AI-Powered Identification of Dental Calculus Across Species

Xuening Yang

Georgetown University

Dr. Courtney Hofman (Supervisor), Alice Lee (Mentor), Dr. Qiwei (Britt) He (Instructor)

Dec 6, 2024

Abstract

Dental calculus, composed of calcified plaque, serves as an exceptional repository of information due to its ability to preserve diverse inclusion morphologies. However, traditional visual analysis methods for identifying these inclusions are often time-consuming and prone to human error. Recent advancements in image processing and machine learning have achieved significant success in image identification across various scientific disciplines. This project explores the potential of machine learning techniques to classify mammalian diet types, species, and the diverse inclusions of scientific interest commonly observed in dental calculus. By quantitatively analyzing the distinct visual features of dental calculus, this study utilizes machine learning algorithms to enhance our understanding of the microbial and mineralogical content within these samples. Ultimately, this approach aims to equip researchers with a robust tool for analyzing SEM images of archaeological substrates, providing deeper insights into ecological and evolutionary contexts.

Keywords: dental calculus, SEM images, image processing, convolution neural network, clustering analysis

1. Introduction

1.1 Historical and Scientific Overview of Dental Calculus

Dental calculus, or tartar, is a mineralized plaque that stubbornly attaches to the surface of teeth, not only in humans but also in most mammals. With the advancement of modern technology, people have become more conscious of improving their quality of life and now place greater importance on dental hygiene and aesthetics. Visiting the dentist has become a regular routine for prophylaxis to remove this calcified plaque. The formation of dental calculus is primarily considered a consequence of poor oral hygiene. Bacteria in the oral cavity mix with food particles and proteins to form dental plaque, which, if not removed, mineralizes and hardens into calculus. This can lead to serious dental health issues such as gum disease. Interestingly, due to its complex composition and the presence of rich biomolecules, dental calculus holds significant scientific value, offering a wealth of information for research. In 1975, dental calculus was discovered by Armitage to have the ability to capture a variety of microscopic dietary and environmental particles, including phytoliths—opaline silica deposits formed in certain plants. These findings, extracted from dental calculus samples, highlight the potential for analyzing the relationship between these inclusions and dental calculus.

Charlier et al. (2010) mentioned that Dobney developed the first method for microscopically examining dental calculus, enabling the identification of food remnants trapped and preserved within the plaque. It allowed for the detection of materials such as cereal fragments, plant fibers, phytoliths, pollen, seeds, animal hairs, parasites, and even accidentally included insects. This demonstrates the necessity for further investigation into the formation of dental calculus and its inclusions, which can reveal valuable historical information. Furthermore, Forshaw (2019) highlighted that dental calculus has increasingly been recognized as a useful tool

in oral health, forensic studies, archaeology, and anthropological research. Specifically, it traps and preserves human and microbial DNA, including microorganisms like bacteria and viruses, over time. This provides critical insights into dietary habits and living conditions. Notably, dental calculus has even been utilized in COVID-19 investigations, as the SARS-CoV-2 virus can be detected in the oral microbiota. This underscores the broader significance of dental calculus and emphasizes the importance of exploring its contents, such as phytoliths, microbes, diatoms, and other micro-remains.

1.2 Literature Review

1.2.1 Advances in Image Processing Techniques

To analyze the inclusions within dental calculus, many researchers observe images captured by high-precision microscopy with the naked eye and apply their expertise to determine the composition and relationships of the inclusions and dental calculus. Nevertheless, relying solely on classical microscopy (OM) is often insufficient. A study by Power et al. (2014) introduced a novel method for identifying starch and other micro remains in intact human and chimpanzee dental calculus using scanning electron microscopy (SEM) coupled with energy-dispersive X-ray spectroscopy (EDX), successfully identifying these remains. This indicates that SEM is one of the advanced and complementary tools widely used, as it generates high-resolution images capable of capturing nanoparticles and their structures, making it ideal for visualizing biological samples. However, Power et al. (2014) also discussed the differences in sensitivity between SEM-EDX and optical microscopy (OM), emphasizing that SEM-EDX was a complementary technique rather than a replacement for OM in the study of dental calculus micro remains. To achieve the highest resolution in detecting these remains, a workflow that begins

with SEM-EDX analysis followed by OM is recommended. Additionally, the approach was noted to be time-intensive, potentially costly, and reliant on careful interpretation, highlighting the limitations of using advanced SEM techniques for experimentation (Power et al., 2014). This raises concerns about modern contamination during sample processing, which could introduce errors into the manual identification process. Manual analysis is indeed prone to errors; therefore, precision instruments and machinery, along with advanced algorithms, can aid in automatically identifying and extracting relevant features with minimal human intervention, reducing cognitive biases and inconsistent judgments when manually analyzing datasets.

An increasing number of research analyses are aided by computer systems and artificial intelligence (AI) to enhance accuracy through automated image processing and reduce the likelihood of human errors. According to Saladra and Kopernik (2016), the use of these image processing algorithms significantly enriched the interpretation of SEM images, improved both qualitative and quantitative assessments, and greatly helped researchers better visualize and measure material defects. Additionally, image processing techniques have been applied in Alzheimer's Disease Computer-Aided Diagnosis using Positron Emission Tomography brain images, before using image-based machine learning (Fu'adah et al., 2021).

Image processing is a crucial task for studies requiring the transformation of images into readable data. In the model training process described by Wang et al. (2023), each image had a height and width, with every pixel occupying a specific position. Additionally, images had a third dimension: the color channel. In grayscale images, this channel represents shades of gray, ranging from 0 (black) to 255 (white). According to the schematic diagram of the model training process, multi-color images like RGB (Red, Green, Blue) had three channels, each corresponding to a primary color. Each channel was represented as a 2D matrix indicating pixel intensity for

that specific color (Wang et al., 2023). When combined, these three matrices formed a 3D representation of the image. This process of breaking an image into multiple matrices allows for efficient image analysis and manipulation, making image processing essential for machine learning and computer vision tasks.

1.2.2 Application of Machine Learning and Deep Learning in SEM Image Classification

After converting images into the data, many research studies have incorporated machine learning techniques to enhance the accuracy of experimental results. Machine learning models are increasingly being applied across various fields, especially in health science research, where clinical diagnostics often require investigations at the nanoscience and biomicro scales. Wang et al. (2023) proposed an image modeling method that combines hyperspectral fluorescence imaging and machine learning to enhance recognition, quantitative assessment, and maintain accuracy in early-stage caries diagnosis. By fusing spectral, texture, and color features, the integrated learning algorithm demonstrated superior performance and stronger generalization capabilities (Wang et al., 2023). The study evaluated various models using four traditional machine learning algorithms: Integrated Learning, Support Vector Machine (SVM), Decision Tree (DT), and Artificial Neural Networks (ANN). This combination of imaging and machine learning revolutionizes clinical diagnostics for dental health, particularly in caries and calculus detection, by excelling in pixel-level classification and enabling early-stage diagnosis.

Deep learning techniques, compared to traditional supervised machine learning methods, offer distinct advantages that traditional algorithms cannot achieve. Aversa et al. (2020) suggested that deep learning models were more capable of accurately predicting unseen target categorical variables with minor representations. One of the deep learning techniques examined

by Aversa et al. (2020) was the use of advanced Artificial Neural Networks, such as Convolutional Neural Networks (CNNs) like Inception-v3, Inception-v4, and Inception-ResNet-v2 for image classification. Feature extraction and fine-tuning were employed to further improve the accuracy of the classification model (Aversa et al., 2020), which explores the application of deep learning techniques, specifically artificial neural networks, to the classification and analysis of SEM images, a powerful tool in nanoscience and materials science.

