Data Science Capstone: Bee Species Identification

Cheadle Center for Biodiversity and Ecological Restoration

Explorations of methods for quantifying differences in wing morphology

Patrick Moon and Jennifer Rink

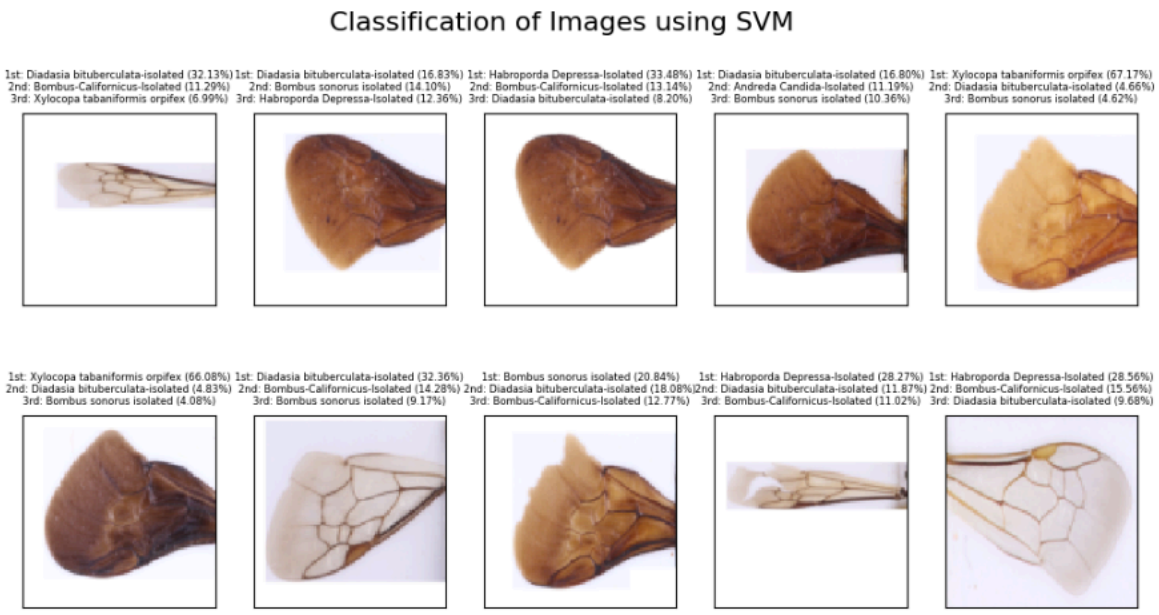

Fig 1. Example of Support Vector Machine output using Spectral Embedding

Introduction

In this report, we explore the significance of differences in wing morphology for bee species identification. Leveraging advanced data science techniques such as Linear Discriminant Analysis (LDA) and Spectral Embedding, our capstone project aims to unravel the unique characteristics within wing-vein structures.

Linear Discriminant Analysis (LDA)

This methodology begins with the extraction of high-level features from images of bee wings using the VGG16 neural network model, which is particularly adept at capturing the intricate patterns in the wing images. For each bee wing image, the model processes the input to generate a set of feature vectors, excluding the top fully connected layers to retain only the most relevant spatial hierarchies for classification tasks.

Upon gathering and preprocessing the data, Linear Discriminant Analysis is applied. LDA serves as a technique to reduce the dimensionality of feature sets while preserving as much class discriminatory information as possible. It works by finding a new axis that maximizes the separation between multiple classes[1].

The transformed features are visualized in a scatter plot, with each point representing a bee wing image, colored and labeled according to its species. This visualization confirms the effectiveness of LDA in distinguishing between species, and it also highlights the natural clustering of similar species, as indicated by their proximity in the plot.

To further our analysis, centroids of each species cluster are calculated, representing the geometric mean of all points belonging to a species in the reduced feature space. These centroids will be used for measuring the Euclidean distances between species, providing a quantitative measure of difference between any two species.

This distance metric enhances our understanding of wing morphology by quantifying how distinct the species are, based on their wing morphology, captured through deep learning and refined through linear discriminant analysis.

