Developing Tools for Species-Level Identification: Computer Vision Enables Rapid Identification of Wild Bee Species Using Wing Venation

By

Jahir Morris



Advisor: Dr. Sarah D. Kocher

Graduate Student Advisor: Michelle White

April 29th, 2024

In partial fulfillment of the requirements for a bachelor of arts degree in Ecology and Evolutionary Biology at Princeton University.

Contents

Acknowledgements **3**
Abstract **5**
Introduction **6**
I. Halictidae 7
II. Bee Classification by Wing Venation 10
III. Machine Learning Techniques 11
IV. Objectives 12
Materials and Methods **13**
I. Photographing Specimens 13
II. Image Processing and Annotation 15
III. Feature Extraction 16
IV. Feature Selection 18
V. Classification Scheme 22
a. Support Vector Machines 22
i. Training I 23
ii. Training II 23
b. Random Forest Classifier 24
VI. Validation 25
VII. Optimization 27
a. SVM Hyperparameters 27
i. “C” 27
ii. Gamma 28
iii. kernel 28
b. RFC Hyperparameters 29
i. N estimators 29
ii. Bootstrap 29
iii. Criterion 30
iv. Max Features 31
v. Max Depth 31
I. How to Interpret Results 32
a. Confusion Matrix 32
b. Performance Metrics 33
c. Precision, Recall, and F1-Score 33
VIII. Limitations 34
a. Data Collection 34
b. Species Representation and Feature Extraction 35
Results **37**
I. Biologically Significant Features 37
II. Classification Accuracy 40
III. Misclassifications 42
IV. Final Application Design 43
Results Summarized **44**
Discussion **45**
Future Work **48**
I. Improvements 48
II. Applications 49
Appendix **51**
References **52**

Acknowledgements

I would like to express my sincere gratitude to the following people and organizations who, through their meaningful contributions, have empowered me to complete this body of work:

For the *High Meadows Environmental Institute*, the *Department of Ecology and Evolutionary Biology* at Princeton, and the *Office of Undergraduate Research* for funding this project.

For *Miriam Richards and the Brock Bee Lab* at Brock University for supplying several of the specimens used in this study.

For *Christopher Lawrence* at the Rubenstein Lab for your mentorship and guidance, especially in learning ML-morph.

For my faculty advisor, *Dr. Sarah Kocher*, for being so encouraging and knowledgeable throughout the entirety of this process. Even when I didn't reach for months at a time, you always made sure that I had all the necessary tools and information to successfully complete our study. Your enthusiastic guidance made even the most difficult moments during the research process seem surmountable. I'm grateful to be able to call you my advisor.

For my graduate student advisor, *Michelle White*. I cannot thank you enough for the amount of time, effort, and energy that you invested in me and this research project. It is your guidance that allowed me to follow this project through to the end, despite how limited my knowledge of machine learning was when we began. A simple 'thank you' does not even begin to capture the amount of gratitude I have for you. Any undergraduate student would be lucky to have you as a mentor.

For my freshman-year high school biology teacher and science research coordinator, *Mrs. Catherine Kenny*. Never would I have guessed that stepping into your classroom on my first day of school would've marked the beginning of such a formative connection. It is through your grace and mentorship that I've been able to pursue an academic career in ecology at Princeton. Your unwavering support through the years has kept me grounded as I continue to push new boundaries and achieve my goals. Thank you for encouraging me to grow.

For my family,

For my parents, *Chriscita and Audley Morris*, for your continued love, support, and guidance over the last twenty-one years of my life. There is something terrifyingly beautiful about a youngest child growing into his relationship with his parents. Every step in this process may not be pretty but each one is just as necessary as the last. Let all the tears and laughter we've shared between the day I was born till now be a testament to the undying love and gratitude that I will always feel for you both. Thank you for everything.

For my brother *Jalil Morris* and my sister *Tyana (Michelle Marshall)*, for being the best siblings a younger brother could ask for. There are no words to express how grateful I am to have the both of you at my side from the very beginning. From late night pokémon sessions to ziplining through the mountains of South Africa, there isn't a single moment I've shared with either of you that didn't make me feel secure, loved, and confident. I wouldn't be who I am today without you both. Thank you for being everything I've ever needed, and so much more.

For my sister, *Sheba*, for your unconditional love and affection from the moment we brought you home, to the day you left this Earth. I am choking back tears as I write this letter, knowing that this may be the closest I'll ever get to saying goodbye. Just know that if there is not a single person left on this Earth who is thinking of you, then I must have come to join you in the afterlife. Thank you for the hugs, kisses, and the many memories which I'll never forget.

For my chosen family,

For my best friend, conscience, and partner in crime, *Nigel Munroe*. I needn't make this section too long because there is only so much left unsaid between us. Despite everything we've been through, you've always been a constant source of love, reassurance, understanding, and support. I am truly lucky to be able to call someone as inspiring as you, my best friend.

For my friends who made Princeton bearable. For the day-ones, twos, and threes, all of whom I couldn't possibly name here. For the *Tems Fanclub*, the *Physics Crew*, *Margarita Mondays*, the *Maybach*, '*Afro-scullians*', *Songline Slammers*, and the *Black Arts Collective*. The life of a Princeton undergraduate has never been an easy one, but the experiences I've been blessed enough to share with you all have made my time at this university unforgettable. Thank you for your loving friendship.

Abstract

Through their role as pollinators, bees continue to be major contributors to the survival and productivity of both native plant communities and the agricultural industry. However, in spite of their ecological and economic importance, estimates of global bee populations have demonstrated rapid declines in species richness and density over the past century. Although these estimates provide information on general population trends, many of them lack specificity and offer little insight into dynamics occurring at the species-level. Additionally, the process of determining species-level identifications for bees relies heavily upon a shrinking number of expert taxonomists who are capable of recognizing subtle morphological differences between species. For these reasons, monitoring programs and large-scale conservation for insect pollinators has been slow, costly, and relatively ineffective. This study proposes the creation of a publicly available database of wing-images for the most common bee species found across New Jersey, US, and the creation of an application which will use these images to deduce accurate species-level identifications of insect pollinators. With the implementation of deep learning frameworks, an application for automatic identification of bee pollinators can be made available for use by researchers, students, and local communities. By generating a publicly available database of forewing images, future users will be able to expand the training set of our model to include new species, and ultimately increase the repertoire of species for which it can produce predictions. Through the integration of recent advances in machine learning technology and citizen science approaches to data collection, this project has the potential to both streamline the pollinator identification process and create opportunities for community outreach and global participation in efforts to conserve remaining populations of bees.

Introduction

Through their role as pollinators, bees continue to be essential drivers of survival and productivity for both native plant communities and the agricultural industry worldwide (Woodard et al. 2020). Insect pollination (carried out primarily by bees) alone is required for the growth of approximately 90% of wild plants (Burd & Kopec 2017) and nearly 75% of crops used directly for human consumption (Potts et al. 2010). Native pollinators (i.e. bumblebees and solitary wild bees) provide essential reproductive services for wild as well as cultivated plants in virtually all terrestrial ecosystems (Ayral et al. 2020). Even among other biotic pollinating agents, bees are the most effective because of their high flower reliability and flower constancy (Roubik 1995; Garibaldi et al. 2013; as cited in Matias et al. 2016). This means that bees actively seek out specific flowers and are able to identify their preferred species even in highly diverse settings. Many flowering plants have even evolved concurrently with their respective pollinators, resulting in highly specialized relationships between a single plant species and (in many cases) an individual bee (Shimizu et al. 2014).

Previous studies have proven that the preservation of a diverse community of wild bee species is necessary for maintaining high levels of ecosystem function within large areas of habitat (Winfree et al. 2018). High pollinator diversity ensures that in the inevitable aftermath of species turnover between habitats, vulnerable ecosystems possess an ample number of remaining species which can fill vacant niches.

However, in light of the mounting evidence behind their ecological significance, estimates show that global bee populations have experienced rapid declines in both species richness and density over the past century (Biesmeijer et al. 2006). These population trends have largely been attributed to forces such as habitat loss, disease, and pesticide use (Lebuhn & Luna

2021). However, many of these estimates lack specificity and offer little to no insight into population shifts occurring at the species-level. The large majority of estimates taken on bee population trends group this highly diverse family of insects into broad geographical regions, degree of sociality, or state of captivity (captive colonies v. wild populations) (Koh 2015). This leaves little room to draw conclusions on the state of particular species, and thus, determine the potential causes of their observed decline. Consequently, initiatives aimed at monitoring bee populations and assessing biodiversity have been slow and often confounded by the practical difficulty associated with determining accurate taxonomic identifications of individuals (Portman et. al 2020).

I. Halictidae

Some of the most abundant insect pollinators for a variety of wild plants and crops in North America belong to the family Halictidae, commonly referred to as sweat bees (Landaverde-Gonzáles 2017). Alone this family consists of 4,500 species and their role as pollinators is of great importance, especially considering that the most dominant crop-pollinating species are also the most widespread species in agricultural landscapes, and that more than half of North American and Hawaiian bee species are in decline (Kopec 2017). Sweat bees are generalist species, meaning that their populations can survive in a wide range of environmental conditions and take advantage of an even wider range of resources. As such, they play a large role in pollinating plant communities in a variety of environments ranging from tropical agricultural landscapes to densely populated urban centers (Mason & Arathi 2019; Landaverde-Gonzáles 2017). Despite how widespread and abundant halictid bees are, species of this highly diverse family of insects are notoriously difficult to differentiate from one another,

even among experienced taxonomists (Mason & Arathi 2019). This has had the effect of drastically limiting the amount of reliable data we possess on species-specific halictid bee population dynamics. This lack of information could have severe consequences in localities where populations of sweat bees, which may be the dominant contributing pollinator to the entire plant community, are experiencing rapid declines (Rodrigo Goméz et al. 2016). Thus, sweat bees represent an appropriate family of study species from which to begin the development of our pollinator identification tool.

Currently, accurate species-level identification of a specimen is dependent upon the time and knowledge provided by a small group of expert taxonomists (Portman et. al 2021). Using subtle morphological differences such as those observed in head length or thorax and abdomen structure, these taxonomists are capable of differentiating between species. Though effective, these methods involve an expensive and time-consuming process of specimen collection, preparation, and examination, as well as demand highly specialized knowledge on several bee species (a knowledge which most taxonomists possess for only a small number of species). This means that more often than not, large-scale monitoring programs will require input from multiple individuals. Additionally, collection and preparation methods for insect pollinators are often invasive and lethal to the specimen (Prendergast 2020). This greatly increases the possibility of restriction and condemnation of monitoring efforts, especially when studying threatened or sensitive species. Thus, there is a great need for alternative methods to produce accurate species-level identification of pollinators.

Recent literature has revealed the potential for automated approaches to bypass the aforementioned obstacles to producing accurate species-level identifications (Spiesman et al. 2021). Many of these approaches are designed to equip non-experts with the necessary tools to

identify several species with notably less investment of both time and financial capital (Milam et al. 2020). In doing so, they also present alternative data collection methods that are considerably less harmful to the study system. New DNA barcoding technologies, for instance, have allowed for nearly any research group to reliably distinguish between even the most cryptic species through the use of genetic barcoding of the cytochrome c oxidase subunit I (COI) (Yang et al. 2018). COI is a gene that is present in most eukaryotes and possesses a slow enough mutation rate so that the gene is highly conserved *within* species, but rapid enough to be distinguishable *between* species, thus making it an ideal gene for DNA barcoding. As a result of these advancements, even the least experienced taxonomists are able to bypass the arduous process of differentiating between minute morphological traits and produce species-level identifications with considerable accuracy. However, these methods still require access to a functioning genomics lab with DNA extraction and polymerase chain reaction (PCR) and sequencing technology. These processes can typically be expensive, time consuming, and difficult to conduct on live specimens. While this still represents an obstacle to creating an accessible system for identifying pollinators, recent efforts to collect and compile DNA barcodes on behalf of most extant species have hinted at the potential for automated computer systems to further streamline the process of identification (Milam et al. 2020). Recent advancements in machine vision and natural variation in wing venation between insect species have revealed a new opportunity to further improve existing methods for species-level classification (Buschbacher et al. 2019).

This project aims to develop the necessary tools for image-based recognition and identification of pollinators by species, and ultimately use these tools to promote increased specificity and accuracy in monitoring programs designed to aid in their conservation.

II. Bee Classification by Wing Venation

Wing venation has long been used as an entomological tool for distinguishing between insects, including bees, at virtually all taxonomic levels (Michener 2007, as cited in Hall 2011). Bees tend to vary substantially in their wing venation patterns. Some morphological features, such as the number of submarginal cells, even vary across genera (Karunaratne 2008). Even a subset of the wing venation patterns, such as those outlining the inner cells, has been proven sufficient to produce accurate species-level identifications (Steinhage et. al 2006). A multitude of morphometric approaches such as: landmarking coordinates, wing cell geometry, and distance ratios, all can be applied directly to wing venation patterns to extract taxonomically distinct characteristics down to the species-level. Although these characteristics are largely indistinguishable at a glance, computer vision possesses the unique ability to quantify subtle morphological features and rapidly parse through the results to produce accurate identifications. Recent studies have also shown that the above morphometric approaches possess a similar capacity to distinguish between bee species as some genomic methodologies (Oleksa and Tofilski 2014). Considering that access to the facilities and genotyping technologies necessary for molecular identification remain limited and relatively costly, bee identification by wing venation has become a viable alternative to the use of modern genomic approaches, especially where expert taxonomists are not readily available.

Currently, wing measurements are almost exclusively derived from digital images. Thus, following the development of the initial machine learning pipeline, species identification by way of wing venation is only dependent upon using a camera system and the resulting images. Camera systems ranging from those belonging to smartphones to digital cameras mounted on a microscope (such as the one used for this study), have proven capable of capturing images with

sufficient detail to satisfy the algorithm's minimum requirements to produce accurate identifications (Stefan et al. 2024, Arbuckle et al. 2001).