Moreover, Artificial Neural Network models are increasingly used for disease classification and detection. Fu'adah et al. (2021) highlighted the application of deep learning, specifically CNN techniques, for analyzing MRI images. The study demonstrated that CNNs could serve as a baseline approach for detecting patterns in MRI data. Fu'adah et al. (2021) successfully classified Alzheimer's disease into four categories: non-demented, very mild demented, mild demented, and moderate demented, using a relatively small dataset of 664 points, with 166 reserved for testing, achieving reliable prediction results. The model was evaluated using metrics such as accuracy, recall, precision, and F1-scores, which were essential for categorical regression models, and it achieved a high accuracy score. However, the choice of the loss function was inappropriate. They chose binary cross-entropy to label Alzheimer's disease stages, which was typically used for binary classification tasks. For multi-class classification, categorical or multi-class cross-entropy would be more appropriate. Consequently, in dental calculus research, binary cross-entropy is appropriate for binary classification tasks, such as predicting whether a species is herbivorous or carnivorous. However, when predicting the species type itself, categorical cross-entropy would provide more accurate evaluation results. What's more, this research primarily utilizes the AlexNet architecture, which consists of five convolutional layers, with max-pooling or average-pooling used for dimensionality reduction

throughout the layers. AlexNet, introduced in 2012, serves as the baseline model for this research. More advanced techniques should be explored to improve and fine-tune deep learning outcomes if computing resources become available.

1.2.3 Clustering Analysis in SEM Image Grouping

Considering the grouping of SEM images without predefined labels, or uncovering morphological or compositional similarities and differences in the dataset can use clustering analysis as it is widely used in biological data in recent advanced research. The DensityCut algorithm was introduced as a novel density-based clustering approach designed for biological data, such as cancer mutation clustering and single-cell analyses, and it could effectively cluster irregular shape synthetic benchmark datasets. DensityCut could broadly be used for exploratory data analysis without making assumptions about the shape, size and the number of clusters (Ding et al, 2016). Though DensityCut relies heavily on density separation within the dataset and it may pose a challenge on handling high-dimensional SEM image datasets, it highlights the efficiency of the clustering approach for categorizing groups.

In parallel, Cohn and Holm (2020) explored transfer learning combined with K-means clustering, to classify images in the Northeastern University Steel Surface Defects Database (NEU-SSDD), which highlighted the potential of unsupervised learning for extracting patterns from unlabeled datasets and achieving high classification performance without requiring extensive labeled data. Therefore, using clustering to group dental calculus morphologies can enhance the ability to group and interpret inclusions effectively. This aligns with the goal to uncover patterns that relate to dietary and ecological influences on dental calculus inclusions.

1.3 The Present Study

In a nutshell, present research focuses on the application of technology to SEM images in the mineralized structure or nanoscience field. However, the identification and assessment of inclusions in dental calculus have not yet been explored. A novel method using more advanced algorithms may uncover patterns within dental calculus, revealing the wealth of information it contains and potentially leading to new applications for its use.

Traditionally, dental calculus had been primarily used in forensic studies, drug entrapment, human DNA analysis, and other areas to provide insights into diets and health, particularly the consumption of plant-based remedies and treatments (Forshaw, 2019). Although the process of extracting and recovering embedded materials, such as microfossils, can be very challenging, the valuable information obtained is worth investigating and will provide exciting data. Forshaw (2019) emphasized that analyzing inclusions in dental calculus is valuable because it contains significant information for understanding diet and environmental influences.

While no current studies have specifically focused on identifying species-specific inclusions in dental calculus, research in related areas, such as the automatic identification and diagnosis of dental caries and calculus using machine learning and hyperspectral fluorescence imaging or the application of machine learning methods to rock samples, had yielded promising results, and using transfer learning and clustering combined method on images had also achieved good performance (Li et al. 2021; Cohn and Holm, 2020). These indicated that applying image processing, machine learning, and deep learning techniques, clustering to SEM images of dental calculus to identify mammalian species, detect patterns, and analyze inclusions is worth exploring. Advanced analytical techniques and tools are leveraged to enhance the scientific

understanding of dental calculus, its analysis, and its relevance in fields such as microbiology, anthropology, and life sciences.

1.4 Objective and Research Questions

To bridge the potential gap, this research aims to explore whether computer vision techniques, such as image processing, machine learning, deep learning, and clustering, can identify the presence of specific dietary inclusions in dental calculus. Additionally, the study will assess whether these techniques can distinguish species-specific patterns in dental calculus, particularly the calcified dental plaque morphology, among different mammals. Using machine learning algorithms to identify these inclusions provides further insight into their relationship with dietary habits and oral microbiome diversity.

RQ1: Can machine learning be utilized to distinguish dental calculus between different species or other distinguishing factors, such as diet types?

RQ2: When analyzing the morphology of inclusions identified in SEM images, can clustering techniques produce meaningful results? Additionally, which clustering methods are most suitable for grouping the inclusions into categories such as phytoliths, microbes, and diatoms?

2. Method

2.1 Data Description

This study utilizes SEM images of dental calculus from non-human mammalian teeth, primarily sponsored by the University of Oklahoma's Laboratories of Molecular Anthropology

and Microbiome Research (LMAMR). The images were originally obtained from the Smithsonian Institution's National Museum of Natural History, Division of Mammals Collection. The animal species are primarily from North and South America, with a few species from Africa and Asia. The sample data consists of SEM images, which are raster graphics stored in .TIF and .JPG formats. These images were represented as matrices that depict two-dimensional pictures through a grid of pixels and can be rendered on a computer display (Robertson, 2001). The images were captured under various SEM settings, with 4 to approximately 20 images per subsample. Each image has different resolutions where most of them are primarily taken at magnifications of 50x, 500x, 2000x, and 5000x. Higher magnifications indicate a zoom-in version of the images which are taken at the same dental calculus but from a different landscape.

The total dataset contains 572 images collected from 29 subsamples across 12 different animal species, as shown in Table 1. *Group* indicates the dietary type, with three categories: Foregut, Hindgut, and Carnivore. *Species Identification* specifies the mammalian species group in detail. *Catalog No.* is a unique identifier assigned to all images taken from a specific subsample of a species. Each individual animal may belong to the same species identification category. *Magn* stands for magnification, with each column representing images captured at that magnification range. *Total* is the sum of images for each catalog, providing an overall count of individual species.

Insert Table 1 around here.

Though the data was collected through a prestigious institution, verified for resolution and focus, and underwent quality control in data annotation by LMAMR, the dataset has limited representation from geographic regions beyond the Americas, which may limit comprehensive

global insights. Besides, there is taxonomic bias due to an imbalance in the number of animals within certain groups; for example, carnivores have fewer species represented. This imbalance may lead to underrepresentation of other mammalian groups, potentially limiting findings related to digestive diversity or causing overfitting on traits specific to more highly represented species. Moreover, the analysis in this study relies on image data for pattern recognition and detection, which could affect the accuracy of the predictive models developed.

In addition, an important attribute of the data is that some SEM images of the same animal are identical but have different adjustments when taken in contrast, brightness, and other image attributes, while others are zoomed-in versions of a different image. For example, in Figure 1, the images are identical but have different contrast levels. This overlap of information across images reduces the dataset's unique information content. Therefore, meticulous calibration and adjustments are necessary for an in-depth analysis.

Insert Figure 1 around here.

