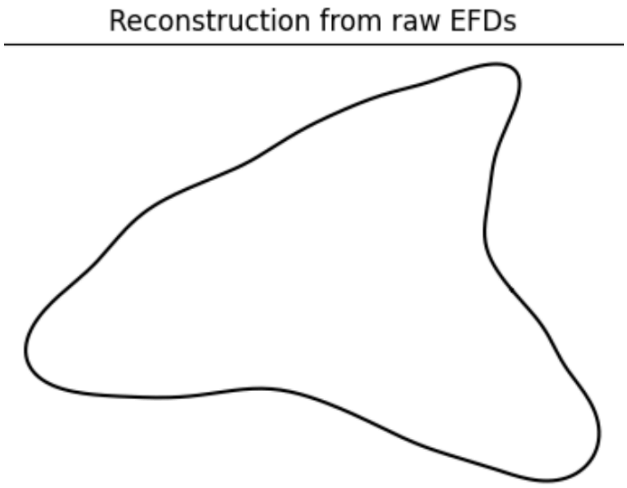


Figure 2: Bull shark tooth contour reconstruction using fifteen harmonics from unnormalized elliptical Fourier descriptors to validate image preprocessing and contour delineation.

Then, I fit these Fourier descriptors to one-dimensional PCA and t-SNE models, effectively reducing the 4D descriptor matrix to a single dimension. For PCA, this is done by assuming a linear relationship across dimensions and maximizing variance along artificial axes or principal components (Goodman et al., 2022). These principal components can be used to describe the variance of the data, preserving information about multiple attributes in one dimension—the core idea behind dimension reduction methods. Alternatively, t-SNE assumes non-linear relationships across attributes and instead converts the distance between points to a probability of two points being "neighbors" (Xiang, 2021). Observations with similar neighbors will be clustered together, preserving local clusters while distorting overall variance, unlike PCA (Xiang, 2021).

The performance of dimension reduction models can be quantified using a silhouette score. The silhouette coefficient can be calculated by taking the difference between the mean value of a predicted data cluster and the mean value of another cluster, over the maximum mean cluster value. This score evaluates how well clusters are separated in a range from -1 (clusters

completely overlap) to 1 (clusters do not overlap), with values close to zero meaning clusters closely border one another. In addition to the silhouette score, we can also calculate the trustworthiness score, which measures how well clusters in high-dimensional space are preserved in our new, reduced dimension. By leveraging these two model metrics, we can quantify the grouping performance across linear and non-linear dimension reduction techniques.

Results

Principal component analysis and t-distributed stochastic neighbor embedding both clearly identify morphological differences between shark species, given a set of unlabeled shark teeth from the same genus, *Carcharhinus*. Taxa clusters and their boundaries are clearly visible, as seen in Figures 3 and 4.