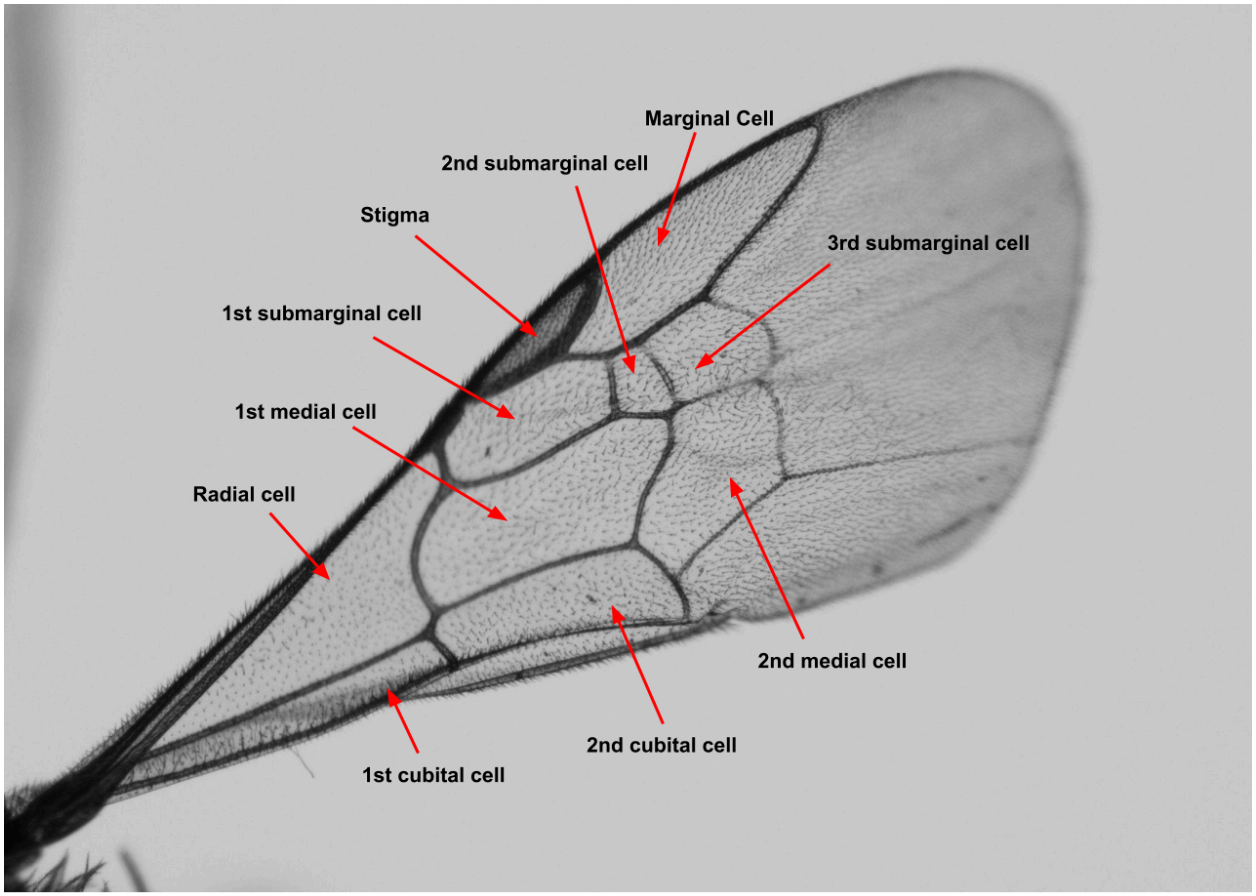


Figure 1: Diagram of halictid forewing (example photo taken of *Lasioglossum atwoodi*). The wing veins create a pattern of cells that can be identified by their position relative to the others. All cells are labeled by their position and whether they are marginal, submarginal, medial, radial, or cubital.

III. Machine Learning Techniques

Identification systems using machine learning (ML) techniques such as convolutional neural networks have proven to be particularly effective at identifying plants and animals captured on camera trap footage (Wäldchen 2018; Norouzzadeh et al. 2018). Convolutional neural networks (CNN) are algorithms which take input images, assign quantitative values to various aspects of

that image (i.e., features), and then use those values to differentiate those images from one another. Previous studies have proven that these techniques can even surpass human capabilities in terms of classification accuracy (Buschbacher et al. 2019).

We will use the characteristics contained in images of sweat bee forewings to train our classification model because they have been proven to be sufficient when identifying individual species, whereas training on full body images has seen highly variable success rates and often have difficulty producing accurate identifications below genus-level (Stefan et al. 2024; Buschbacher et al. 2019). There are a number of existing frameworks for ML encoded in python which have been designed for the purpose of solving classification problems such as this (python v3.12.3). There is a growing body of research highlighting the potential of multinomial classification frameworks such as support vector machines and random forest classifiers to process high-dimensional image data and solve complex identification problems in biology (Yang 2004; Pal 2003).

Through the use of these frameworks and ML techniques, a comprehensive database of bee species can be created in an effort to develop an accessible and resource efficient approach to accurately monitoring - and ultimately conserving - dwindling communities of halictid pollinators.

IV. Objectives

Our study aims to first generate a dataset of wing images representing wild bee species native to New Jersey and the wider northeastern United States. This dataset should have sufficient

representation of each species so as to capture intraspecific variation among individuals. Secondly, we plan to identify a set of morphological features which are derived from the wing venation patterns of photographed specimens, which can be used to reliably distinguish between species. Lastly, our study aims to use these features to develop a classification model with the following qualities:

- Accurate (prediction accuracy above what can be expected of random chance)
- Reliable (i.e. precise in its identifications)
- Makes predictions based off of a single wing image
- Requires little to no user knowledge of bee morphology or computer programming
- Requires little to no user input (beyond submitting a wing image)
- Is accessible to those outside of academic researchers and melittologists

Materials and Methods

I. Photographing Specimens

Data collection for this study began in May 2023 and continued through August 2023. The dataset used in this study consists of 755 wing images, representing 16 different species of bees from across 4 genera found throughout New Jersey and the larger mid-atlantic region. We photographed 50 individuals per species (excluding *Lasioglossum paradmirandum*, *L. cinctipes*, for which we had only 30 and 29 respectively, as well as *L. ephialtum*, *Halictus confusus*, *H. rubicundus*, and *Augochlora pura*, for which we had 49 individuals each). This number of replicates is necessary in order to accurately capture the amount of individual variation within each of the species. Previous literature suggests that anywhere between 20 and 30 wing images

per species is sufficient to train an accurate identification model (Buschbacher et al. 2019). All individuals were sourced either from the collections at the Kocher Lab at Princeton University or the Brock Bee Lab at Brock University. Specimens supplied by these labs have been collected from a variety of habitats ranging from densely wooded forest to extensively developed agricultural fields and subsequently pinned, genotyped, barcoded, and labeled with species identifications by expert taxonomists.

Family	**Subfamily**	**Tribe**	**Genus**	**Species**	**Count**
Halictidae	Halictinae	Augochlorini	*Augochlorella*	*aurata*	50
Halictidae	Halictinae	Augochlorini	*Augochlora*	*pura*	49
Halictidae	Halictinae	Halictini	*Agapostemon*	*virescens*	50
Halictidae	Halictinae	Halictini	*Halictus*	*ligatus*	50
Halictidae	Halictinae	Halictini	*Halictus*	*rubicundus*	49
Halictidae	Halictinae	Halictini	*Halictus*	*confusus*	49
Halictidae	Halictinae	Halictini	*Lasioglossum*	*imitatum*	50
Halictidae	Halictinae	Halictini	*Lasioglossum*	*coreacium*	50
Halictidae	Halictinae	Halictini	*Lasioglossum*	*nymphaearum*	50
Halictidae	Halictinae	Halictini	*Lasioglossum*	*versatum*	50
Halictidae	Halictinae	Halictini	*Lasioglossum*	*paradmirandum*	30
Halictidae	Halictinae	Halictini	*Lasioglossum*	*cinctipes*	29
Halictidae	Halictinae	Halictini	*Lasioglossum (Dialictus)*	*ephialtum*	49
Halictidae	Halictinae	Halictini	*Lasioglossum (Dialictus)*	*admirandum*	50
Halictidae	Halictinae	Halictini	*Lasioglossum (Dialictus)*	*atwoodi*	50
Halictidae	Halictinae	Halictini	*Lasioglossum (Dialictus)*	*hitchensi*	50

Table 1: List of Species Included in Our Dataset of Wing Images.

We used a Nikon SMZ1270 digital microscope to photograph one of two forewings belonging to each specimen. No preference was given to either the right or left wing due to the lack observable difference between their respective morphologies. Although the background remained constant, images were taken in different lighting conditions, zoom, and resolution, so as to maximize visibility of the venation pattern of each distinct wing. We used sub-stage lighting in order to eliminate reflection of light off the wings. This allowed for faint wing venation patterns to be clearly visible within each image.

We converted all images to grayscale and made increases to image contrast and sharpness when necessary to better distinguish veins. All necessary adjustments were made directly through the Nikon microscope's user interface and required no post-hoc processing.

II. Image Processing and Annotation

Using imglab software, we landmarked the wing images using both bounding box and individual XY landmarking methods. Although the initial dataset was annotated manually, the bee species identification pipeline uses ML-morph, a supervised learning-based phenotyping framework which employs object detection and shape prediction models to automatically place landmarks on never-before-seen images (Porto and Voje 2020). By outlining the parameter of the wing into what's referred to as a bounding box, we indicate what content from the image is of interest for the object detector model. Then, by placing an XY coordinate point on each vein joint, making a total of 16 coordinate points per wing image, we dictated what specific components of the object the shape prediction model will recognize and ultimately landmark. An .xml file composed of

these coordinates was downloaded for each wing image to be used when training the object classification model for the species identification pipeline.

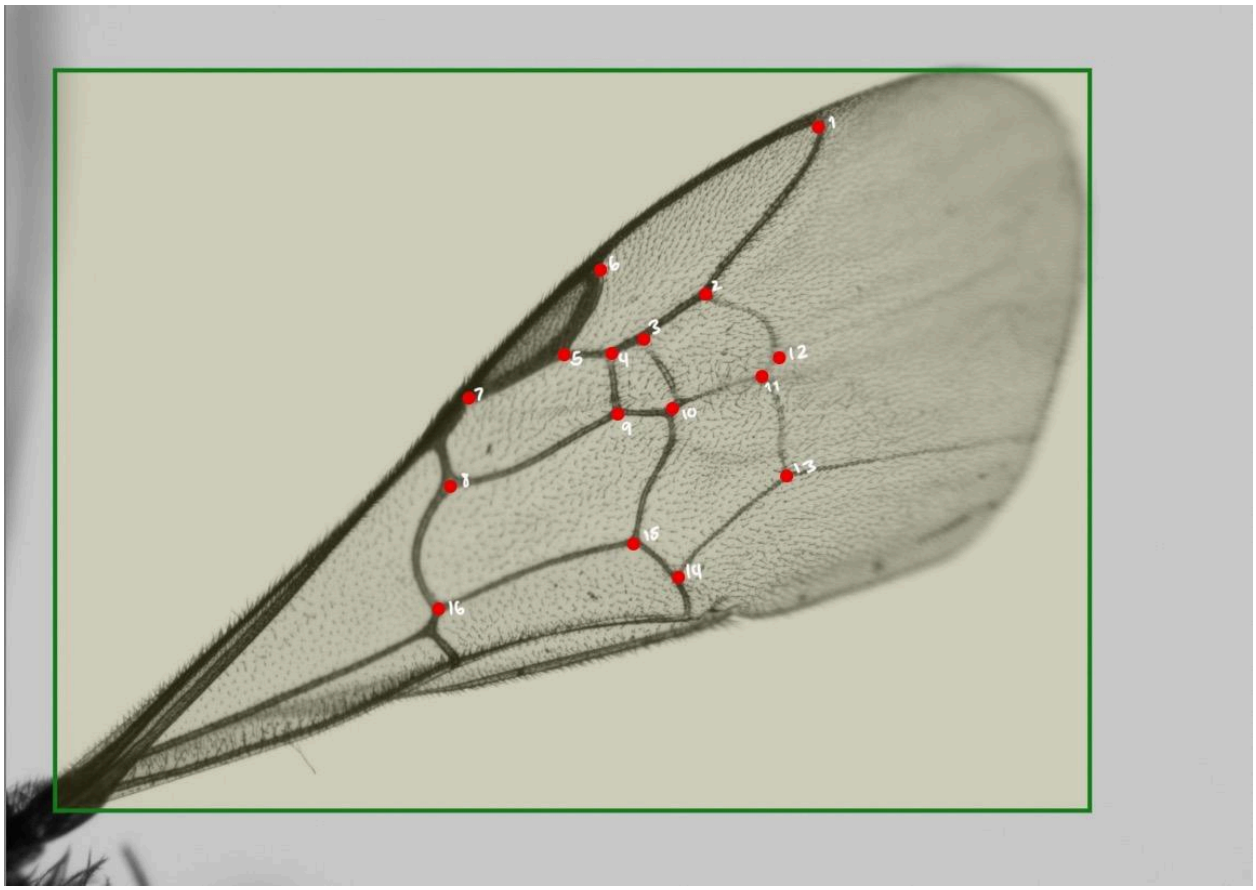


Figure 2: *Lasioglossum atwoodi* forewing with labeled vein joints from Imglab. We placed 16 coordinate points on each image to represent the positions of vein joints.

III. Feature Extraction

In order for any classification model to produce accurate identifications, it's necessary that the model be based on a high quality set of quantifiable features which are derived from the images. Ideal characteristics for a high quality feature include: the ability to be measured easily from the source data (in this case, XY coordinate points), can have its measurements replicated across all

images in the dataset, and there is a reliable distinction between the taxonomic units of interest (species) based on that feature.