2.2 Instruments

According to the LMAMR, dental calculus samples were selected for sub-sampling based on their weight. Samples with higher weights (in grams) were sub-sampled, as they presumably contained enough material to allow for both analysis and retention for microbiome studies. The samples were mounted onto sterilized aluminum disks for SEM imaging at the Ancient DNA Laboratory at the LMAMR. Image collection was based on four standard magnifications stated previously (50x, 500x, 2000x, and 5000x), with additional exploratory images taken to capture any distinct features observed on the calculus during SEM viewing by lab researchers.

To identify the primary diet groups, the original labels were initially categorized into three groups: Foregut, Hindgut, and Carnivore, as previously shown in Table 1. In Table 2, these categories are listed under *Old Label*, where they were reclassified into a *New Label* for binary classification. Foregut and Hindgut were combined into a single Herbivores category, while Carnivore remained unchanged. This binary classification enables the use of SEM image data to predict whether a species is herbivorous or carnivorous. *Number of Images* represents the total count of images for each species, while *Percentage* shows the proportion of each dietary type within the total sample, revealing an imbalance between the two categories. To address this imbalance, image augmentation was applied to increase the data size for the Carnivore category, thereby expanding the overall dataset.

Insert Table 2 around here.

To identify specific animal species, this involves a multi-class classification task with a total of 12 species, using images of their dental calculus, as shown in Table 3. *Species* lists the different animal species regrouped from the *Species Identification* column in Table 1. *Number of Images* represents the total count of images for each species. *Percentage* indicates the proportion of images for each species relative to the total dataset.

The images are mostly evenly distributed, with the exception of Antilocapra and Ovis Mexicana. If the images were uniformly distributed, each species would represent approximately 8.5% of the dataset. However, Rhinoceros Unicornis has the highest representation at 14%, while Ovis Mexicana has the lowest at 2%. Given a total of 572 images across 12 classes, the dataset size may be insufficient for robust machine learning tasks. This aligns with the purpose of applying image augmentation to increase the dataset size and enhance model performance.

Insert Table 3 around here.

2.3 Design / Procedure

2.3.1 Image Processing

In earlier years, Tan, Zhang, and Gao (1997) applied image processing to analyze food structures. Through correlation and regression analyses, the results of image processing were compared with manual measurements, demonstrating that SEM processing algorithms are highly effective for quantitatively analyzing the structure of puffed foods and extracting image features for quantitative analysis (Tan, Zhang, & Gao, 1997). Therefore, in this study, image processing techniques are employed to enhance the quality and visibility of high-resolution dental calculus SEM images and to extract feature information for subsequent deep learning analysis. In this context, dental calculus image data appeared to be in grayscale format, typically with only one channel, where pixel values ranged from black (0) to white (255), representing various shades of gray (Robertson, 2001). In terms of dimensions, a grayscale image is usually represented as Height (H) x Width (W) x Channel (C), where the channel is 1, indicating a single channel for grayscale intensity (Kumar, Brennan, Mileo, & Bendechache, 2024).

Each SEM digital image of dental calculus can be thought of as a two-dimensional surface plot with x- and y-coordinates (Robertson, 2001). To proceed to the preprocessing step, the SEM images of dental calculus are manipulated and processed using the OpenCV library in Python, a powerful tool for handling image processing tasks (Bradski, 2000). Additionally, the Keras library (Chollet, 2015), and the torch and torchvision packages from PyTorch in Python are used for detailed adjustments of the image data (PyTorch, n.d.). All steps in image processing

primarily rely on these Python packages to properly handle the image data before applying deep learning models.

Image data augmentation. Data augmentation in the image processing step was crucial for increasing dataset diversity, preventing overfitting, and improving the generalization of machine learning models by applying various transformations to the existing data (Kumar, Brennan, Mileo, & Bendechache, 2024). Due to limited labeled data in this study, data augmentation is applied to enhance model validation, as Goodfellow, Bengio, & Courville (2016) suggested deep neural networks generally required large datasets to learn complex patterns effectively.

Two types of image manipulations, cropping and resizing are common preprocessing data augmentation techniques. These methods could crop images randomly or at the center of the image (Kumar, Brennan, Mileo, & Bendechache, 2024). To better preserve features and maximize the capture of different inclusions, this study crops the images based on the presence of obvious features or inclusions. To achieve this, one approach is to use RandomCrop(size=(height, width)) from torchvision, which randomly crops the image to the specified height and width. To better preserve information, cropping is also conducted using manual slicing on the image data, where the indices are carefully defined. This allows the image to be cropped into smaller sections, each containing representative information. This approach not only increases the training data size but also maintains the essential information in the original image. For resizing, a default dimensions of (224, 224) were applied to images because this is a standard input size for many convolutional neural network architectures (Krizhevsky, Sutskever, & Hinton, 2012). Another non-geometric manipulation method is the use of kernel filters to enhance or soften the image, such as applying a Gaussian blur or an edge filter. These

filters must be carefully applied, as Kumar, Brennan, Mileo, & Bendechache (2024) implied that they might introduce noise and artifacts. In this study, edge detection is used to identify features of interest and create clear boundaries around dental calculus and inclusion structures for morphological analysis. This process enables the algorithm to recognize and classify different patterns more effectively, aiding in species identification tasks.

Performing transformations on existing dental calculus SEM images as part of image data augmentation increases the size and diversity of the training set before applying deep learning models. The random_split function from torch.utils.data is used to split the dataset into 80% training and 20% test data. This process helps improve model performance by making the training data more representative of species-specific patterns in dental calculus, thereby aiding in the identification of these patterns in real microbiome cases.

Image magnification. Image magnification refers to the degree of zoom applied to each SEM image captured using optical microscopy (OM). A lower magnification indicates that the image was taken from a higher level, capturing more comprehensive information about the dental calculus from a broader, holistic perspective. Conversely, a higher magnification means the SEM image is zoomed in on a specific area of the dental calculus. As a result, the entire dental calculus is not visible in the image; instead, it focuses on detailed features at a specific location. In this research, we grouped the SEM images by their magnification and manually selected images with the same magnification for specific parts of the analysis. This approach ensured that the information contained in each image was comparable, maintaining consistency in scale and perspective to ensure that the comparisons were both meaningful and accurate.

2.3.2 Convolutional Neural Network (CNN)

The deep learning model used in this study is the convolutional neural network (CNN). The basic algorithm discussed by LeCun et al. (1989) identified local features of an image by breaking it down into pixels and aggregating the information to extract higher-level features. The neural network consists of multiple hidden layers, where each layer combines local information from the previous layer to achieve the desired classification outcome.

AlexNet. One of the CNN models used in this study is AlexNet. The AlexNet architecture consists of 8 layers: 5 convolutional layers and 3 fully connected layers (Krizhevsky, Sutskever, & Hinton, 2012). Figure 2 illustrates a modified basic structure of the AlexNet model, showing the dimensions of each convolutional and fully connected layer. According to Krizhevsky, Sutskever, and Hinton (2012), the first convolutional layer filters a 224×224×3 input image with 96 kernels of size 11×11×3 and a stride of 4 pixels. The second layer applies 256 kernels of size 5×5×48 to the pooled and normalized output of the first layer. The third, fourth, and fifth convolutional layers are connected sequentially without pooling or normalization, using 384 kernels of size 3×3×256, 384 kernels of size 3×3×192, and 256 kernels of size 3×3×192, respectively. Each of the final fully connected layers contains 4096 neurons (Krizhevsky, Sutskever, & Hinton, 2012).