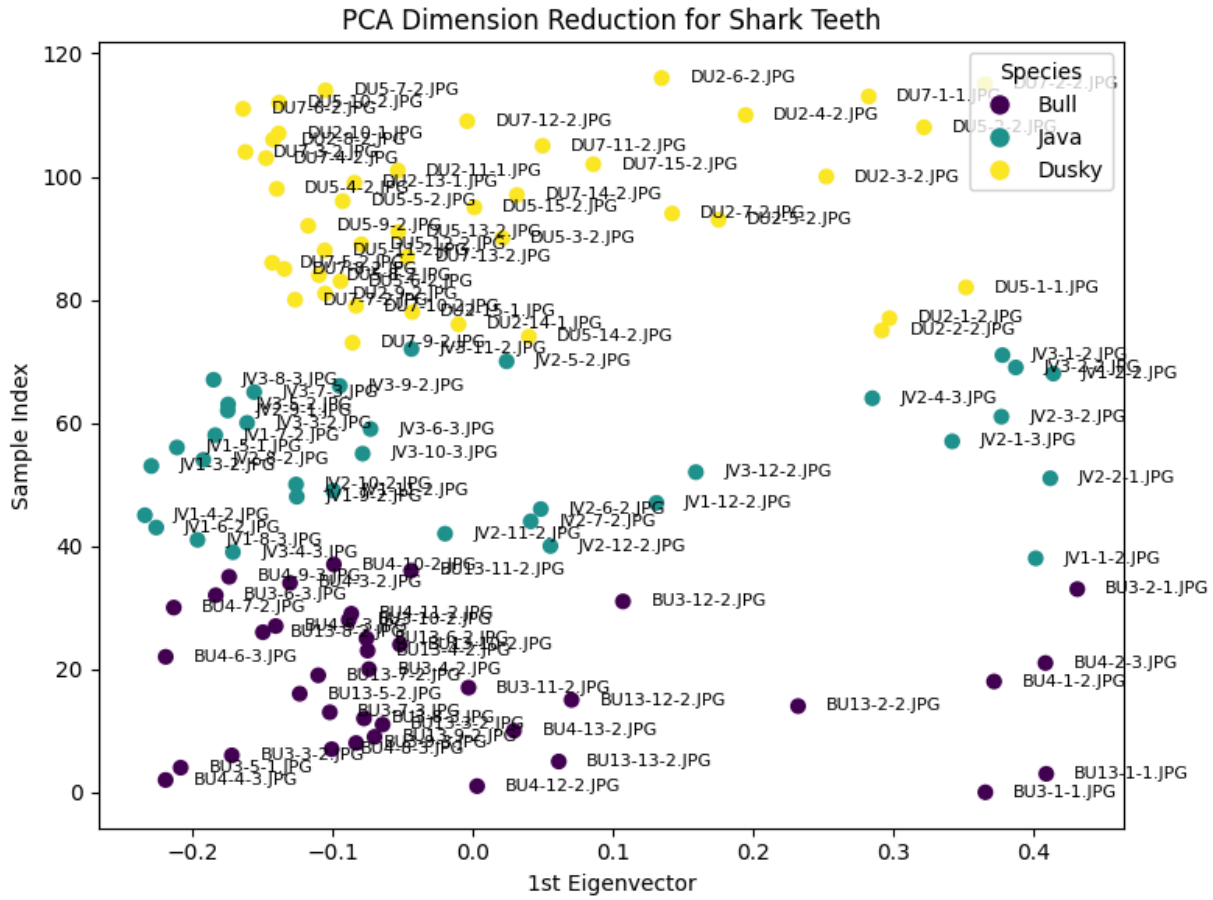


Figure 3: PCA dimension reduction for all shark teeth samples, mapped by species to color and labeled with their associated image files.

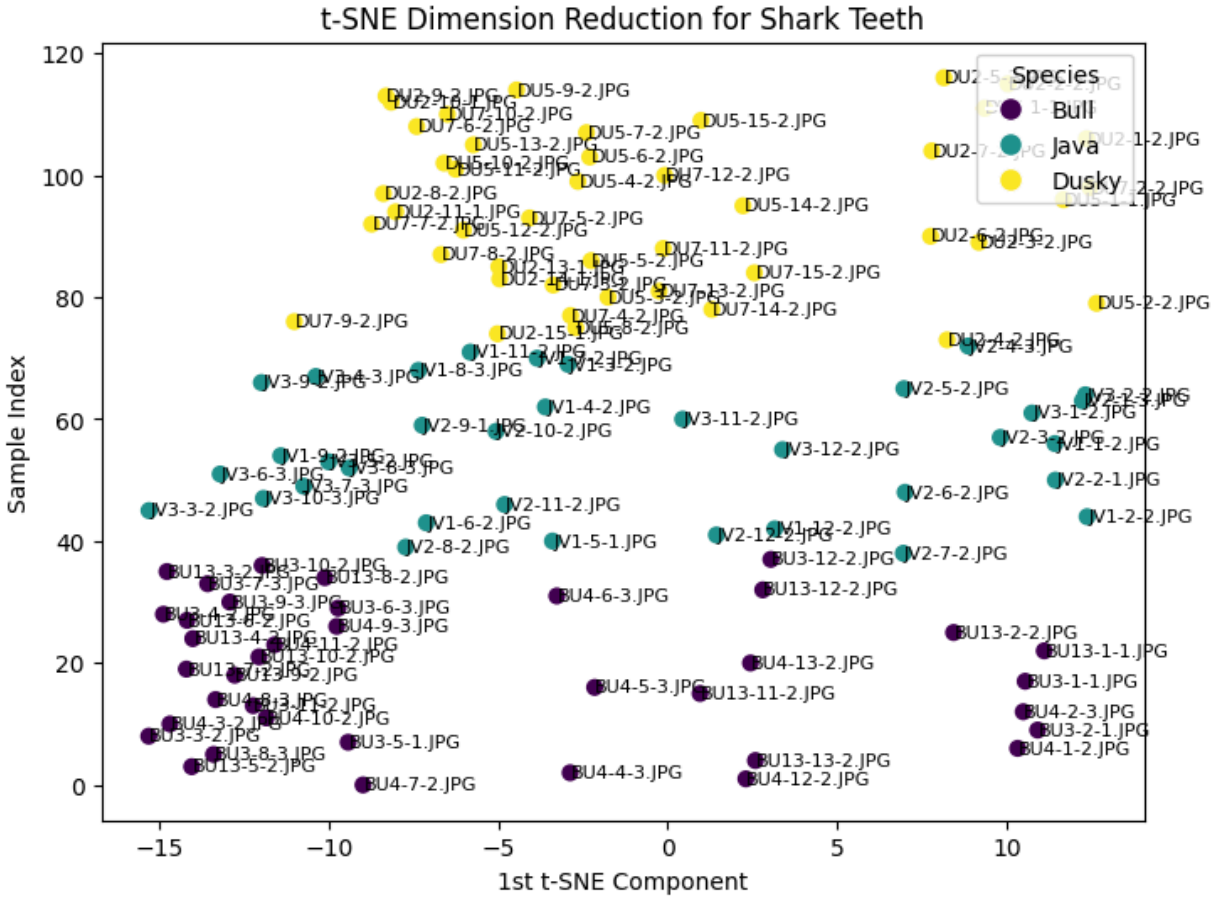
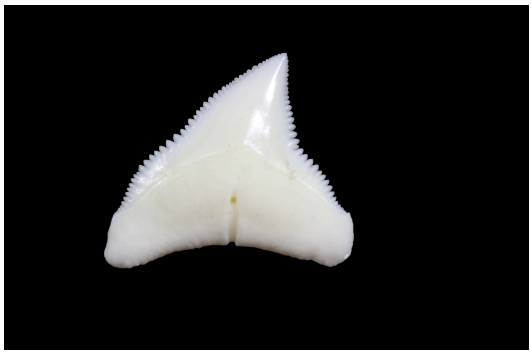
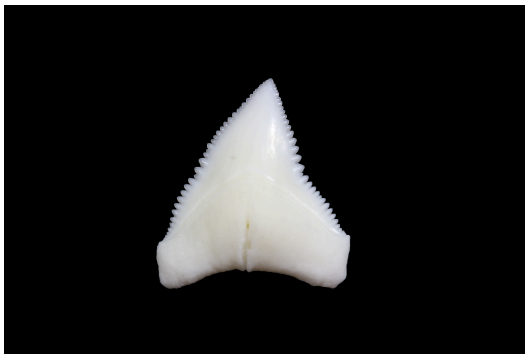