Since image preprocessing and annotation of the vein joints have been completed, extracting features from the resulting coordinate points can be accomplished through a number of methods. Using wing venation patterns as the source for our data points means that our dataset is representative of a collection of wing cells (polygons) and their respective vein joints (vertices). Consequently, we considered the following methods to quantify their geometries into features:

- Area of interior cells
- Perimeter of interior cells
- Cross-sectional length of interior cells

These measurements alone could be sufficient for the classification model. However, this assumes that the images were all taken at standard conditions with identical zoom and orientation (therefore making the pixels in each image equivalent in terms of distance measurements). Because these conditions were not met for this study, it was necessary that we take the ratio of the above morphological measurements to one another, and use those values as the features from which our species identifications are derived.

We used the following formula to calculate the distances between all pairwise combinations of each labeled vein joint:

$$d = \sqrt{(x_2 - x_1) + (y_2 - y_1)}$$

We followed this by calculating the ratio of all of these distances to one another, and appending those values to our original XY feature set. With our distances and distance ratios calculated, we have a total of 7,260 features.

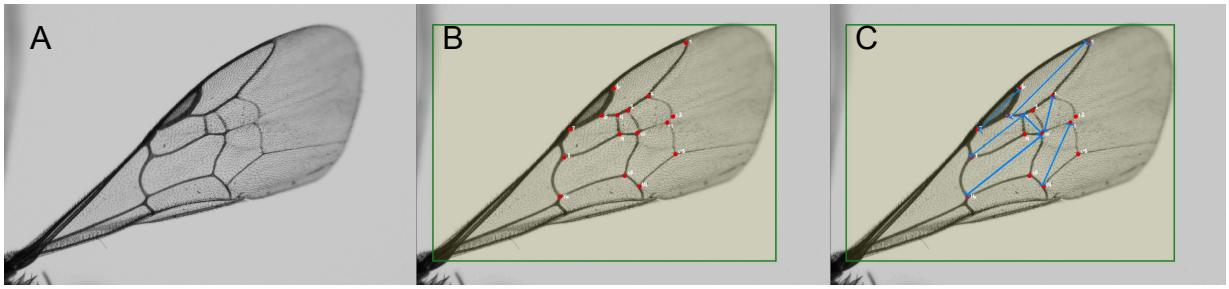


Figure 3: (*Lasioglossum atwoodi*) Wing Image Annotation and Feature Extraction Procedure. Fig.3A shows an image captured by the SMZ1270 microscope. Fig.3B shows that same image with landmarked vein joints (placed in sequence from 1 to 16). Fig.3C shows a subset of morphological measurements that we calculated from the landmarks.

IV. Feature Selection

Before building our model, we needed to determine which features out of those we've extracted are useful for making species classifications. The first step in this process requires that we determine how correlated, if at all, each of our features are to one another. Ideally, we desire our features to be highly correlated to the target classes (i.e. species). This would imply that those features can reliably dictate which species of bee a wing image must belong to. However, if those features are highly correlated with one another, this implies that they have comparable ability to discriminate between species. This makes it pointless to use any more than one of those features to train the classifier, and may even be harmful to the model's overall performance. For this reason, we created a correlation matrix to visualize the extent to which each of our features were correlated with one another (see Figure 3). Based on the results of our matrix, our features were largely uncorrelated.

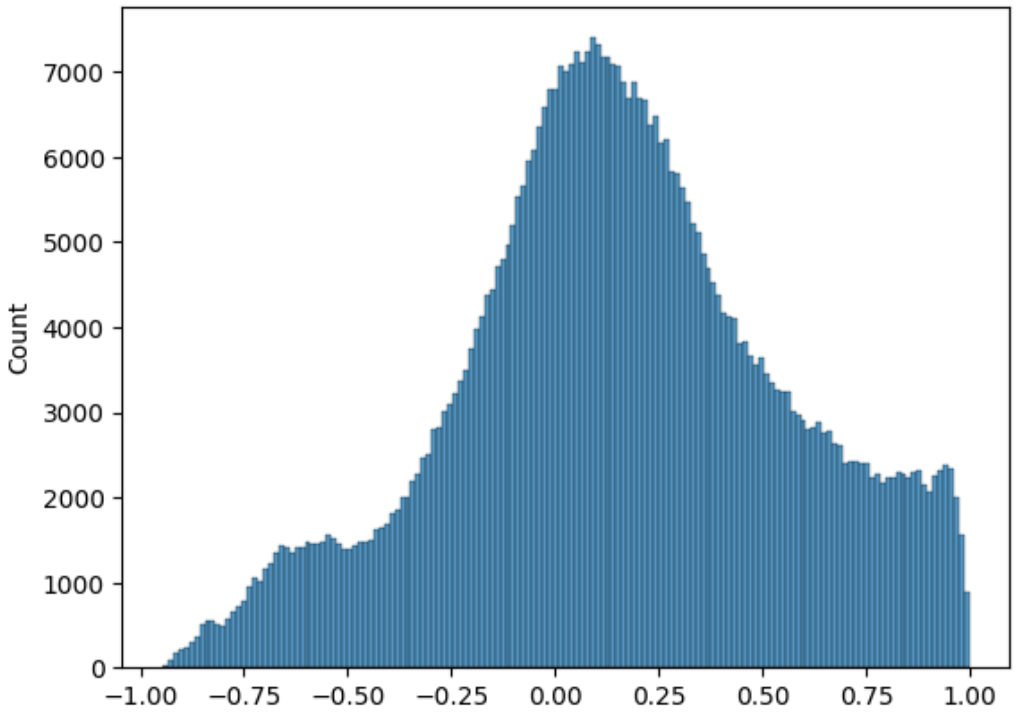


Figure 4: Plot of Pairwise Correlations Between All Features. This distribution is approximately normal, therefore ~96% of features have low correlations.

Afterwards, we ran a principal components analysis (PCA) in order to determine which and to what extent features most effectively distinguish between species. A PCA is a dimensionality reduction tool which effectively groups individual data points into feature-based categories. In an effort to reduce computation time, we ran our analysis with eight components derived from the first 250 features in our list of previously extracted distance ratios. As a result, the PCA plots only represent six of the 16 total species from this study.

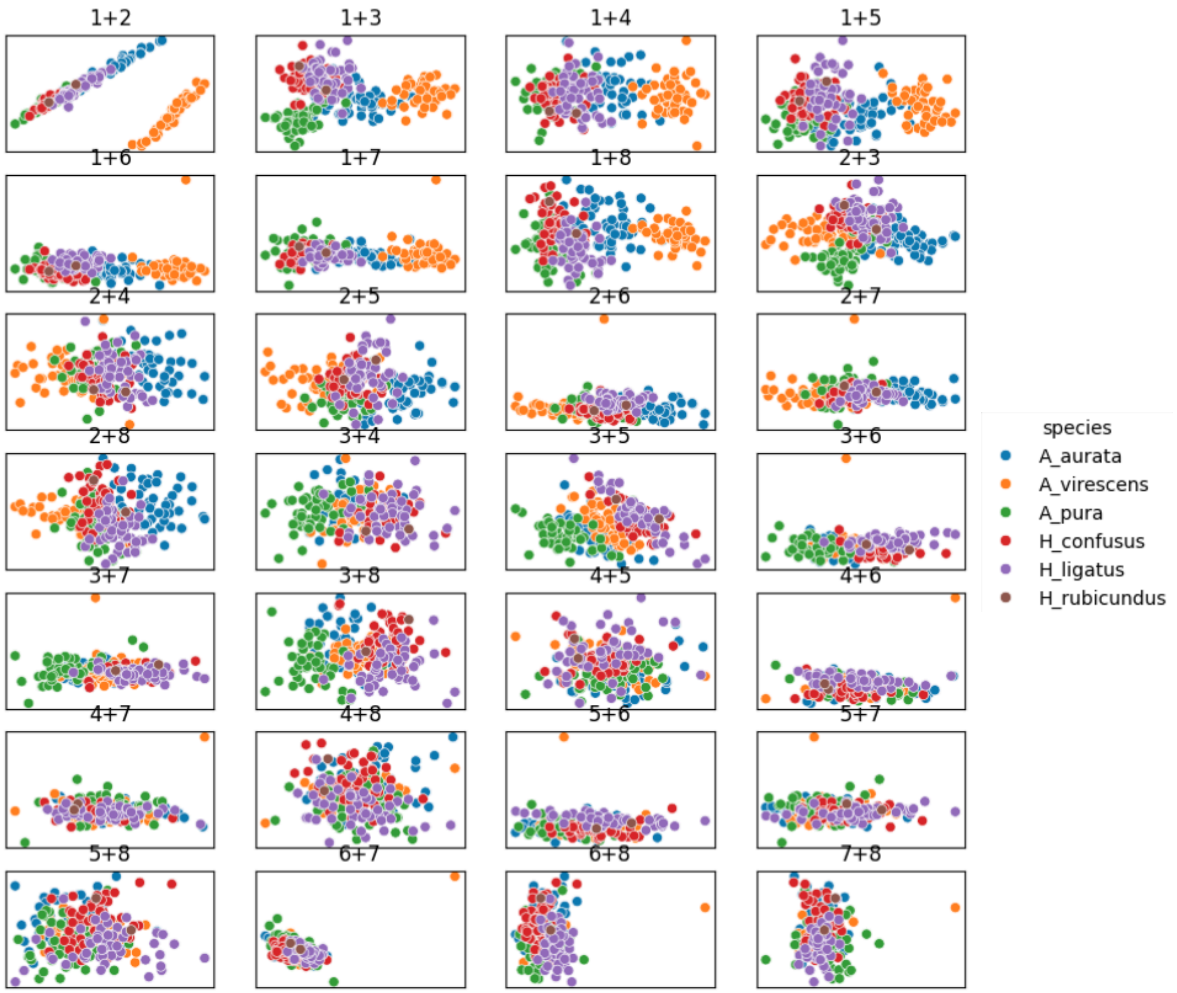


Figure 5: PCA (n components = 8) based on 250 features. The title of each plot describes the principal component along the x-axis + the principal component along the y-axis.

Although PCAs are a powerful tool for categorizing complex data, they assume that there is a linear relationship between the features upon which the analysis was run (i.e. the data points can be separated along a two-dimensional space). This is a byproduct of the fact that PCAs are based on the Pearson's correlation coefficients, and therefore assumes that there is a linear relationship between variables. Due to the non-linear nature of our extracted features (which we know from the low correlation rates reported in figure 4), we fitted our pca to the model of a

three-dimensional t-SNE (T-distributed stochastic neighbor embedding) plot to better visualize the differentiation between species.

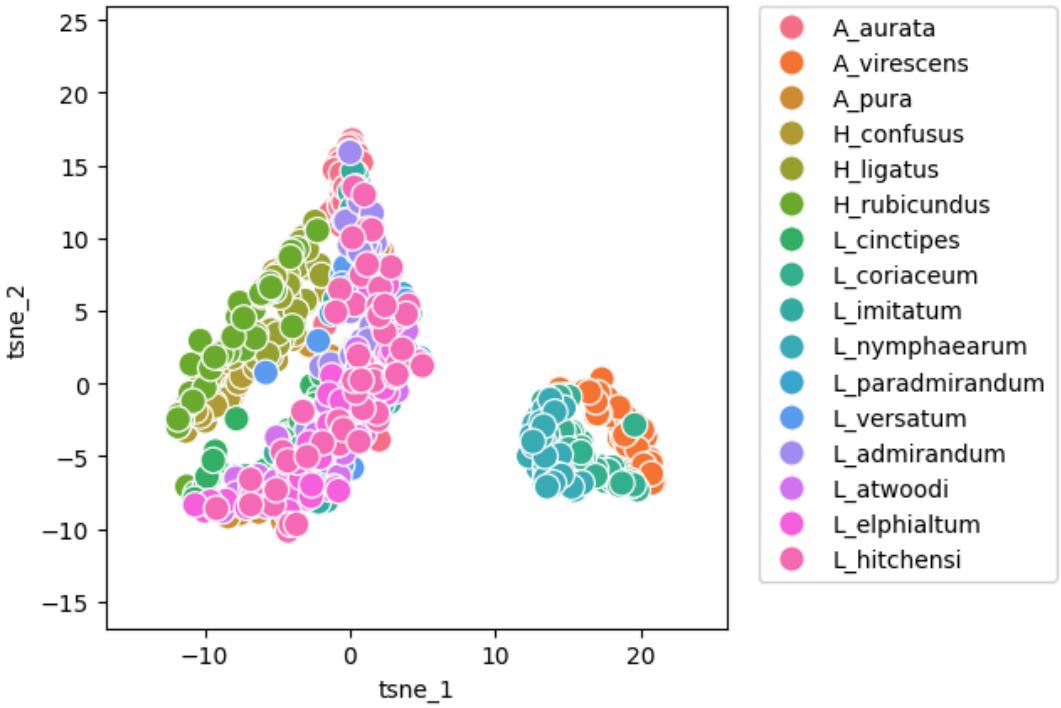


Figure 6: TSNE Plot of Species.

When developing a classification model, it's often assumed that by increasing the number of features indefinitely, one can increase the overall accuracy of the model. However, previous studies have shown that in practice, increasing the number of features beyond a certain threshold can cause an inadvertent reduction in identification accuracy instead (Hall 2011; Hua et al. 2005). For this reason we first trained the classification model on the initial set of extracted features (pixel distances and distance ratios), and on the components used in our PCA second. Ultimately, we compared the maximum validation scores returned by [1]the model trained using the raw features and [2]the model using components from a PCA.