A slight modification to the AlexNet architecture, as shown in Figure 2**,** is applied to adapt it for dental calculus SEM images. Since these images are grayscale and contain only one channel, the architecture is modified by setting the input layer's depth to 1 instead of 3. Consequently, the first convolutional layer filters a 224×224×1 input image, enabling AlexNet to effectively process grayscale SEM images.

Insert Figure 2 around here.

Rectified Linear Units (ReLUs), originally introduced by Nair and Hinton in 2010, were used in AlexNet as the activation function (Krizhevsky, Sutskever, & Hinton, 2012). Represented by the formula f(x)=max(0,x), ReLU takes an input x and outputs the maximum of x and 0 (Nair & Hinton, 2010). The ReLU activation function was applied after each convolutional layer, as well as the output layer (Krizhevsky, Sutskever, & Hinton, 2012). The pooling layers in AlexNet are Max-pooling layers. In AlexNet, Max-pooling was applied after response normalization layers and after the fifth convolutional layer (Krizhevsky, Sutskever, & Hinton, 2012). In this study, an open-source AlexNet package based on PyTorch, alexnet_pytorch, is used. This package replicates the model structure introduced by Krizhevsky, Sutskever, and Hinton. A pre-trained AlexNet model is fine-tuned on the dataset to perform both binary and multi-class classification tasks (PyTorch Community, n.d.).

2.3.3 Clustering Analysis

The goal of clustering analysis is to group all images into distinct clusters. Images within each cluster share similar characteristics, while those in different clusters exhibit distinct differences. K-means clustering, one of the most popular clustering techniques used today, was originally introduced by Kaufman & Rousseeuw (1990). It calculates the distance of each data point from the initially assigned centroids and associates the data with the centroid that has the shortest distance. The centroids are then recalculated based on the data points assigned to each group. This process is repeated iteratively until the clusters stabilize, eventually forming k

clusters based on the initially chosen value of k . This results in grouping all images into well-defined clusters.

As an alternative, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was developed by Ester et al. in 1990. This clustering method offered advantages over traditional clustering approaches, as it required only a single input parameter and provided guidance to users in selecting an appropriate value (Ester et al., 1990). DBSCAN could identify clusters of arbitrary shapes and remains efficient even when applied to large spatial datasets (Ester et al., 1990). Additionally, it effectively addresses the challenge of determining the optimal number of clusters in advance, instead deriving the number of clusters directly from the analysis results. Both clustering methods were employed in this research to define labels for the images and inclusions extracted from the SEM images.

2.3.4 Logistic Regression

We chose the logistic regression model as an alternative shallow supervised machine learning approach, contrasting it with the deep learning convolutional neural network. The machine learning model uses the values from each pixel of the images as one dimension of the input data to predict the final classification group. Logistic regression, originally introduced by Cramer (2002), builds on the formula of a linear regression model and applies a logistic function to transform the output range into [0,1], making it suitable for classification tasks.

$$p(X) = \frac{1}{1+e^{-(aX+b)}}$$

where X is the feature vectors, a is the coefficient vector for the feature space, b is the bias term

The final classified group is determined by a threshold: if the predicted probability is greater than or equal to the threshold, the data is classified into group 1; otherwise, it is classified into group 0. This shallow supervised machine learning method serves as a good substitute for the original deep learning models, requiring less training data and computational power while still maintaining strong predictive capabilities for distinguishing differences between SEM images across different groups.

3. Results

3.1 RQ1: Deep Learning Model for Classification of SEM Images by Species or Diet Types

After processing all SEM images, the pre-trained deep learning model, AlexNet, was evaluated for its classification performance. In the first part of the classification task, the labels were divided into two groups: carnivores and herbivores. The confusion matrix, shown in Appendix Figure 1, revealed that the model classified 115 out of 115 images all as herbivores and none of the images as carnivores. Notably, all 34 carnivore images (29.5% of the dataset) were misclassified as herbivores. Although the accuracy was 81/115 = 70.4%, the results demonstrate the model's inability to distinguish between carnivores and herbivores. Even when additional weight was assigned to the carnivore group, the model still failed to classify the images into two distinct groups. In the second part of the classification tasks, additional labels were created based on specific species types. Six species groups were created for this multi-class classification task. The confusion matrix, shown in Appendix Figure 2, exhibited similar results to the binary classification task. The model was able to predict species in Group 1 and Group 3 but failed to differentiate between the other species groups.

3.1.1 Clustering Analysis for Determining Image Class Labels

To verify the optimal grouping of mammalian types in the dataset, an unsupervised learning approach, clustering, was employed. Three clustering methods, including DBSCAN Clustering, K-Means Clustering, and Agglomerative Clustering, were applied to classify the dataset into distinct groups. DBSCAN was able to choose the optimal number of clusters and initially separated the dataset into three clusters labeled as -1, 0, and 1. PCA dimensionality reduction was performed for simplicity in visualization (Appendix, Figure 2). The majority of the subsample images belonged to Cluster -1, comprising approximately 97 out of 130 images. Cluster 0 contained 23 images, and Cluster 1 included 10 images, both representing a small portion of the dataset. The cluster labeled as 1 represented noise or outliers, as it resulted from a single image.

Thus, the clustering process was refined to group the dataset into two clusters and two clusters were used as the input parameters for K-Means Clustering and Agglomerative Clustering. For enhanced interpretability, PCA dimensionality reduction was applied to project the data into a two-dimensional space. For K-Means, the data points in the reduced feature space appeared to be evenly distributed across the clusters and were effectively distinguished into two groups. Similarly, Agglomerative Clustering (Appendix, Figure 4) achieved clustering separation comparable to that of K-Means Clustering (Appendix, Figure 3). Multiple linkage criteria, including "ward," "complete," "average," and "single," were tested to evaluate the performance of Agglomerative Clustering. Among these, the "ward" linkage produced the most distinct separation between the clusters. This result was consistent with the performance of K-Means Clustering. The two clusters were compared with the original "carnivores" and "herbivores"

groups to determine the reliability of the original binary class labels. As shown in Figure 3, the left plot visualized the true labels in the dataset, where Class 1 and Class 0 represented the two predefined categories. The right plot illustrated the clustering results, where the dataset was grouped into Cluster 1 and Cluster 0 based on the applied K-means clustering algorithm. The results indicated that the two clusters generated by K-Means and Agglomerative Clustering align well with the original "carnivores" and "herbivores" groups.

Insert Figure 3 around here.

3.1.2 Shallow Method: Logistic Regression for Classifying Herbivore and Carnivore

Clustering methods provide confidence that the original binary class labels were good indicators for the SEM images. Other supervised machine learning methods, such as logistic regression, were applied following the clustering analysis to further distinguish the differences between the images. The 130 total subsamples were split into training and testing datasets using an 80/20 training-to-test split. In Table 4, logistic regression achieved an accuracy of 73% across all 26 test samples, indicating a reasonable ability to classify herbivore and carnivore samples based on the PCA-reduced features. The higher F1 score for Class 1 suggested that the model performed better at identifying herbivore samples compared to carnivore samples, which may reflect underlying differences in the feature distributions.

Insert Table 4 around here.

The confusion matrix, shown in Figure 4 (Left), summarized the classification results for 26 test samples obtained via PCA. The model correctly classified 16 herbivore samples (true

positives) and three carnivore samples (true negatives). Five carnivore samples were misclassified as herbivores (false positives), and two herbivore samples were misclassified as carnivores (false negatives). To further evaluate the model, the Receiver Operating Characteristic (ROC) curve, shown in Figure 4 (Right), yielded an Area Under the Curve (AUC) of 0.87. The AUC value demonstrated that the model had a strong overall ability to distinguish between the two classes. This result aligned with the relatively high recall for herbivores and the reasonable classification accuracy.