Figure 4: t-SNE dimension reduction for all shark teeth samples, mapped by species to color and labeled with their associated image files.

In addition to clearly defined boundaries, we notice the horizontal variation within each cluster, seen in Figure 5. We can attribute intraspecific variance to morphological differences between posterior and anterior teeth observed in the *Carcharhinus* genus (Goodman et al., 2022). Despite variation in tooth morphology within a species, both models were still able to detect interspecific differences in structure across taxa using EFA.

(a)  (b) 

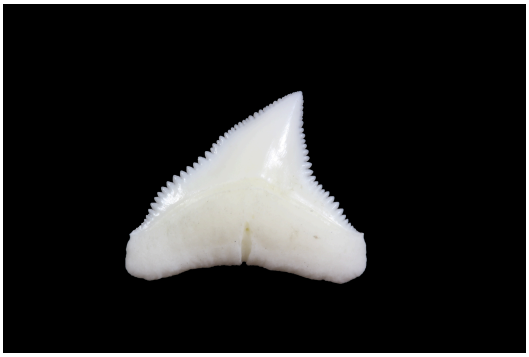
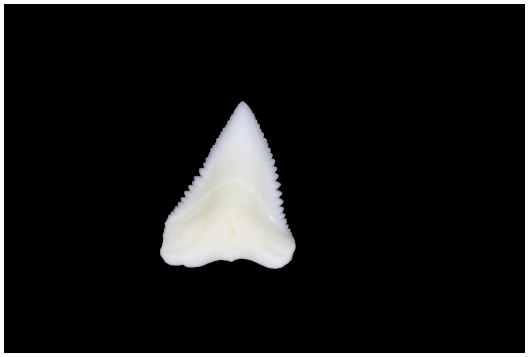
(c)  (d) 

Figure 5: Tooth images with the most variance in the Dusky shark cluster according to PCA and t-SNE graphs, Figures 3 and 4. Both reduction methods recognize the maximum variance within a species to be the variation in tooth curvature between posterior (more curved) and anterior (more straight) teeth. Images a, b, c, and d are "Dusky/DU7-6-2.JPG" (PCA, posterior), "Dusky/DU7-2-2.JPG" (PCA, anterior), "Dusky/DU7-9-2.JPG" (t-SNE, posterior), and "Dusky/DU2-1-2.JPG" (t-SNE, anterior), respectively.