V. Classification Scheme

a. Support Vector Machines

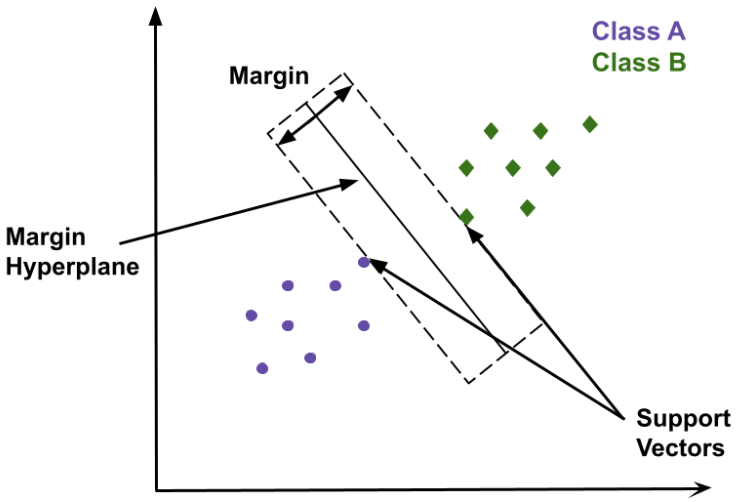


Figure 7: Diagram of how Support Vector Machines determine class distinctions.

For the creation and training of our classification model, we initially used the support vector machines (SVM) built into the scikit-learn package in python (scikit-learn 1.4.2). Support vector machines are a set of supervised learning algorithms which use a series of mathematical functions to determine where along an n-dimensional plane a collection of individual data points can be segregated into a predetermined number of classes. It accomplishes this by plotting the data points along a multi-dimensional plane in which each axis corresponds to a feature-based component that maximizes the distance between the most closely related points between each class (referred to as the margin). Within that margin is a line (or plane if using a three-dimensional space), referred to as the hyperplane or decision boundary, which denotes the

boundary that differentiates between two distinct classes. A data point will receive a different class prediction dependent on which side of that boundary it falls.

i. Training I

To begin training the SVM classification model you need two variables, one composed of the previously extracted features, and another composed of the correct species identifications, referred to as the response variable. Like all supervised learning methods, SVMs require that the above variables each be subdivided into a training and test dataset. For the purposes of this study, we divided the feature and response variables into training and test sets representing 80% and 20% of each dataset respectively, with each dataset also receiving an equal balance of each species. This means that the classification model will use the value from the training feature variable and corresponding identifications from the training response variable to make informed predictions about the species identifications of the remaining test features.

ii. Training II

We repeated the above procedure but instead of training and testing the model using the raw feature values, we used components from a PCA. We repeatedly ran the SVM pipeline based on PCAs iterated from 1 to 120 components in order to determine the minimum number of components necessary to maximize the prediction accuracy of the classification model.

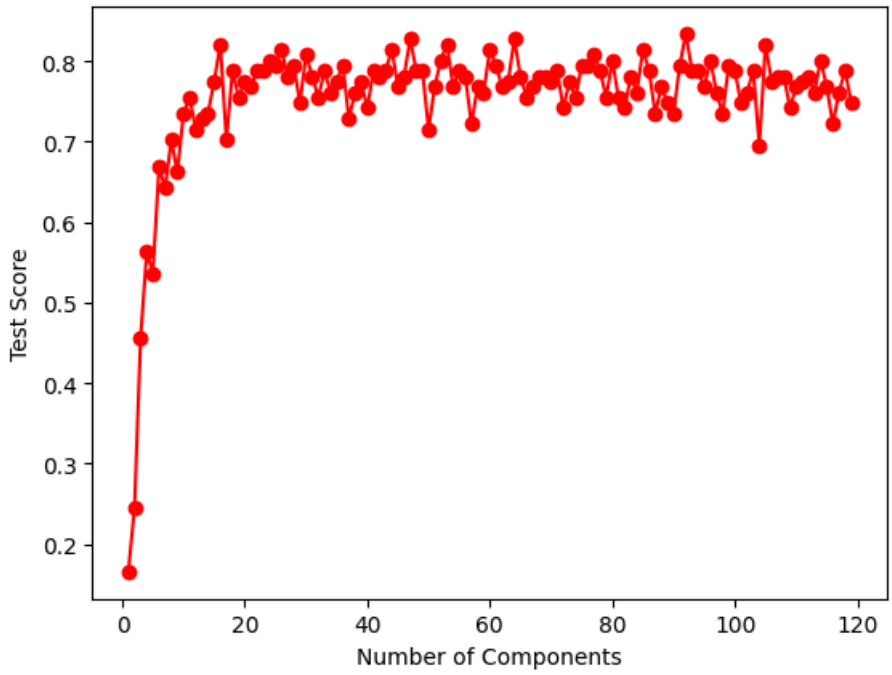


Figure 8: Test scores achieved by the SVM classification model as a function of PCA components.

b. Random Forest Classifier (RFC)

RFCs are widely accepted classification models, especially amongst computational biologists, for their accuracy and robustness. RFCs make predictions through the formation of 'decision trees'—each of which is generated by an algorithm, forming a "forest" of classifiers which all vote on which class should be assigned to a given observation (i.e which species a wing image belongs to) (Petkovic 2018). A diagram of the RFC decision-making process is shown below in Figure 9.

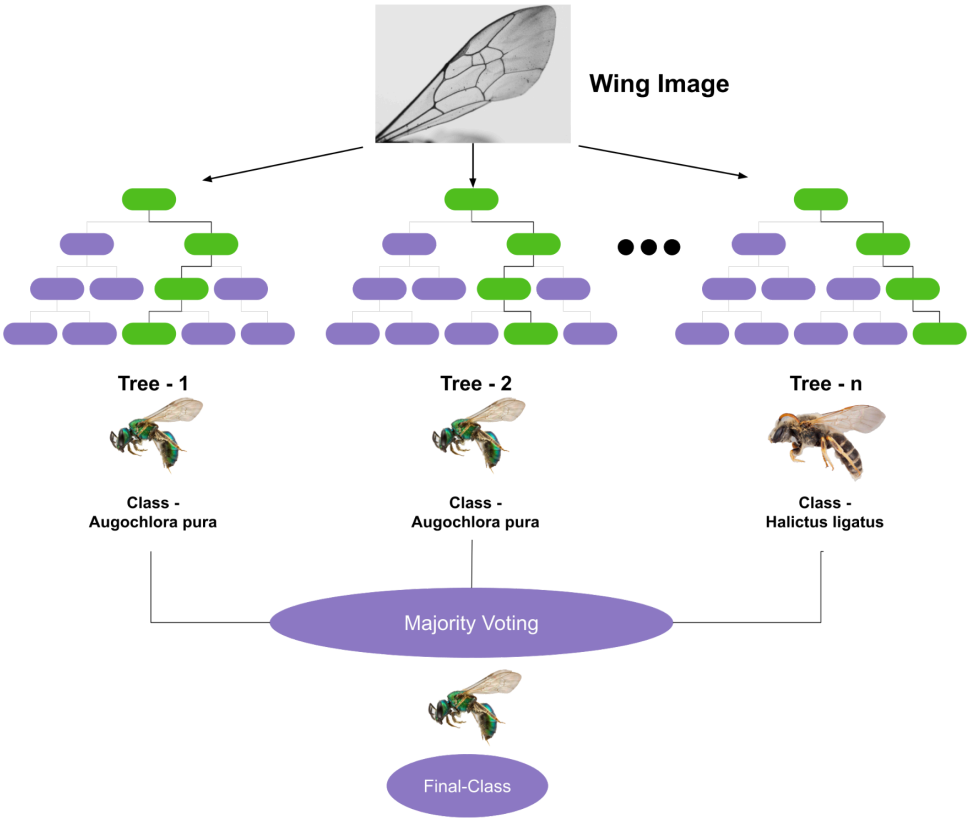


Figure 9: Diagram showing how the random forest classifier makes predictions.

VI. Validation

When training a classification model based on a set of predetermined features, it's possible to over-fit the model to the training data. This severely limits the model's generalizability and may render it ineffective when determining identifications for never-before-seen images. This issue is especially pervasive amongst models where a relatively large number of features is used to differentiate between a small number of classes. The classification model proposed in this study will ideally become a framework for a system which can deduce accurate species-level identifications for bees found across the United States, if not globally. However, our model is currently trained on a dataset that represents only 3.2% of the approximate number of halictid

bees native to North America, and only 0.4% of all bee species native to the United States (Woodard et al. 2020, Moisett and Buchmann 2011). Considering that our model was trained using features from only a fraction of the total potential classes (i.e., species), there was a chance that it may have been overfitted to our data.

One method used to avoid overfitting is known as k-fold cross validation. This procedure works by systematically dividing the dataset into groups such that each data point is included in the testing set a fixed number of times. After each division, the contents of the training and testing sets change. 'K' refers to the number of repetitions to be performed by the cross validation test. Ultimately, the test will return the model's average score across all the iterations. If there is a large disparity between the cross validation score and the initial training score of the model, it's assumed that the model must be overfitted to the training set. However, if the two scores are similar, it's assumed that the model is not overfitted and is therefore generalizable.

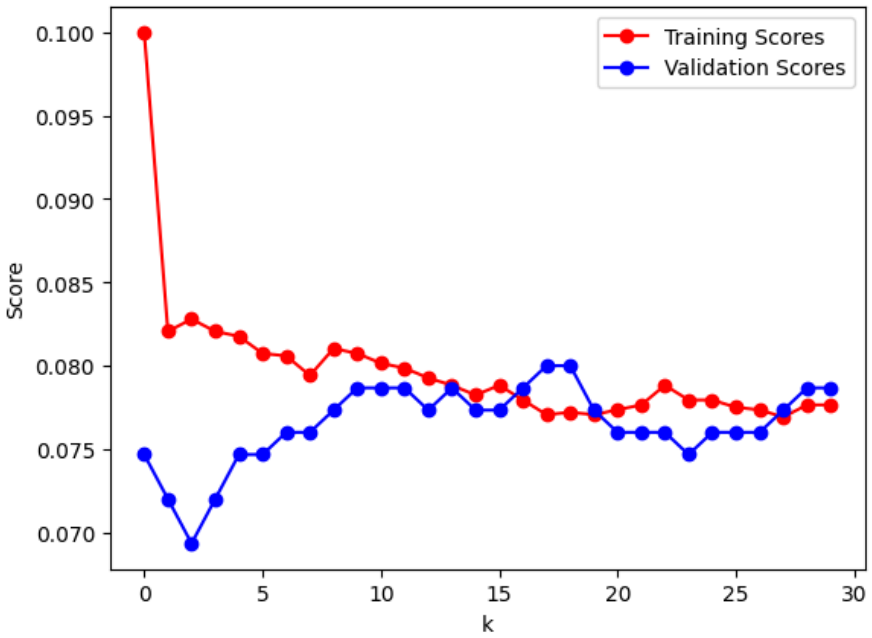


Figure 10: Plot of Training Scores and K-fold Cross Validation Scores as a function of K.

VII. Optimization

After checking whether or not the model was overfitted to our data, we ran additional analyses to further maximize the accuracy of its predictions. One such analysis is known as a grid search. In this procedure, a classification model is tested using various combinations of different hyperparameters defined within some predetermined range or grid. The term 'hyperparameters' refers to the variables or settings of a classification model which are set prior to training. Examples of these settings include: kernel, max depth, criterion, etc. Finding the optimal combination of these hyperparameters has the potential to greatly increase the prediction accuracy of the model. This often makes test scores reported prior to performing an exhaustive grid search misleading but valuable when making comparisons across different parameters.

a. SVM Hyperparameters

i. "C"

For our SVM classifier, we defined a grid which contained a range of values for the following hyperparameters: C , gamma, and kernel. C, referred to as the regularization parameter, controls the tradeoff between maximizing the size of the margin and minimizing training errors. A low C value maximizes the margin, potentially allowing for some misclassifications (i.e, a soft margin), and typically prevents overfitting the data, whereas high C values prioritize maximizing classification accuracy despite running the risk of overfitting the data. In our grid search we parsed through the following values: 0.1, 1, 10, 100, and 1000.

ii. Gamma

The parameter gamma represents the relative amount of influence each data point will have on the shape of the classifier's decision boundary, depending upon how close that point is to where the decision boundary is positioned along the feature space. High gamma values ensure that the shape of the decision boundary is almost entirely dependent on the few data points closests to the decision boundary. Conversely, low gamma values will give greater consideration to data points that are located further away from the decision boundary. We parsed through the following gamma values: 1, 0.1, 0.01, 0.001, and 0.0001.

iii. kernel

A kernel represents a series of mathematical functions that transform training data in such a way that allows for instances (i.e, wing images) to be separated linearly within a high-dimensional feature space. There are various types of kernels, each suited for different types of datasets. The kernels defined in our grid-search are as follows: linear, polynomial (poly), and radial basis function (RBF). As the name suggests, linear kernels assume that the training data is already linearly separable and thus, does not transform the data in any way. Linear kernels are renown for their computation speed and simplicity, but are ineffective when faced with complex, non-linear datasets. Polynomial kernels are more versatile than linear kernels and function with non-linear datasets. These kernels also function well with dense datasets, giving them an advantage in cases where additional features may be necessary to increase model prediction accuracy. Lastly, RBF kernels are similar to polynomial kernels in many ways, and both are commonly used for applications in computer vision and object recognition. Although RBF

kernels are often the default parameter for SVM models, performing a grid search is common practice to ensure that optimal parameters are chosen for a given model.