Insert Figure 4 around here.

Clustering analysis revealed that the current images did not show strong distinctions between species groups. However, the class labels for diet types (carnivores and herbivores) were validated through unsupervised clustering methods. Logistic Regression, used as a shallow classification method, successfully predicted diet types, effectively differentiating between carnivores and herbivores based on the current images. By analyzing the differences in the selected images at 50× magnification, research could investigate the biological distinctions between species of the two diet types.

3.2 RQ2: Unsupervised Learning for Grouping Inclusions Morphologies

In addition to the classification tasks using machine learning methods to assist in distinguishing differences between SEM images, creating groups of inclusion morphologies was another research direction. A total of 36 inclusion images were cropped from the original SEM images, and each image was also randomly cropped into 10 smaller 100 × 100 images. After feeding all 360 dental calculus SEM images into the clustering algorithm, the results are shown

in Figure 5. The data points were grouped into five clusters (Cluster 0 to Cluster 4) based on PCA-reduced two-dimensional features derived from an original 10,000-dimensional feature vector. The clustering was performed without predefined labels to explore potential groupings of morphological features or inclusions within the dataset. The results did not show clear distinctions or separations corresponding to specific morphological categories.

Insert Figure 5 around here.

---

4. Discussion

The current study demonstrated that dental calculus SEM images can be classified into dietary groups using computer vision techniques at 50x magnification. By comparing the results with predefined labels, logistic regression achieved an AUC-ROC score exceeding 80%, validating its effectiveness in distinguishing between the two dietary groups: herbivore and carnivore. Cropping images to focus on the center at 50x magnification, where key objects are typically located, further enhanced the accuracy of the classification. However, the morphological grouping of inclusions using clustering methods failed to produce meaningful distinctions, highlighting the need for validation with domain-specific knowledge to enhance interpretability and reliability.

Unlike previous studies, this research did not achieve good classification results using the pre-trained AlexNet model applied to grayscale SEM images. The AlexNet structure, as used in Fu'adah et al. (2021), was considered less suitable for this dataset due to its relatively small size. As a result, a shallow method, logistic regression, was adopted as a simpler and more effective approach for handling the limited data. Furthermore, while this study employed similar image preprocessing techniques to those described by Wang et al. (2023), the focus differed

significantly. Wang et al. (2023) utilized a machine-learning-based approach combining hyperspectral fluorescence imaging with multiple models to automatically identify and diagnose dental caries and calculus. In contrast, our study relied solely on SEM images and focused on dietary group classification rather than disease diagnosis.

4.1 Limitations and Future Research

Although this study yielded promising results, several challenges remain in utilizing computer vision to recognize and classify the complex structures of dental calculus across species, especially when images are captured at varying magnifications. Addressing these limitations is essential for advancing this field of research.

First, the incorporation of domain knowledge from bioarchaeologists and paleontologists could significantly enhance the analysis. Expert-annotated labels for inclusions, such as phytoliths, microbes, and diatoms, would provide a more robust foundation for identifying features with highly variable morphologies. Additionally, improving the precision of image cropping to preserve key objects in SEM images could enhance the accuracy of morphological group comparisons, reducing the risk of losing critical structural information during preprocessing. Second, exploring more advanced combination methods, such as those similar to DensityCut proposed by Ding et al. (2016) or a hybrid approach combining CNNs with clustering methods, as assessed and found reliable by Cohn and Holm (2020), could yield promising results. While the idea of combining unsupervised and supervised learning remains reasonable for testing, the specific statement about integrating methods from the cited studies may require further refinement. Future research should evaluate whether a hybrid methodology can enhance the classification of complex patterns in dental calculus. Third, the dataset used in

this study, collected from LMAMR at the University of Oklahoma and sourced from the Smithsonian Institution's National Museum of Natural History Division, was limited in size and diversity. Deep learning models typically require extensive and varied training data to achieve high performance. Expanding the dataset to include larger and more diverse samples would enhance model robustness and generalizability across different species and environmental contexts.

By addressing these limitations, future research can build upon the findings of this study to develop more effective computer vision tools for analyzing dental calculus. This work holds the potential to advance our understanding of historical diets and environments, bridging the fields of bioarchaeology, paleontology, and artificial intelligence.

4.2 Conclusion and Implications

Despite its limitations, this study demonstrates that machine learning can assist in classifying species into dietary groups by identifying significant differences analyzed from SEM images, while reducing susceptibility to human errors. Image augmentation is a critical component of data preprocessing, often involving detailed and careful selection to achieve optimal machine learning results. This step typically requires substantial time and effort to fine-tune and adjust the images effectively.

This research lays the groundwork for future studies utilizing robust machine learning algorithms to detect inclusions such as phytoliths, microbes, and diatoms in dental calculus. By advancing these techniques, researchers can refine automated analysis methods to more accurately identify complex morphological patterns. These advancements highlight the potential of combining computer vision with domain-specific knowledge to analyze SEM images and

deepen our understanding of historical dietary habits, ecological interactions, and species-specific traits, paving the way for broader interdisciplinary applications in bioarchaeology and paleontology.

5. Acknowledgments

I sincerely thank Dr. Courtney Hofman and Alice Lee for their invaluable expert guidance and unwavering support throughout this project. I am also deeply grateful to Dr. Qiwei (Britt) He for her insightful leadership and guidance as the capstone course lead. Their contributions and encouragement have been instrumental to the completion of this research.

References

Aversa, R., Coronica, P., De Nobili, C., & Cozzini, S. (2020). Deep learning, feature learning, and clustering analysis for SEM image classification. *Data Intelligence, 2*(4), 513-528. https://doi.org/10.1162/dint_a_00062

Armitage, P. L. (1975). The extraction and identification of opal phytoliths from the teeth of ungulates. Journal of Archaeological Science, 2, 187–197.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Chollet, F. (2015). *Keras*. GitHub. Retrieved from https://keras.io

Cramer, J. S. (2002). The Origins of Logistic Regression (Discussion Paper No. 02-119/4). Tinbergen Institute. Retrieved from https://papers.tinbergen.nl/02119.pdf

Charlier, P., Huynh-Charlier, I., Munoz, O., Billard, M., Brun, L., & Lorin de la Grandmaison, G. (2010). The microscopic (optical and SEM) examination of dental calculus deposits (DCD): Potential interest in forensic anthropology of a bio-archaeological method. *Legal Medicine, 12*(4), 163–171. https://doi.org/10.1016/j.legalmed.2010.03.003

Cohn, R., & Holm, E. (2020). Unsupervised machine learning via transfer learning and k-means clustering to classify materials image data. Integrating Materials and Manufacturing Innovation. https://doi.org/10.1007/s40192-020-00177-1

Ding, J., Shah, S., & Condon, A. (2016). DensityCut: An efficient and versatile topological approach for automatic clustering of biological data. Bioinformatics, 32(17), 2567–2576. https://doi.org/10.1093/bioinformatics/btw227

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD), 226–231.

Forshaw, R. (2022). Dental calculus: Oral health, forensic studies, and archaeology: A review. British Dental Journal, 233(11), 961–967. https://doi.org/10.1038/s41415-022-5266-7

Fu'adah, Y. N., Wijayanto, I., Pratiwi, N. K. C., Taliningsih, F. F., Rizal, S., & Pramudito, M. A. (2021). Automated classification of Alzheimer's disease based on MRI image processing using convolutional neural network (CNN) with AlexNet architecture. *Journal of Physics: Conference Series*, *1844*(1), 012020. https://doi.org/10.1088/1742-6596/1844/1/012020

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *ImageNet classification with deep convolutional neural networks*. In *Advances in neural information processing systems* (Vol. 25)

Kaufman, L., Rousseeuw, P. (1990). Finding Groups in Data: An Introduction To Cluster Analysis. 10.2307/2532178.