The silhouette score is a model performance metric that quantifies how well a model derives distinct clusters within a set of samples. Values near one represent well-defined clusters with no overlap, while values closer to zero represent clusters with ambiguous boundaries between one another. t-SNE performed one order of magnitude better (-0.0027) than PCA (-0.0614) at defining distinct species clusters measured by the silhouette score, as seen in Table 1. This is likely because t-SNE aims to preserve local neighborhoods while PCA aims to preserve global variance. The silhouette score measures how well neighborhoods are clustered, which t-SNE prioritizes. That being said, it is no surprise that t-SNE performed better than PCA by this metric. However, the silhouette score may not be the best measure of grouping

performance between models due to both the nature of the models and their different intended purposes.

The trustworthiness score is another model performance metric that quantifies the probability of a sample in a given cluster to be sorted into its respective neighborhood, or known taxon. Rather than comparing the distinct boundaries between clusters, as with the silhouette score, the trustworthiness score determines how accurately the model groups a given sample into its assigned neighborhood. By this metric, t-SNE outperforms PCA, correctly sorting samples into their respective taxon groups with 97% accuracy, compared to 79.6% accuracy using PCA, as shown in Table 1.

Dimension Reduction Technique	*Silhouette Score [-1, 1]*	*Trustworthiness Score [0, 1]*
Principal Component Analysis (PCA)	-0.0614	0.7956
t-Distributed Stochastic Neighbor Embedding (t-SNE)	-0.0027	0.9702

Table 1: Comparison of silhouette and trustworthiness score across dimension reduction techniques.

Discussion

My findings suggest that t-SNE can be a powerful tool in unsupervised learning to delineate species from an unknown set of tooth samples in the future. t-SNE and its cluster derivations have applications in the field of paleobiology to distinguish between unknown, ancient shark species, in addition to comparisons between modern sharks to conclude phylogenetic relationships. Intracluster variance can also be utilized to infer about a tooth's jaw placement, such as anterior and posterior positions, as observed in Figure 5.

Both PCA and t-SNE methods identify morphological differences between different species from the same genus that demonstrate just how powerful these tools can be at detecting

small structural variations unseen by the human eye. Although t-SNE performs better at delineating clusters from unlabeled tooth images, its reduction data has little physical meaning, unlike linear models that are intended to preserve global variance. Thus, researchers in the future should select a dimension reduction method (linear or non-linear) depending on their intended purposes. Based on my results, I recommend that t-SNE should be utilized for visualization and clustering for unknown teeth to infer their taxonomic classification. Conversely, PCA should be utilized when performing quantitative analysis on reduction data, where physical significance is required to derive mathematical relationships, as opposed to classification alone.

References

Blidh, H. (2016). PyEFD documentation (Version 1.6.0) [Computer software documentation]. Read the Docs. https://pyefd.readthedocs.io/en/latest/index.html

Cullen, J., Marshall, C. (2019). *Do sharks exhibit heterodonty by tooth position and over ontogeny? A comparison using elliptic Fourier analysis.* Journal of morphology, 280(5), 687-700. https://doi.org/10.1002/jmor.20975

Goodman, K., Niella, Y., Bliss-Henaghan, T., Harcourt, R., Smoothey, A. F., & Peddemors, V. M. (2022). *Ontogenetic changes in the tooth morphology of bull sharks (Carcharhinus leucas).* Journal of fish biology, 101(4), 1033–1046. https://doi.org/10.1111/jfb.15170

Marramà, G., & Kriwet, J. (2017). *Principal component and discriminant analyses as powerful tools to support taxonomic identification and their use for functional and phylogenetic signal detection of isolated fossil shark teeth*. PloS one, 12(11), e0188806. https://doi.org/10.1371/journal.pone.0188806

OpenCV Contributors. (n.d.). Image thresholding – OpenCV (4.x) tutorial [Computer software documentation]. OpenCV. https://docs.opencv.org/4.x/d7/d4d/tutorial_py_thresholding.html

OpenCV Contributors. (n.d.). Structural analysis and shape descriptors – OpenCV (3.4) documentation [Computer software documentation]. OpenCV. https://docs.opencv.org/3.4/d3/dc0/group__imgproc__shape.html

Purdy, R. (1995). Fossil Shark Teeth. The Paleontological Society.

ScienceDirect. (Accessed December 1, 2025). Silhouette Coefficient - an overview. ScienceDirect Topics. https://www.sciencedirect.com/topics/computer-science/silhouette-coefficient

Xiang, R., Wang, W., Yang, L., Wang, S., Xu, C., & Chen, X. (2021). *A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data. Frontiers in genetics*. 12, 646936. https://doi.org/10.3389/fgene.2021.646936