The final set of parameters we defined for our SVM model are as follows:

- **kernel** = RBF
- **C** = 1000
- **gamma** = 0.0001

```
svm = SVC(kernel = 'rbf', C = 1000, gamma = 0.0001, probability = True)
```

b. RFC Hyperparameters

i. N estimators

"N_estimators" specifies the number of decision-making trees included in the forest of the model (shown in Figure 9). The default value for this parameter is usually 10. Generally, the higher the number of trees encoded in the classifier, the more effectively the model learns the data. However, a large number of trees can considerably slow the training process. Therefore, a grid search will determine the number of trees necessary to maximize the prediction accuracy of the model without causing a large increase in training time. The range of n_estimator values that we considered in our grid search are as follows: 200, 300, 400, 500, 600, 700, 800, 900, 1000.

ii. Bootstrap

Bootstrapping is a statistical resampling technique in which instances are randomly sampled from the original training set *with replacement* in order to produce multiple separate training sets. The creation of these subsets are necessary because the training process for a random forest

classifier requires that each decision tree be trained on a unique dataset which has been derived from the original. Thus, the bootstrap parameter determines whether or not the RFC will sample the initial dataset with replacement during the training phase for each decision tree. Consequently, the parameter is satisfied with one of two logical values: True or False.

iii. Criterion

At each node in a decision tree (shown in Figure 9), the sample data is split according to the range of values they represent for a particular feature. For instance, samples that follow the right branch of the node may have values between 0-200 for a given morphological feature, whereas samples that follow the left branch have values greater than 200. The criterion parameter determines the feature variables and the threshold values used to determine that split. Included in our grid search are the criterion: gini (gini impurity), and entropy. The gini criterion assigns a number referred to as a 'gini impurity value" to each node on a decision tree that assesses the ability of the threshold assigned at that node to effectively split the samples into two sets. Gini impurity values range from 0 to 0.5, where a 0 or "pure" values indicate that the sample does not need to be split any further, and a 0.5, the most "impure" value, indicates that the sample must continue to be divided based on additional criteria. The entropy criterion measures a nearly identical performance metric. Entropy values range from 0 to 1 with 0 being the most pure and 1 being the most impure. Although entropy values tend to be more accurate than gini impurity, its values require more time to compute. As a result, a grid search is the appropriate analysis to determine the optimal tradeoff between accuracy and computational efficiency.

iv. Max Features

"Max_features" is the parameter which determines the number of extracted features the model will use when determining how to best split the sample dataset at each node in a decision tree. For our grid search we considered the following arguments: "sqrt", which sets the maximum number of features as equal to the square root of our total feature set, and "log2", which sets the maximum as $\log_2$ of the total feature set. We chose to explore these two options because they're already encoded into the set arguments for the parameter, and they both allow us to reduce the relative number of features we extracted for this study without jeopardizing the model's performance.

v. Max Depth

Each decision tree in a RFC possesses a "depth" which determines the number of splits they must perform per observation (i.e., wing image). Generally, the more splits a tree has, the information it captures about the observation. However, increasing the max depth of the decision trees indefinitely runs the risk of overfitting the data. For our grid search we considered the following range of max depth values: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, none. Setting the max depth as "none" means that each decision tree will continue to split until each node reaches the minimum number of samples (i.e., each node is considered "pure"; see section on criterion for further explanation on node purity).

The final set of parameters we defined for our RFC model are as follows:

- **N_estimators** = 800
- **Bootstrap** = True
- **Criterion** = 'entropy'

- **Max_features** = ‘sqrt’
- **Max_depth** = None

```
rfc = RandomForestClassifier(bootstrap = True, criterion = 'entropy', max_depth = None, max_features = 'sqrt', min_samples_leaf = 2, min_samples_split = 5, n_estimators = 800)
```

I. How to Interpret Results

a. Confusion Matrix

Visualization plays an important role in our ability to understand and compare the capacities of classification models to address a given problem. There are a number of different visualization methods, all of which convey separate aspects of the decision making process which results in the model’s identifications. For multinomial classifiers, like those employed in this study, confusion matrices are a standard visualization tool for reporting model performance (Alsallakh 2014). In a confusion matrix, a coordinate grid is formed with rows and columns that represent either the predicted class labels or the true class labels for each observation. Ideally, all results displayed on the matrix will appear directly along the diagonal of the plot. This indicates that the classifier correctly predicted the class of each test image 100% of the time. However, when values appear off the diagonal, this indicates the number of mis-classifications or instances where the model was ‘confused’. Since confusion matrices display mis-classifications, they have the potential to inform viewers about what classes the model is having a difficult time differentiating between. This is valuable information when re-evaluating the meaningfulness of the features used to train the classifier.

b. Performance Metrics

The overall performance of a classification model is usually summarized as the weighted average of the classifier prediction accuracy across all classes, a metric often referred to as the 'test score'. By adjusting the weight of each prediction accuracy with respect to class size, test scores provide a robust, albeit terse, representation of a classifier's capabilities.

c. Precision, Recall, and F1-Score

There are several metrics that, when reported in addition to the test score, provide a more comprehensive representation of a classification model's performance. These include but are not limited to metrics such as precision, recall, and f1-score. Precision refers to the proportion of true-positive cases which the model correctly predicted to be positive (e.g. the ratio of *Halictus ligatus* cases correctly predicted to be *H. ligatus* to the total number of cases predicted to be *H. ligatus*). Recall is the proportion of correctly identified true positives out of all true positives + false negatives for a given class. This metric is often used as a measure of how well a model can accurately identify relevant data (i.e. model sensitivity). Depending on the nature of the research question, it may be necessary to prioritize either greater recall or precision when evaluating the performance of a classification model. However, when developing a multi-species identifier such as this, it's essential that the model be able to reliably recognize a species as well as differentiate it from others. If not, the model would be an ineffectual tool for monitoring a family of insects as diverse and widespread as halictid bees. For this reason, we require a performance metric which serves as an indicator of both the model's precision and recall simultaneously. This metric is known as the f1-score and is calculated by finding the harmonic mean of the precision and recall

values. Therefore, a satisfactory f1-score is reflective of precision and recall values of similar quality.

VIII. Limitations

Throughout this study, we discovered a number of factors which may have limited our model's ability to produce the most accurate identifications. Discussed below is a list of such limitations and the ways in which they may have impacted our results.

a. Data Collection

The initial dataset of wing images that we generated to train our classification model is composed of images taken of full-bodied pinned specimens. This presented the most sustainable and least invasive opportunity to collect wing images because we didn't need to dismember the insects, allowing them to remain intact for future studies. This also meant that we were able to crowdsource specimens from multiple institutions without damaging their collections. However, this method may have negatively impacted our model's performance. Using full-body specimens requires that the bee, and its limbs, be continuously re-oriented to better display their forewings. However, bee forewings are not naturally positioned flat along the insects' body. As a result, we were not able to focus on every wing vein or vein joint under our digital microscope, especially those near the outer parameter of the wing. This directly limits the ability of our automated annotation software, ML-morph, to predict the position of those vein joints with high precision. This is due to the lack of clarity at those vein joints in our training images (Porto and Voje 2020).

Additionally, using full-bodied specimens meant that if the forewings of an insect were too folded or damaged to be photographed (especially at the positions of vein joints), we were forced to remove that individual from our dataset, ultimately reducing the size of our training set.

These problems could be solved by photographing forewings that had been detached, flattened, and secured to a medium prior to the beginning of the study. However, this would require the mutilation of several specimens, which would severely limit the number of species that could be included in the training set. Training using photographs of detached wings would also be antithetical to the ultimate goal of our model, which is to make accurate identifications of live specimens. Additionally, most institutions do not have collections of various species of wild bees with 50 or more specimens each with cataloged wings. The number of viable species becomes even smaller when you consider that many rare species of wild bees do not have extensive collections of specimens, and those that do likely cannot be damaged.

b. Species Representation and Feature Extraction

The number of species represented in this study is also relatively low compared to the number of wild sweat bee species native to the United States (~500 species) (Woodard et al. 2020, Moisett and Buchmann 2011). Although we trained our model so that it's generalizable, in the sense that it is not overfitted to our training data, a lack of species representation is still relevant when considering the practicability of our model and the quality of our chosen features. Our current model cannot make predictions on any species not included in our training set. This means that without adding additional images, our model can only predict the species of a fraction of Halictidae native to the US. Although we cannot overlook the importance of gaining more

granular data on the status of sweat bee pollinators, this is a clear example of the limited usability of our current model. However, these limitations can be surpassed in the future with the addition of wing images from new species to our dataset. Only a single wing image with a verified species identity is necessary to familiarize the model to a new species. In light of that, our goal is to have our model's repertoire of identifiable species gradually grow as users update the training set of wing images with at least a single image representing a new species.

In addition to its impact on our model's practicability, a lack of species representation can have implications on the quality of features used to make predictions. Classification models are entirely dependent on their ability to distinguish between species along a feature space. This means that the generalizability of our model could be threatened by overfitting our classifier to our current dataset, or by poorly choosing features. We eliminated the possibility of overfitting through the use of k-fold cross validation, so only the quality of our features remains a question. Ideally, the features we chose to inform our classification model will segregate species with little to no overlap between them. A subset of our features were capable of effectively segregating species (see Fig. 6, Fig. 11). However, it's possible that when faced with a set of species that is much larger than what it was initially trained on, the features we chose, and ultimately our model, may not perform as well.

Results

I. Biologically Significant Features

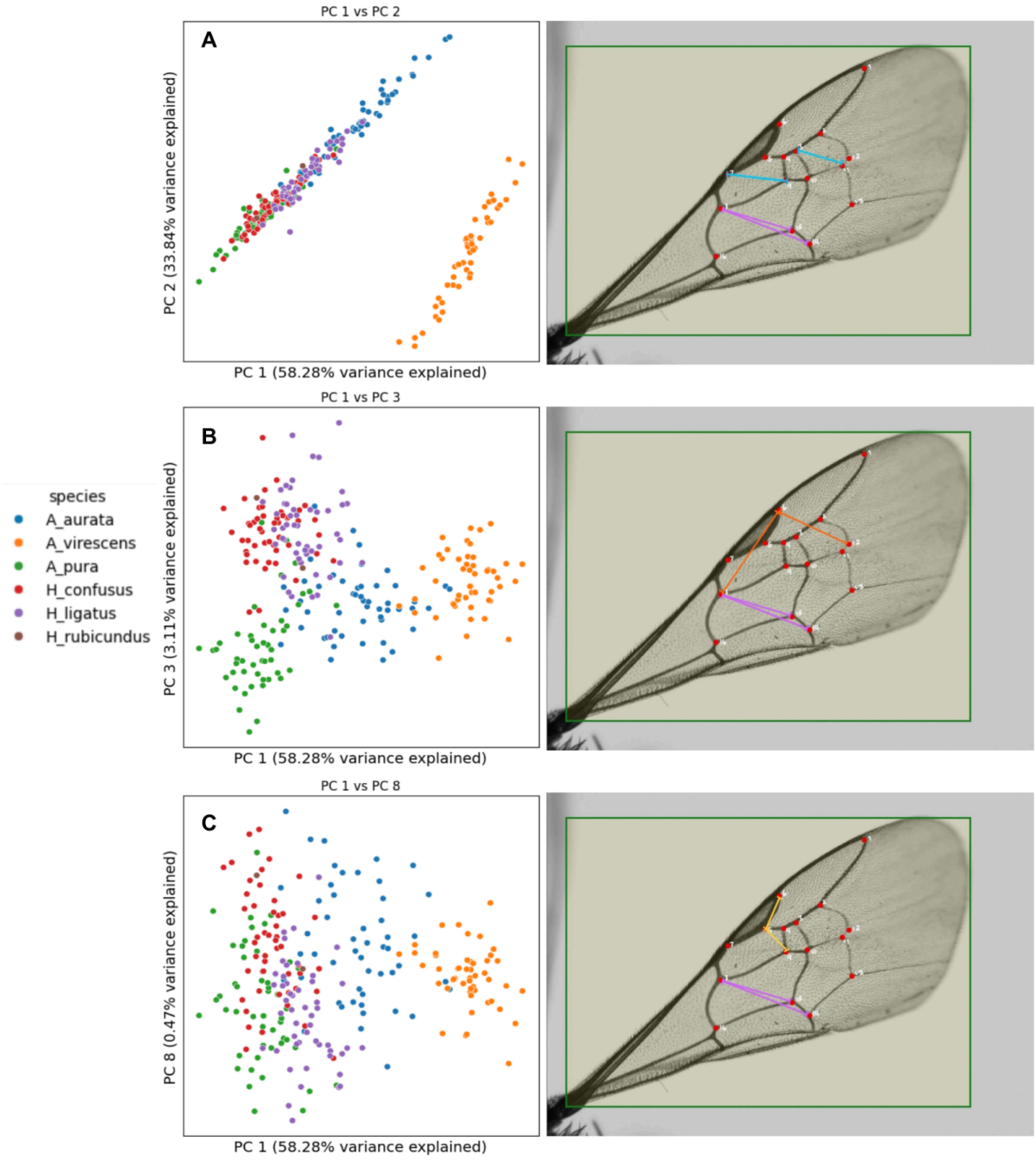
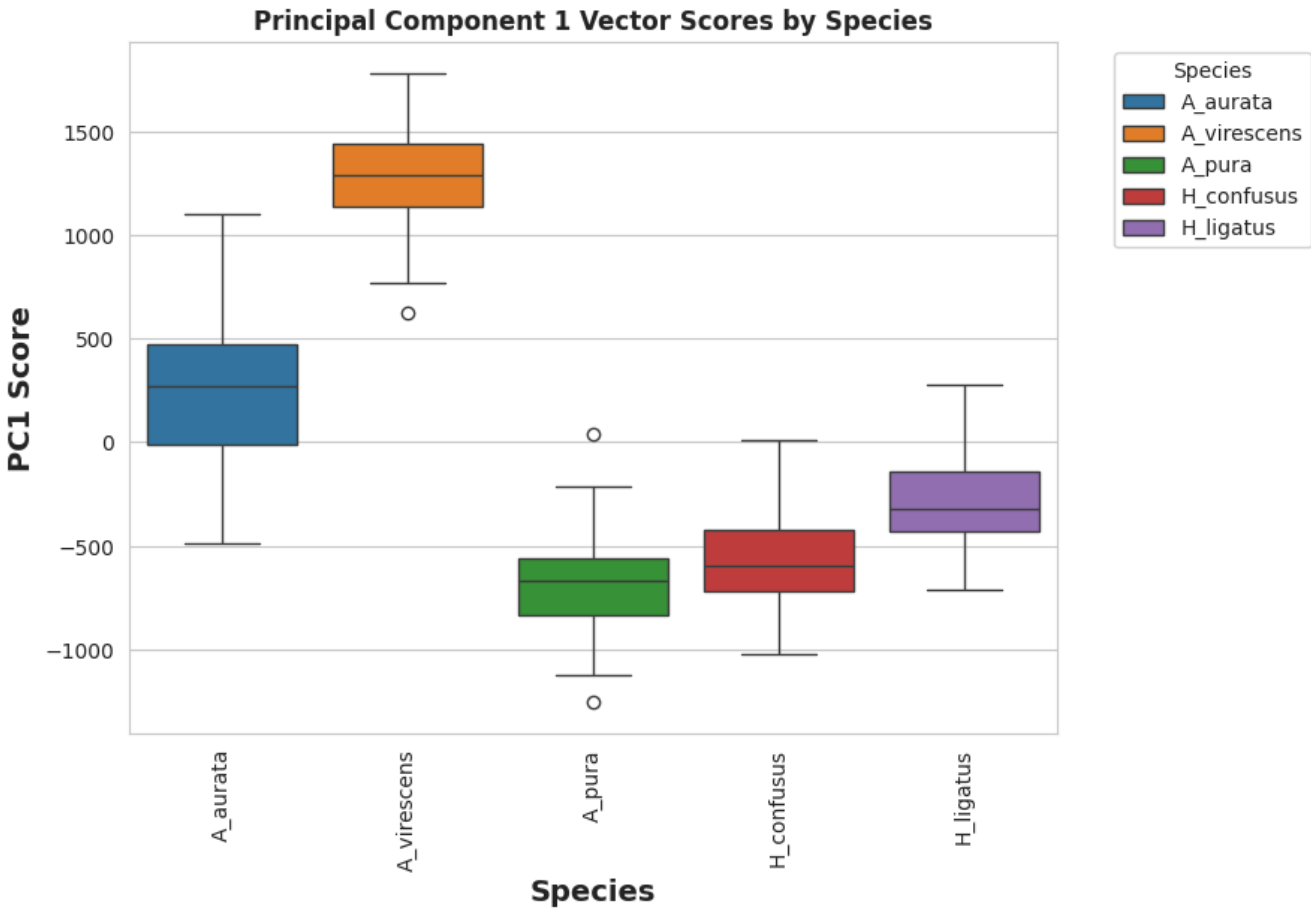