Kumar, T., Brennan, R., Mileo, A., & Bendechache, M. (2024). Image data augmentation approaches: A comprehensive survey and future directions. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3470122

Li, C., Wang, D., & Kong, L. (2021). Application of machine learning techniques in mineral classification for scanning electron microscopy - Energy dispersive X-ray spectroscopy (SEM-EDS) images. *Journal of Petroleum Science and Engineering, 200,* 108178. https://doi.org/10.1016/j.petrol.2020.108178

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). *Backpropagation applied to handwritten zip code recognition*. AT&T Bell Laboratories. Retrieved from https://henriquetmaia.github.io/pdf/papers/lecun1989.pdf

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*

O'Shea, K., & Nash, R. (2015). *An introduction to convolutional neural networks*. arXiv preprint arXiv:1511.08458.

Power, R. C., Salazar-García, D. C., Wittig, R. M., & Henry, A. G. (2014). Assessing use and suitability of scanning electron microscopy in the analysis of micro remains in dental calculus. Journal of Archaeological Science, 49, 160-169. https://doi.org/10.1016/j.jas.2014.04.016

PyTorch Community. (n.d.). *alexnet-pytorch* (Version 0.2.0) [Computer software]. PyPI. https://pypi.org/project/alexnet-pytorch/

PyTorch. (n.d.). *Torchvision: PyTorch's computer vision library*. Retrieved from https://pytorch.org/vision/stable/index.html

Robertson, D. (2001). *Introduction to computer graphics and the OpenGL application programming interface*. Hobart and William Smith Colleges. Retrieved from https://math.hws.edu/graphicsbook/c1/s1.html

Saladra, D., & Kopernik, M. (2016). Qualitative and quantitative interpretation of SEM images using digital image processing. *Journal of Microscopy, 264*(1), 102-124. https://doi.org/10.1111/jmi.12431

Strang, G. (2023) Introduction to Linear Algebra. 6th Edition, Wellesley Cambridge Press.

Stewart, J. (2015). *Calculus: Early Transcendentals* (8th ed.). Brooks Cole.

Tan, J., Zhang, H., & Gao, X. (1997). SEM image processing for food structure analysis. *Journal of Texture Studies, 28*(6), 657–672. https://doi.org/10.1111/j.1745-4603.1997.tb00145.x

Wang, C., Zhang, R., Wei, X., Wang, L., Xu, W., & Yao, Q. (2023). Machine learning-based

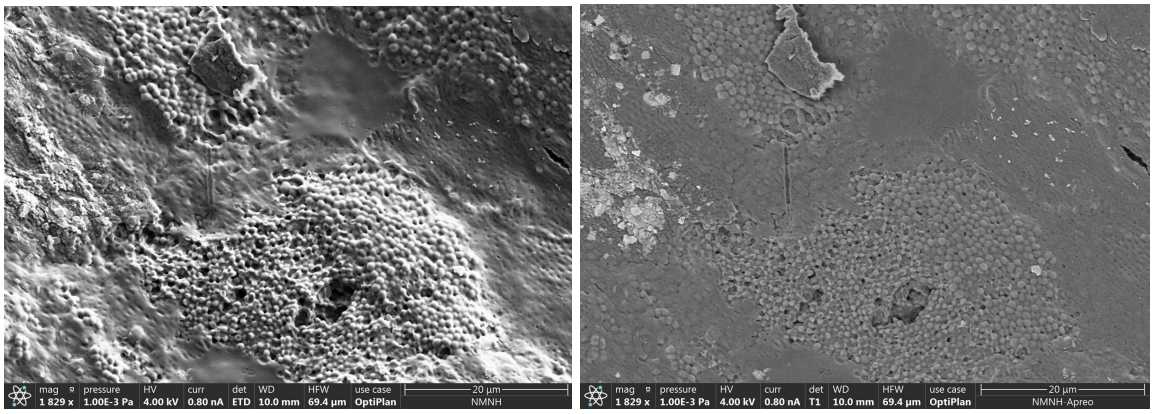
automatic identification and diagnosis of dental caries and calculus using hyperspectral fluorescence imaging. *Photodiagnosis and Photodynamic Therapy, 41*, 103217. https://doi.org/10.1016/j.pdpdt.2022.103217:contentReference[oaicite:0]{index=0}

Table 1. Sample description by group, species identification, catalog number, five different magnification scales, and total number of images for each catalog.

Group	Species Identification	Catalog No.	Magn 0-50	Magn 51-500	Magn 501-2000	Magn 2001-5000	Magn 5001+	Total
Foregut	Ovis canadensis canadensis	209419	3	7	8	1	0	19
	Ovis canadensis mexicana	118257	1	5	3	3	0	12
	Ovis canadensis cremnobates	139722	0	18	6	6	2	32
	Ovis canadensis cremnobates	147511	2	7	9	4	3	25
	Ovis canadensis canadensis	240289	2	5	5	5	2	19
	Ovis canadensis canadensis	240959	1	1	4	2	4	12
	Bison bison	122731	2	2	2	2	3	11
	Bison bison	122732	2	3	9	5	4	23
	Bison bison	102039	2	3	3	5	8	21
	Antilocapra americana americana	205804	1	4	4	2	1	12
	Antilocapra americana americana	A3447	1	2	4	1	1	9
Hindgut	Rhinoceros unicornis	464963	2	4	5	2	3	16
	Rhinoceros unicornis	545847	1	4	6	3	4	18
	Rhinoceros unicornis	540042	1	14	4	9	2	30
	Rhinoceros unicornis	336953	2	5	2	3	5	17
	Diceros bicornis	199708	8	13	9	3	0	33
	Diceros bicornis	271189	1	9	5	4	0	19
	Tapirus terrestris	261025	3	5	5	3	6	22
	Tapirus terrestris	292150	2	6	6	4	0	18
	Tapirus terrestris	218778	1	4	2	3	4	14
	Lepus americanus americanus	179425	1	11	13	8	0	33
Carnivore	Chrysocyon brachyurus	271567	2	4	5	4	7	22
	Chrysocyon brachyurus	588223	2	8	5	8	2	25
	Chrysocyon brachyurus	588425	1	5	4	2	2	14
	Canis lupus baileyi	234499	2	5	5	4	3	19
	Canis lupus baileyi	529679	1	3	7	1	1	13
	Canis lupus baileyi	529682	2	3	4	7	3	19
	Panthera leo	A22705	2	6	10	5	6	29
	Panthera leo	A12319	3	2	5	3	3	16

Note: Magnification ranges are categorized as follows: 0–50x, 51–500x, 501–2000x, 2001–5000x, and 5001x+. Each value represents the number of SEM images captured for each catalog number at the specified magnification scale.

Figure 1. Example of images taken under the same conditions with different contrast settings



Note: Both left and right images are from Catalog 179425, Lepus americanus americanus.