[1] Kavlakoglu, Eda. *Implementing Linear Discriminant Analysis (LDA) in Python*, 18 Mar. 2024, developer.ibm.com/tutorials/awb-implementing-linear-discriminant-analysis-python/.

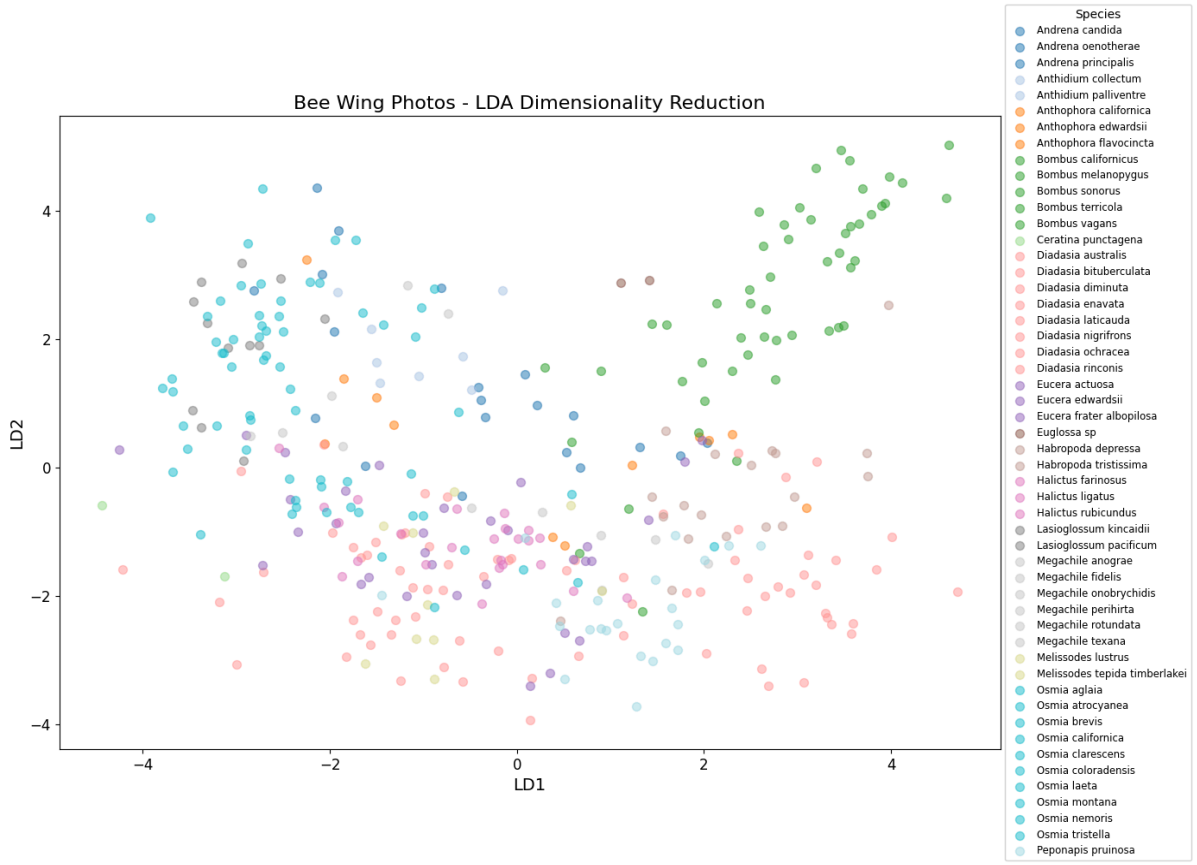

Fig 2. LDA graph visually depicting quantified differences between species

Spectral Embedding

Discovering patterns and connections is an essential part of our exploration in quantifying differences in wing morphology, and the utilization of spectral embedding shows promise in aiding this problem.

Spectral Embedding is a dimensionality reduction technique that is frequently applied in data analysis. It shares a close relationship with Principal Component Analysis (PCA), but unlike PCA, spectral embedding can be used to analyze high dimensional, non-linear data structures.