Figure 11: PCA plots showing the principal components which best segregate specimens by species and their corresponding morphological features (shown on a *Lasioglossum atwoodi* forewing). Figure 11A shows principal components (PC) 1 and 2, Figure 11B shows PC1 and 3, and Figure 11C shows PC1 and 8.



ANOVA F-value: 425.539
ANOVA p-value: 1.879×10^{-108}

Figure 12: Principal Component 1 Vectors Scores Across Five Species. This figure shows that there are significant differences between halictid species based on the features of wing morphology represented by PC 1.

After running our initial PCAs, we visually selected the components that were most successful at segregating the individual wing images by species. The morphological features (distance and distance ratios) which have the highest loadings on the principal components (PC) in the analysis are shown to the right of each plot.

PC1 was present in all three plots. PC1 and PC2, which accounted for 58.24% and 33.84% of the variance observed in our dataset (shown in Figure 11A), respectively, were most

successful at distinguishing *Agapostemon virescens* from the remainder of the species captured in the analysis. Although these components could not successfully segregate species aside from *A. virescens*, the results from this analysis may indicate that *A. virescens* possesses some adaptive feature of wing morphology that distinguishes it from the remaining species. This discovery could become the basis for future studies which explore the potential adaptive qualities and underlying mechanisms behind *A. virescens* wing morphology.

PC1 and PC3 (3.11%) (shown in Figure 11B) were most successful at clustering the members of each species into visibly distinct regions in our feature space. Similar to figure 9B, PC1 and PC8 (0.47%) (shown in Figure 11C) were successful at loosely segregating each species into distinct regions. However, it failed to cluster the individuals of each species compactly, which increased the amount of overlap between them.

The presence of PC1 in all three of our analyses indicates that the set of features which constitute this principal component may be a powerful morphological marker for distinguishing species of bees from one another. A similar conclusion can be drawn for PC3 considering that when plotted against PC1, the two components were able to segregate all represented species with minimal overlap. We extracted the features that correspond to each principal component, and found that the morphological feature represented by PC1 is the ratio of the distance between the upper left joint of the 1st medial cell (joint 8) and the terminal vein joint of the 2nd medial cell (joint 14) to the distance between the upper left joint of the 1st medial cell (joint 8) and bottom right vein joint of the 1st submarginal cell (joint 15). The wing feature that is represented by PC3 is the ratio of the distance between the upper most vein joint of the stigma (joint 6) and its midpoint (joint

5), to the distance between the midpoint of the stigma (joint 5) and the bottom left joint of the 2nd submarginal cell (joint 9).

The feature that corresponds to PC8 is the ratio of the distance between the upper most vein joint of the stigma (joint 6) and the upper left joint of the 1st medial cell (joint 8), to the distance between the upper most vein joint of the stigma (joint 6) and left most vein joint of the 3rd submarginal cell (joint 12). Although it wasn't as successful as its counterparts, the features that constitute PC8 may still represent key interspecies differences in bee wing morphology.

II. Classification Accuracy

The final test scores we reported for this study were achieved using the Random Forest Classifier (RFC) model instead of the Support Vector Machine (SVM). After performing a grid search and cross validations for both RFC and SVM, we compared their respective prediction accuracies. The final test score returned by the RFC model was 0.83 (shown in table 2), whereas the SVM model returned a value of 0.76 (when trained on raw features) and 0.78 (when trained on principal components) (see sections on training I and training II). Therefore, the RFC model shows greater accuracy and precision than the SVM model when predicting species identities.

species	precision	recall	f1-score	support
Augochlorella aurata	1.00	1.00	1.00	10
Augochlora pura	0.91	1.00	0.95	10
Agapostemon virescens	1.00	1.00	1.00	10
Halictus confusus	1.00	1.00	1.00	10
Halictus ligatus	1.00	1.00	1.00	10
Halictus rubincunclus	1.00	1.00	1.00	10
Lasioglossum admirandum	0.56	1.00	0.71	10
Lasioglossum atwoodi	0.40	0.40	0.40	10
Lasioglossum cinctipes	1.00	0.80	0.89	5
Lasioglossum coriaceum	0.91	1.00	0.95	10
Lasioglossum elphialtum	0.67	0.60	0.63	10
Lasioglossum hitchensi	0.55	0.60	0.57	10
Lasioglossum imitatum	1.00	1.00	1.00	10
Lasioglossum nyphaearum	1.00	0.90	0.95	10
Lasioglossum paradmirandum	0.00	0.00	0.00	6
Lasioglossum versatum	0.75	0.60	0.67	10
Weighted average	0.81	0.83	0.81	151
Accuracy	0.83			

Table 2: RFC Model performance metrics for each species.

The results from training an RFC model to identify species from our collected set of wing images were encouraging. On average, the RFC model returned an 83% prediction accuracy. The model also returned an f1-score of above 90% for nine out of the sixteen species included in this study. However, three species returned relatively f1-scores (<70%): *Lasioglossum atwoodi, L.*

elphialtum, and L. hitchensi. The individual performance metrics for each species are reported above in Table 2.

III. Misclassifications

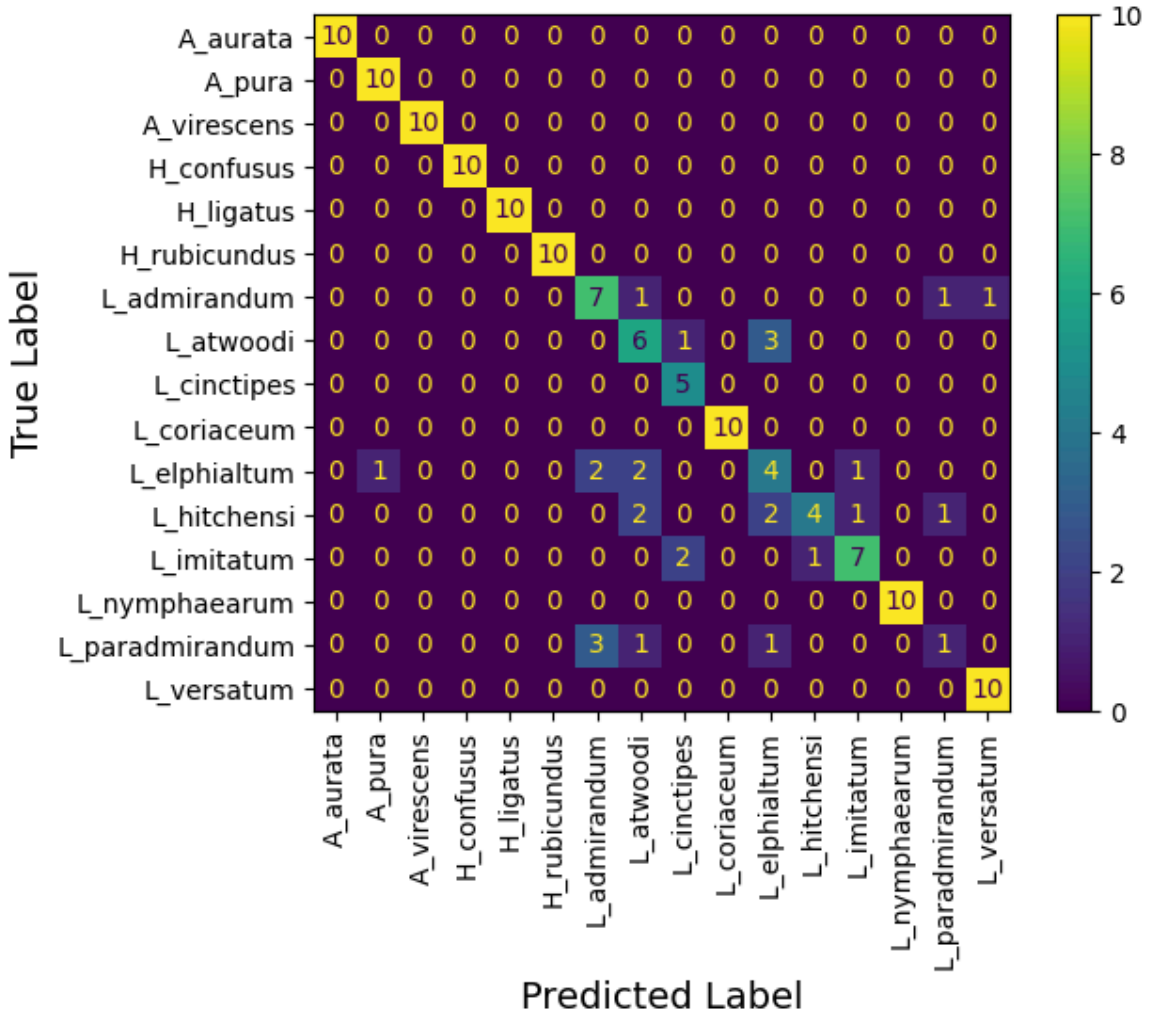


Figure 13: Confusion Matrix of Mis-classifications made by the classification Model. The matrix shows that our model is capable of predicting the identities of up to 10 different species with 100% accuracy (i.e., no misclassifications).

IV. Final Application Design

The final design for our bee identification app combines the trained RFC model and ML-morph pipeline in a software which allows the user to submit (or capture) images of bee wings and be returned with a species identification.

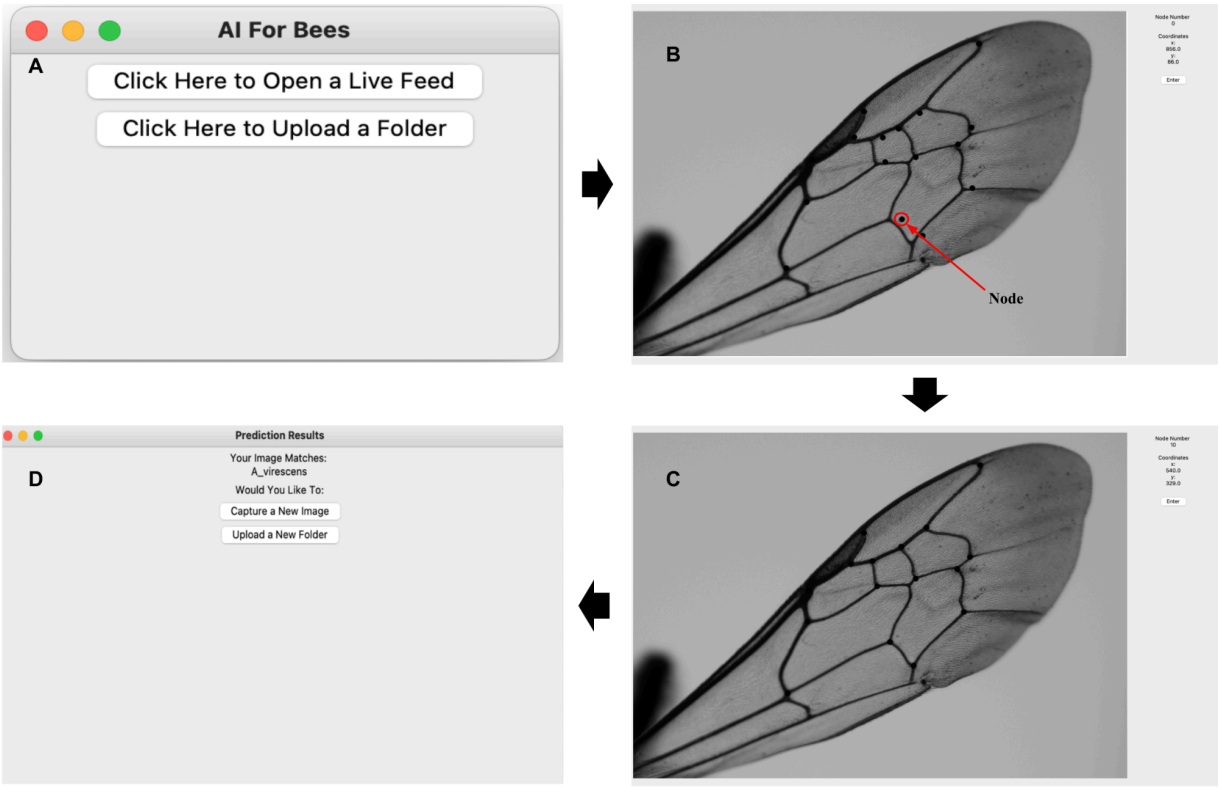


Figure 14: How to use our Bee Identification pipeline. *Figure 14A*: When first running our Bee Identification model, you will be prompted to either capture a live image of a bee forewing or upload a .png file of a forewing that was taken previously. *Figure 14B*: After uploading or capturing an image, the automated annotation software (ml-morph v.1.0.0) will place preliminary landmarks along the wing's vein joints. The xy coordinates of each point,, referred to as "nodes", are reported to the right of the image. At this stage, you're able to adjust the positions of each node to better align with the veins joints on the wing. *Figure 14C*: After correcting the positions of the nodes, press "enter" to feed that data to the bee classifier. *Figure 14D*: After submitting your image and landmarks to the classifier, the model will return a species prediction in a new window. In this example, the model correctly predicted the forewing as belonging to *Agapostemon virescens*.