Table 2. Distribution of dietary type labels in SEM images

Old Label	New Label	Number of Images	Percentage
Carnivore	Carnivore	157	27%
Foregut	Herbivore	195	34%
Hindgut	Herbivore	220	39%

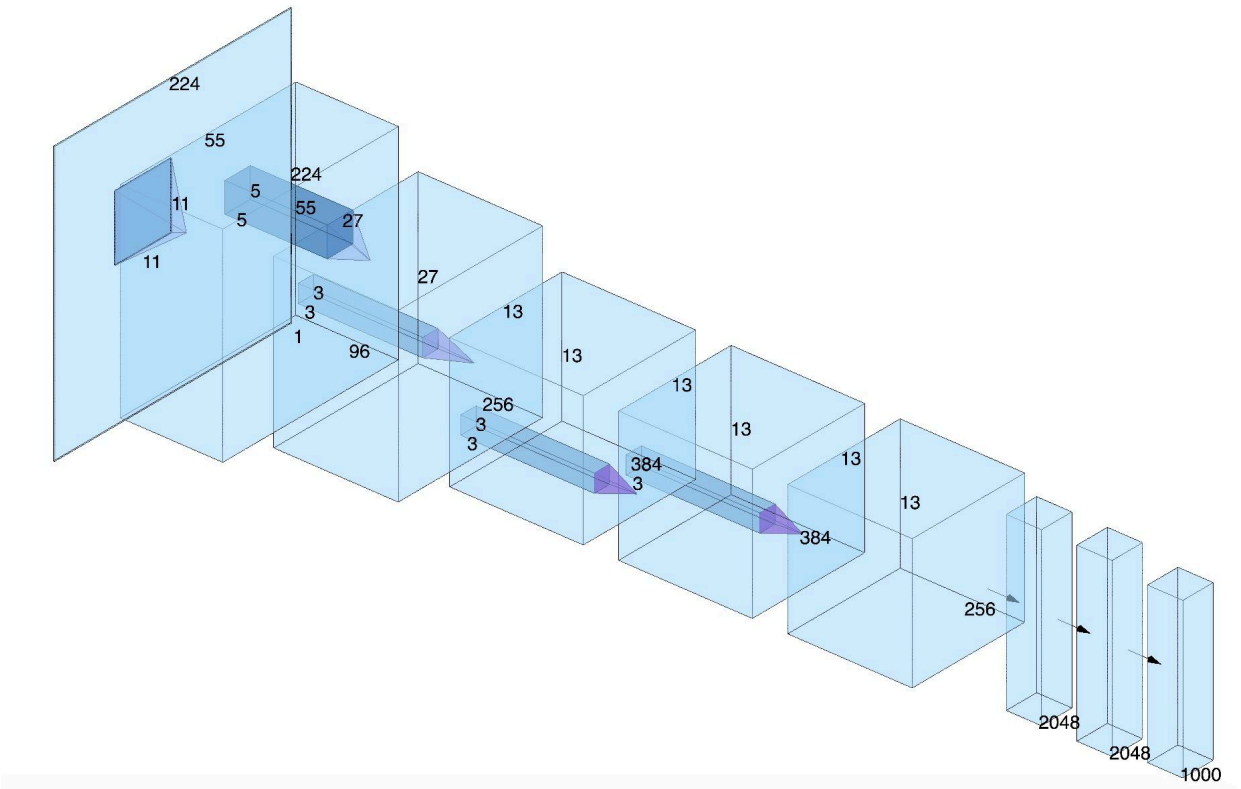
Note: 'Old Label' information is derived from Table 1 in the original dataset.

Table 3. Distribution of species labels in SEM images

Species	Number of Images	Percentage
Antilocapra americana americana	21	4%
Bison bison	55	10%
Canis lupus baileyi	51	9%
Chrysocyon brachyurus	61	11%
Diceros bicornis	52	9%
Lepus americanus americanus	33	6%
Ovis canadensis canadensis	50	9%
Ovis canadensis cremnobates	57	10%
Ovis canadensis mexicana	12	2%
Panthera leo	45	8%
Rhinoceros unicornis	81	14%
Tapirus terrestris	54	9%

Note: Species information is derived from Table 1. The total number of images is 572.

Figure 2. Illustration of AlexNet Architecture in this study



Note: AlexNet was originally designed for RGB images with three color channels (red, green, and blue). For this study, it has been adapted to work with grayscale images, which have a single color channel.

Figure 3. K-Means Clustering Results with PCA Dimensionality Reduction

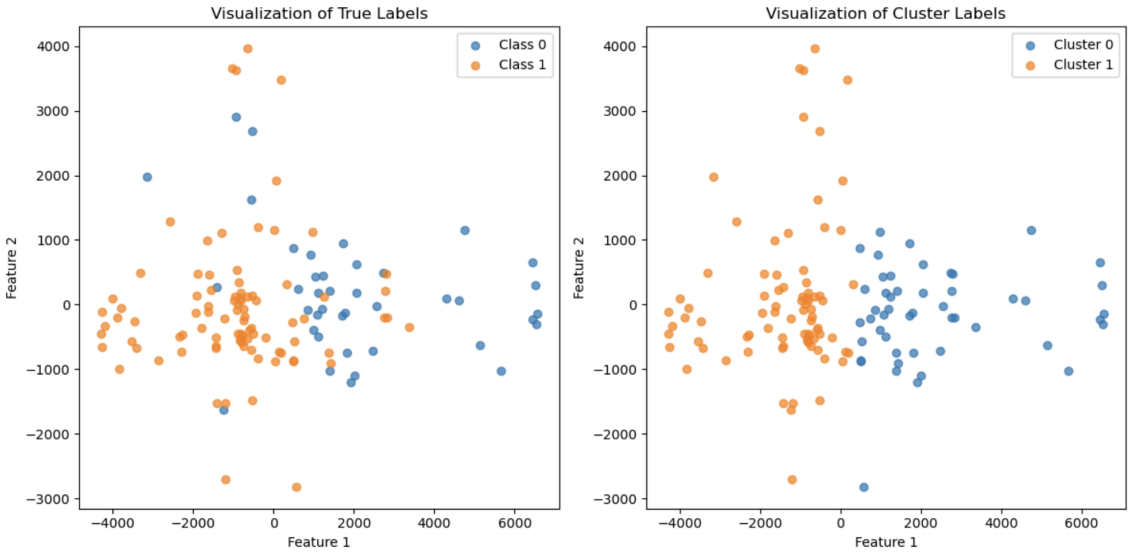


Table 4. Performance Metrics of Logistic Regression for Classifying Carnivores and Herbivores

	precision	recall	f1-score	support
0	0.60	0.38	0.46	8
1	0.76	0.89	0.82	18
accuracy			0.73	26

Note: The logistic regression model achieved a precision of 0.60 for carnivores (Class 0), meaning 60% of the samples predicted as carnivores were correct. The recall for Class 0 was 0.38, indicating that only 38% of the actual carnivore samples were correctly identified, resulting in an F1-score of 0.46. For herbivores (Class 1), the model achieved a higher precision of 0.76 and a recall of 0.89, correctly identifying 89% of the actual herbivore samples, with an F1-score of 0.82. The total support included 8 samples for Class 0 and 18 samples for Class 1, leading to an overall model accuracy.