Spectral Embedding can provide a graphical representation of the similarities between images. Using the scikit-learn library in Python, we were able to use the Spectral Embedding function alongside a Support Vector Machine for classification to perform our analysis on the cropped wing images. In an article analyzing the MNIST dataset using similar techniques[2] , a linear kernel was used in the support vector machine because the data could be separated by a hyperplane. In our case, the data has complex multi-dimensional features that cannot be separated linearly, so we employ a Radial Basis Function (RBF) kernel in our support vector machine. RBF kernels are more flexible and can capture complex relationships between features, making them suitable for a wide range of classification tasks, including those with overlapping classes or nonlinear decision boundaries[3].

However, through our code, we found that spectral embedding does not seem to capture the complex structures of the patterns in the wing veins as efficiently and accurately as LDA does.

[2] Pewsey, Matt. "MNIST Handwritten Digit Classification." *Matt Pewsey*, 28 Sept. 2021, mpewsey.github.io/2021/09/28/mnist-handwritten-digit-classification.html.

[3] B. -C. Kuo, H. -H. Ho, C. -H. Li, C. -C. Hung and J. -S. Taur, "A Kernel-Based Feature Selection Method for SVM With RBF Kernel for Hyperspectral Image Classification," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 1, pp. 317-326, Jan. 2014, doi: 10.1109/JSTARS.2013.2262926. keywords: {Support vector machines;Kernel;Training;Hyperspectral imaging;Educational institutions;Feature extraction;Feature selection;hyperspectral image classification;kernel-based feature selection;radial basis function;support vector machines}

Fine tuning the Support Vector Machine could potentially present more accurate results, but with a simpler solution like LDA, it may not be necessary to expend effort trying to optimize the SVM from Spectral Embedding. Below are descriptions of each parameter that can be altered.

1. **Kernel Type:**

   kernel='poly': Polynomial kernel with parameters like degree and coef0.
   kernel='rbf': Radial basis function (RBF) kernel with parameter gamma.
   kernel='sigmoid': Sigmoid kernel with parameters gamma and coef0.

2. **Regularization Parameter (C):**
   Adjust the regularization parameter C to control the trade-off between maximizing the margin and minimizing the classification error.
   Higher values of C penalize misclassification more severely, which can lead to overfitting. Conversely, lower values of C encourage a wider margin but may increase bias.
3. **Gamma Parameter (RBF Kernel):**

   If using an RBF kernel like we did initially, tune the gamma parameter, which determines the influence of each training example. Higher values of gamma lead to more complex decision boundaries, potentially causing overfitting. Lower values of gamma result in smoother decision boundaries, which may improve generalization.

In conclusion, LDA provides a more intuitive and accurate visualization of the differences in species than Spectral Embedding. However, our most successful exploration of differences in wing morphology was our landmark model that provides a very accurate and interesting decision tree.

Find it here on the complete GitHub: https://github.com/PSTAT197-F23/capstone-CCBER

References

B. -C. Kuo, H. -H. Ho, C. -H. Li, C. -C. Hung and J. -S. Taur, "A Kernel-Based Feature Selection Method for SVM With RBF Kernel for Hyperspectral Image Classification," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 1, pp. 317-326, Jan. 2014, doi: 10.1109/JSTARS.2013.2262926. keywords: {Support vector machines;Kernel;Training;Hyperspectral imaging;Educational institutions;Feature extraction;Feature selection;hyperspectral image classification;kernel-based feature selection;radial basis function;support vector machines}

Kavlakoglu, Eda. *Implementing Linear Discriminant Analysis (LDA) in Python*, 18 Mar. 2024, developer.ibm.com/tutorials/awb-implementing-linear-discriminant-analysis-python/.

Pewsey, Matt. "MNIST Handwritten Digit Classification." *Matt Pewsey*, 28 Sept. 2021, mpewsey.github.io/2021/09/28/mnist-handwritten-digit-classification.html.