Results Summarized

Wing venation patterns provide sufficient information to differentiate between several species of wild sweat bees. The first principal component in our analysis accounts for a disproportionately large amount of the variance which exists in our dataset (58.28%). This indicates that the morphological features which constitute that component are an essential contributor to our classification model's ability to distinguish between species.

Overall, the performance of our classification model is promising despite its imperfections. Using a random forest classifier, we were able to achieve a prediction accuracy of over eighty percent when trained on only 16 different species. The model was also able to identify ten species with one-hundred percent accuracy, including all members of the *Halictus*, *Augochlora, Augochlorella,* and *Agapostemon* genera. However, all of the lowest scoring classes (i.e. species) belonged to the *Lasioglossum* genus, including *Lasioglossum paradmirandum* which received only a single correct classification by our model. It's possible that the model's inability to correctly classify this species could be rectified by increasing the size of its training set. In this study, *L. paradmirandum* possessed the minimum number of training images required to successfully train a classification model (30 images as opposed to our recommended 50). However, *L. cinctipes* possessed a training set of comparable size (29 images) and the model performed much better than on its counterpart. Therefore, it's likely that while the size of the training set may be a factor in the model performance, there are additional factors which have contributed to such a large disparity in classification accuracy. Additionally, the majority of mis-classified images of *L. paradmirandum* were classified as *L. admirandum*. This pattern is reflective of how morphologically similar individuals from these two species are when compared

to the remainder of species represented in this study. Even expert taxonomists find it difficult to distinguish between the *L. paradmirandum* and *admirandum*, which ultimately solidified their placement in a group of species known for being generally indistinguishable from one another, *Lasioglossum viridatum* (Rhoades et al. 2011).

Discussion

Insect pollinators play a pivotal role in our effort to maintain the health and productivity of natural ecosystems and agricultural landscapes. As of 2005, insect pollination, of which bees are the primary contributors, provided nearly 10% of the total economic value of agricultural production used directly for human consumption (Khalifa et al. 2021; Gallai et al. 2009). However, the value of insect pollinators goes well beyond that of their provisioning services. Nearly 80% of wild plants are dependent upon pollination by bees, making them indispensable as benefactors to plant diversity in virtually all terrestrial ecosystems (Kopec & Burd 2017; Potts et al. 2010). This links the status of bee communities to the biodiversity of native plant communities, and therefore links them to ecosystem stability and potential for carbon storage (Ayral et al. 2020, Woodard et al. 2020). For these reasons, insect pollinators are of high priority on an extensive list of taxa that require immediate protections. Despite this, global estimates of insect pollinator communities have shown rapid declines in both aggregate abundance and species richness (Turley et al. 2022; Cameron et al. 2011, Biesmeijer et al. 2006). Previous studies show that several species of wild bees are significantly less abundant or absent from many more localities than would be predicted from natural history collections (Cameron et al. 2011).

In light of their decreasing numbers, conservation efforts have been slow and relatively ineffective at providing the necessary protections to insect pollinators (Colla 2022). The majority of management strategies prioritize the non-native european honey bee, *Apis mellifera*, and neglect the highly diverse communities of wild bees. A predominant factor behind the relative ineffectiveness of wild bee conservation is the lack of specificity and detail in estimates of bee populations, especially regarding species-level population dynamics. Initiatives aimed at monitoring bee populations and assessing biodiversity (i.e., producing the necessary data from which to base conservation strategy) have been sluggish and confounded by the practical difficulty of manually identifying individual specimens (Portman et al. 2020). Current methods of identifying pollinator species revolve around the expert opinion of a small number of taxonomists capable of differentiating between bee species by minute morphological distinctions or through molecular (DNA) barcoding which requires access to a genomics lab and an existing database of species-level identifications. By developing a machine learning tool that can provide rapid taxonomic identifications for bees, we begin to dismantle these barriers to expansive pollinator monitoring.

In our study, we found that the venation patterns which outline the forewings of wild bees represent a significant morphological feature capable of differentiating between species. Although previous literature has examined the potential of wing images to inform honey bee identifications, few studies have evidenced the potential for wing venation to distinguish native wild bee species (Oleksa and Tofilski 2014; Oleska et al. 2023; Arbuckle et al. 2001). This finding illuminates the potential for wing image-based identifications to bridge a widening information gap between what's known about the conversation status of the non-native and

intensely managed honey bee and the community of wild bees which constitute the majority of insect pollinators.

Additionally, we were successful in using bee wings and their corresponding venation patterns to train a classification model which can accurately identify wild bees down to the species-level. With an overall accuracy of above 80%, our model is capable of reliably distinguishing between what is currently a set of 16 different species of wild sweat bees. Our use of wing venation patterns allows us to gain much higher resolution on species identities than would typically be possible with approaches that use full body images. Full-body capture typically requires the deployment of camera traps to record video footage of pollinators at a variety of flower visitation sites (Stefan et al. 2024). However, there is a trade-off between the ease of placing cameras to monitor pollinator populations, and the degree of taxonomic detail which can be retrieved from their footage. Although camera traps may be able to collect large scale data on pollinator communities with minimal time investment from researchers, their inability to classify species to the lowest taxonomic level make them ineffective when answering questions about species-specific dynamics. This implies that the best strategy for conserving insect pollinators may involve a combination of methods including both wing venation-based classification (our model), and full-body imaging methodologies.

The current pipeline for our model requires little to no specialized knowledge of programming nor bee morphology in order to use. A built-in automated image annotation software both streamlines the identification process and removes barriers to those who may have limited knowledge on insect wing morphology. Thus, both melittologists and those in unrelated fields or professions will be able to rapidly amass precise and accurate, species-level data on local pollinator communities. Using our current model, we will gain greater insight into the

conservation status, population dynamics, and dispersal of US native sweat bees. With this data, effective conservation strategies can be implemented with the goal of protecting and restoring populations of these essential insect pollinators. Without this data, crop lands and wild plant communities across the US face the threat of losing a number of pollinator species upon which they have become largely dependent. By hastening the process by which wild bees are identified, and liberating opportunities for citizen science-based approaches to bee monitoring, this software will have a profound positive impact on the future of insect pollinator conservation.

Future Work

I. Improvements

Although the software we've produced is far from perfect, some of its best qualities lie in its widespread applications and its ability to be improved upon by the user. Currently the dataset we generated for the purpose of training the classification model is reflective of only 16 species. Although this number is negligible in comparison to the number of species native to New Jersey, the dataset can easily be expanded upon through the addition of new wing images. As the name suggests, machine learning algorithms possess the ability to constantly improve their functionality in response to the introduction of new information (i.e. new observations or wing images from currently unrepresented species). As users continue to add wing images to the training set, the classification model will both become familiarized with new species (so long as at least one image has a verified identity belonging to that species), and become more refined in its predictions of familiar species. However, there will be challenges to adjusting the current model to accommodate a training set large enough to apply to all native wild bees in the United

States, and eventually the world. It will likely become necessary to extract additional features from our wing images to ensure our model's ability to distinguish between an increasing number of species (some of which may possess morphological similarities which are not currently discernible by the model). Similarly, it may become necessary to replace the current classifier used in our model (random forest classifier) with one that is better suited for such a large range of species.

Lastly, it's possible that the dataset of wing images will reach a number of species where bees can no longer be identified based solely upon wing venation. Should our model be tasked with identifying a majority of known bee species, this may become the reality. In this case, it will be necessary to allow the user to input additional data such as: region/locality, habitat type, and time of year when the specimen was observed. This information will allow the model to eliminate potential species which may have similar wing morphologies as the specimen, but occupy different spatio-temporal niches. This would further streamline the identification process and address concerns regarding the model's generalizability. Identical procedures have already proven to be effective in identification tools such iNaturalist, a nonprofit social network of biologists and citizen scientists built around a software which identifies and maps observations of various species across the globe (https://www.inaturalist.org/).

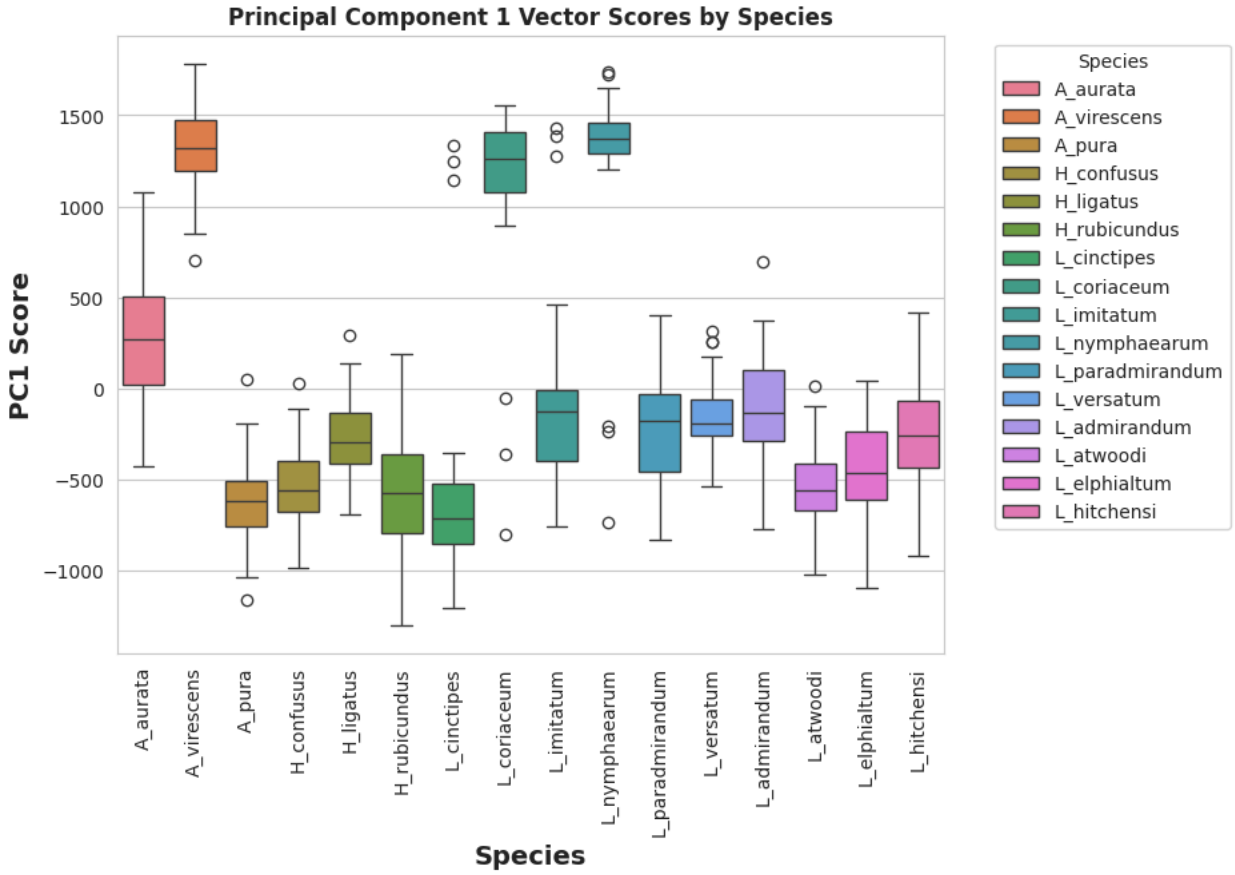
II. Applications

There is a growing body of literature which argues for the potential of machine learning based object recognition and smartphones to play a pivotal role in the field of conservation monitoring (Nart & Costa 2022; Stefan et al. 2024). Smartphones are well renowned as tools for large-scale data collection due to their ubiquity and ability to collect data on various habitat variables through integrated sensors for sound, location, ambient light, atmospheric pressure, and

temperature (Stefan et al. 2024). This makes them an ideal tool for citizen science based approaches in biology. Thus, one of the immediate next steps in our study is to incorporate our bee classification model into a mobile application that can be readily downloaded onto a smartphone. This will greatly increase the accessibility of our model beyond that of highly-informed researchers and taxonomists, and will create a platform to collect pollinator census data from various localities. Our goal is to create an application that aligns with our initial objective of developing an identification tool that is both robust and accessible to those without experience in bee monitoring or computer programming.