Figure 4. Confusion Matrix (Left) and ROC Curve (Right) for Logistic Regression

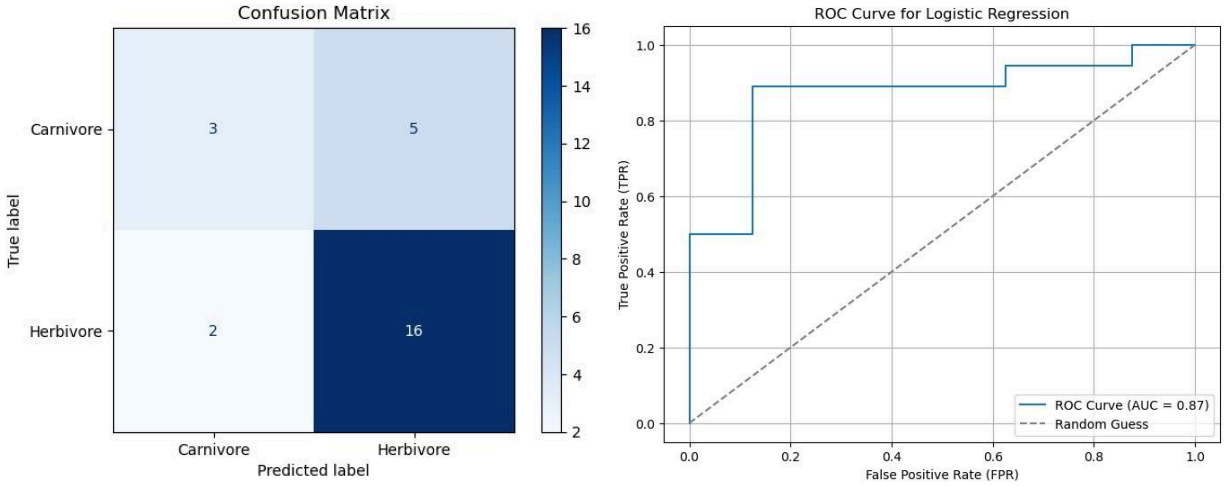
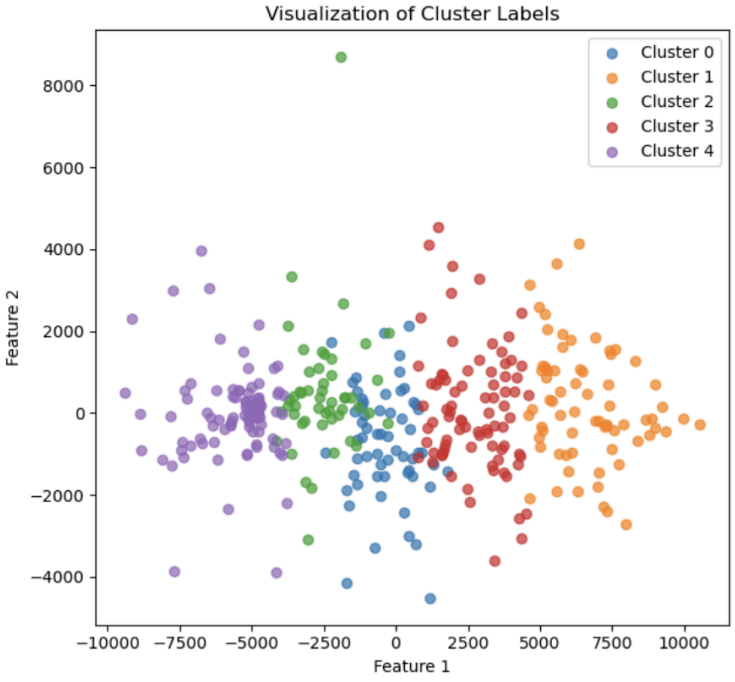


Figure 5. Clustering Results for Morphological Groupings



Appendix

Figure 1. Confusion Matrix of the AlexNet Model with Binary Class Labels

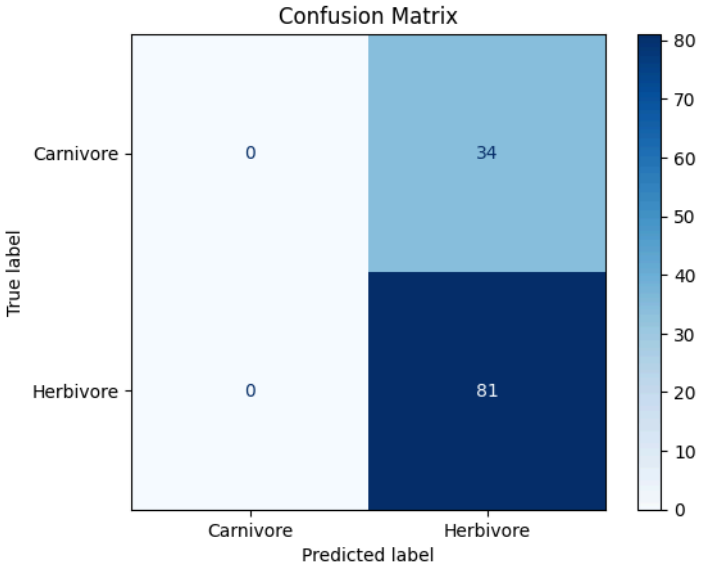


Figure 2. Confusion Matrix of the AlexNet Model with Multi-Class Labels

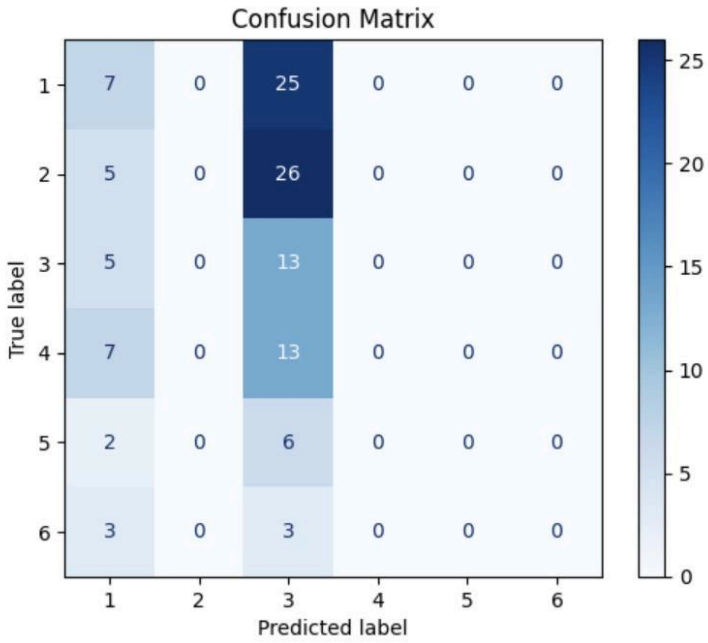


Figure 3. DBSCAN Clustering Results with PCA Dimensionality Reduction

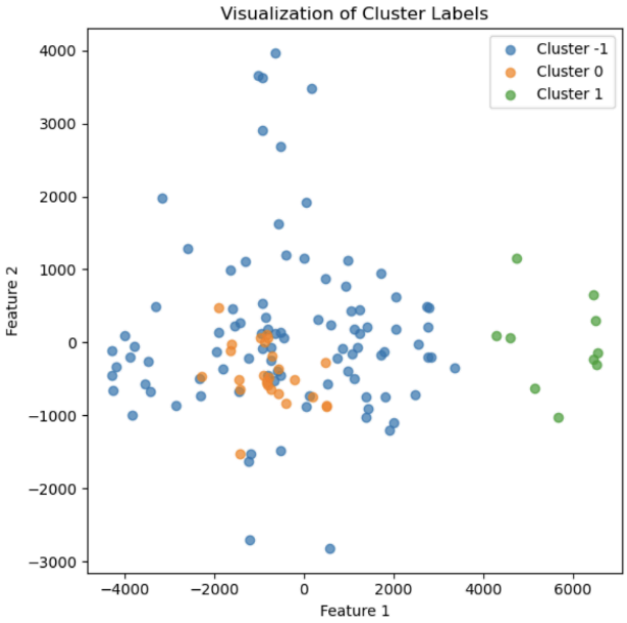


Figure 4. Agglomerative Clustering Results with PCA Dimensionality Reduction