In addition to becoming a mobile application, our model has the potential to contribute greatly to the field of education and undergraduate research. As a tool, we designed our model to be used by anyone who is interested in pollinator monitoring, which includes students and other young scientists. We plan to incorporate our model, and its future iterations, into the research projects and field courses of undergraduate students. Using this tool, students will be capable of gathering data on pollinator biodiversity across a variety of landscapes and land-use regimes. This would add to what is currently a scarce collection of literature on the impact of environmental stressors and land-use types on the populations of specific insect pollinators. Additionally, students would gain the opportunity to use a novel technology to pursue research questions in a currently lacking field of study.

Appendix



ANOVA F-value: 196.228
ANOVA p-value: 6.076×10^{-245}

Figure 15: Principal Component 1 Vectors Scores Across All Species.

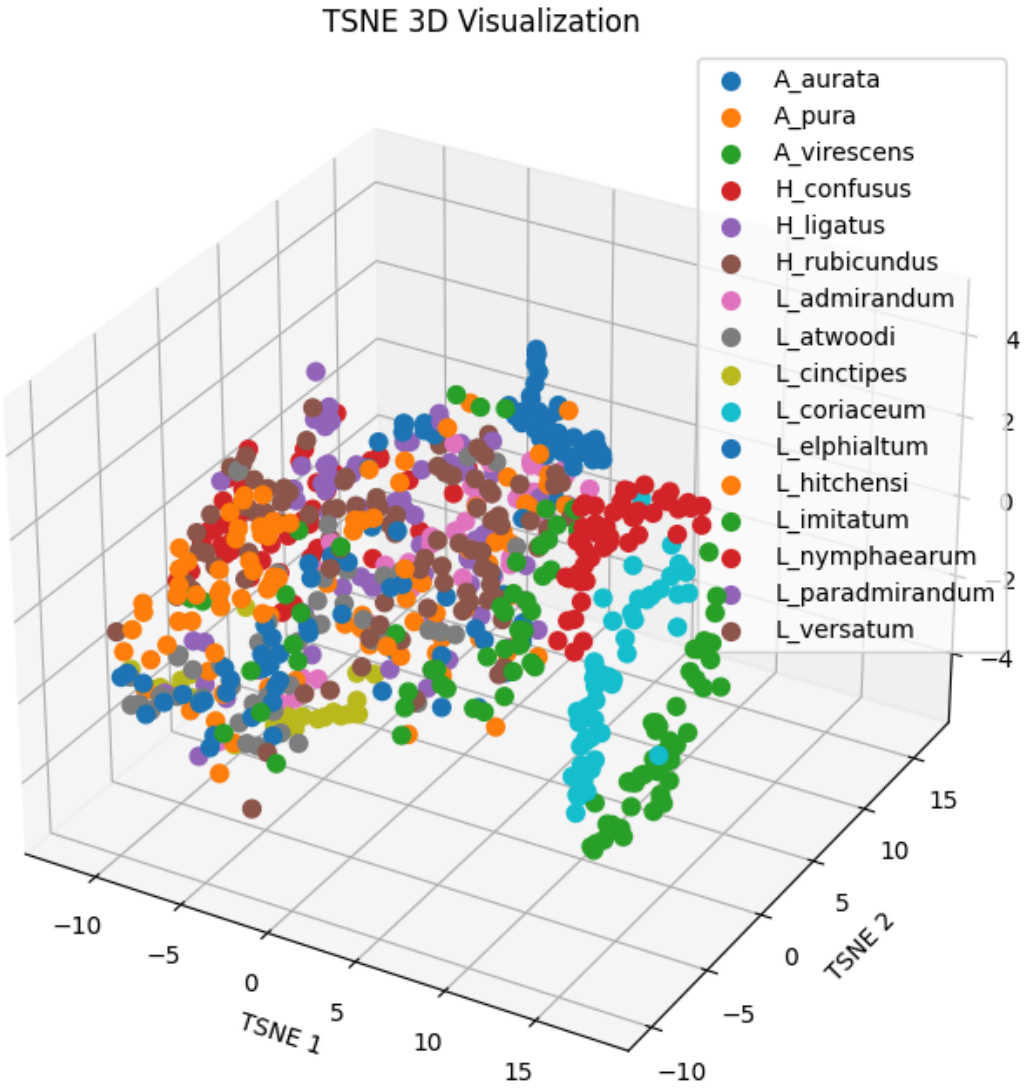


Figure 16: 3D Visualization of TSNE Plot of Halictid Species

References

AAA, Algethami AF, Musharraf SG, AlAjmi MF, Zhao C, Masry SHD, Abdel-Daim MM, Halabi MF, Kai G, Al Naggar Y, Bishr M, Diab MAM, El-Seedi HR. Overview of Bee Pollination and Its Economic Value for Crop Production. Insects. 2021 Jul 31;12(8):688. doi: 10.3390/insects12080688. PMID: 34442255; PMCID: PMC8396518.

Alsallakh, Bilal & Hanbury, Allan & Hauser, Helwig & Miksch, Silvia & Rauber, Andreas. (2014). Visual Methods for Analyzing Probabilistic Classification Data. IEEE Transactions on Visualization and Computer Graphics. 20. 1703-1712. 10.1109/TVCG.2014.2346660.

Arbuckle, Tom & Schröder, Stefan & Steinhage, Volker & Wittmann, Dieter. (2002). Biodiversity Informatics in Action: Identification and Monitoring of Bee Species using ABIS. Proc. 15th Int. Symp. Informatics for Environmental Protection.

Cameron, Sydney A., et al. "Patterns of Widespread Decline in North American Bumble Bees." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 2, 2011, pp. 662–67. *JSTOR*, http://www.jstor.org/stable/25770838. Accessed 24 Apr. 2024.

Colla, Sheila R. The potential consequences of 'bee washing' on wild bee health and conservation, International Journal for Parasitology: Parasites and Wildlife, Volume 18, 2022, Pages 30-32, ISSN 2213-2244, https://doi.org/10.1016/j.ijppaw.2022.03.011.

De Nart, D., Costa, C., Di Prisco, G. *et al.* Image recognition using convolutional neural networks for classification of honey bee subspecies. *Apidologie* **53**, 5 (2022). https://doi.org/10.1007/s13592-022-00918-5

Erickson, Bradley J., and Felipe Kitamura. "Magician's Corner: 9. performance metrics for machine learning models." *Radiology: Artificial Intelligence*, vol. 3, no. 3, 1 May 2021, https://doi.org/10.1148/ryai.2021200126.

Garibaldi, Lucas A. *et al.* ,Wild Pollinators Enhance Fruit Set of Crops Regardless of Honey Bee Abundance. *Science***339**,1608-1611(2013).DOI:10.1126/science.1230200

Ghosh, Sampat & Aryal, Sunil & Jung, Chuleui. (2020). Ecosystem Services of Honey Bees; Regulating, Provisioning, and Cultural Functions. Journal of Apiculture. 35. 10.17519/apiculture.2020.06.35.2.119.

Inoka, WA & Karunaratne, Inoka & Edirisinghe, Jayanthi. (2008). Key to the identification of common bees of Sri Lanka. Journal of The National Science Foundation of Sri Lanka - J NATL SCI FOUND SRI LANKA. 36. 10.4038/jnsfsr.v36i1.134.

Jianping Hua, Zixiang Xiong, James Lowey, Edward Suh, Edward R. Dougherty, Optimal number of features as a function of sample size for various classification rules, *Bioinformatics*, Volume 21, Issue 8, April 2005, Pages 1509–1515, https://doi.org/10.1093/bioinformatics/bti171

Kopec, Kelsey, and Lori Ann Burd. "Pollinators in Peril: A Systematic Status Review of North American and Hawaiian Native Bees." *Search Issue Lab*, Issue Lab, 1 Feb. 2017, search.issuelab.org/resource/pollinators-in-peril-a-systematic-status-review-of-north-american-and-hawaiian-native-bees.html.

Landaverde-González P, Quezada-Euán JJG, Theodorou P, Murray TE, Husemann M, Ayala R, Moo-Valle H, Vandame R, Paxton RJ. Sweat bees on hot chillies: provision of pollination services by native bees in traditional slash-and-burn agriculture in the Yucatán Peninsula of tropical Mexico. J Appl Ecol. 2017 Dec;54(6):1814-1824. doi: 10.1111/1365-2664.12860. Epub 2017 Jan 27. PMID: 29200497; PMCID: PMC5697652.

Mason, Lisa, Arathi, H.S., Assessing the efficacy of citizen scientists monitoring native bees in urban areas, Global Ecology and Conservation, Volume 17, 2019, e00561, ISSN 2351-9894, https://doi.org/10.1016/j.gecco.2019.e00561.

Matias, D.M.S., Leventon, J., Rau, AL. *et al.* A review of ecosystem service benefits from wild bees across social contexts. *Ambio* **46**, 456–467 (2017). https://doi.org/10.1007/s13280-016-0844-z

Michener, Charles Duncan. *The Bees of the World*. Johns Hopkins University Press, 2007.

Moissett, Beatriz, et al. *Bee Basics: An Introduction to Our Native Bees*. USDA, Forest Service, 2010.

Nicola Gallai, Jean-Michel Salles, Josef Settele, Bernard E. Vaissière, Economic valuation of the vulnerability of world agriculture confronted with pollinator decline, Ecological Economics, Volume 68, Issue 3, 2009, Pages 810-821, ISSN 0921-8009, https://doi.org/10.1016/j.ecolecon.2008.06.014.

Oleksa A, Căuia E, Siceanu A, Puškadija Z, Kovačić M, Pinto MA, Rodrigues PJ, Hatjina F, Charistos L, Bouga M, Prešern J, Kandemir İ, Rašić S, Kusza S, Tofilski A. Honey bee (Apis mellifera) wing images: a tool for identification and conservation. Gigascience. 2023 Mar 20;12:giad019. doi: 10.1093/gigascience/giad019. PMID: 36971293; PMCID: PMC10041535.

Oleksa, A., Tofilski, A. Wing geometric morphometrics and microsatellite analysis provide similar discrimination of honey bee subspecies. *Apidologie* **46**, 49–60 (2015). https://doi.org/10.1007/s13592-014-0300-7

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217–222. https://doi.org/10.1080/01431160412331269698

Petkovic D, Altman R, Wong M, Vigil A. Improving the explainability of Random Forest classifier - user centered approach. Pac Symp Biocomput. 2018;23:204-215. PMID: 29218882; PMCID: PMC5728671.

Portman, Zachary & Bruninga-Socolar, Bethanne & Cariveau, Daniel. (2020). The State of Bee Monitoring in the United States: A Call to Refocus Away From Bowl Traps and Towards More Effective Methods. Annals of the Entomological Society of America. 113. 10.1093/aesa/saaa010.

Porto A, Voje KL. ML-morph: A fast, accurate and general approach for automated detection and landmarking of biological structures in images. *Methods Ecol Evol*. 2020; 11: 500–512. https://doi.org/10.1111/2041-210X.13373

Potts, Simon G., Biesmeijer, Jacobus C., Kremen, Claire, Neumann, Peter, Schweiger, Oliver, Kunin, William E., Global pollinator declines: trends, impacts and drivers, Trends in Ecology & Evolution, Volume 25, Issue 6, 2010, Pages 345-353, ISSN 0169-5347, https://doi.org/10.1016/j.tree.2010.01.007.

Python Software Foundation. Python Language Reference, version 3.12.3. Available at http://www.python.org

Rhoades, Paul R. , Klingeman, William E., Trigiano, Robert N., Skinner, John A., "Evaluating Pollination Biology of *Cornus florida* L. and *C. kousa* (Buerger ex. Miq.) Hance (Cornaceae: Cornales)," Journal of the Kansas Entomological Society, 84(4), 285-297, (1 October 2011)

Rodrigo Gómez, S., Ornosa, C., Selfa, J., Guara, M., and Polidori, C. (2016) Small sweat bees (Hymenoptera: Halictidae) as potential major pollinators of melon (*Cucumis melo*) in the Mediterranean. *Entomological Science*, 19: 55–66. doi: 10.1111/ens.12168.

Roubik, David W. *Pollination of Cultivated Plants in the Tropics*. Food and Agriculture Organization of the United Nations, 1995.

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

Ștefan, Valentin, et al. Utilising Affordable Smartphones and Open-Source Time-Lapse Photography for Monitoring Pollinators, 2 Feb. 2024, https://doi.org/10.1101/2024.01.31.578173.

Steinhage, V., Schröder, S., Roth, V., Cremers, A.B., Drescher, W. and Wittmann, D. (2006), The Science of “Fingerprinting” Bees. German Research, 28: 19-21. https://doi.org/10.1002/germ.200690003

Tomar, Riya, et al. “Deep learning-powered camouflaged object recognition model.” *2024 International Conference on Integrated Circuits and Communication Systems (ICICACS)*, 23 Feb. 2024, https://doi.org/10.1109/icicacs60521.2024.10498339.

Turley NE, Biddinger DJ, Joshi NK, López-Uribe MM. Six years of wild bee monitoring shows changes in biodiversity within and across years and declines in abundance. Ecol Evol. 2022 Aug 12;12(8):e9190. doi: 10.1002/ece3.9190. PMID: 35983174; PMCID: PMC9374588.

Winfree, Rachael *et al.* ,Species turnover promotes the importance of bee diversity for crop pollination at regional scales.*Science***359**,791-793(2018).DOI:10.1126/science.aao2117

Woodard, S. Hollis, Federman, Sarah, James, Rosalind R., et al.,Towards a U.S. national program for monitoring native bees, Biological Conservation,Volume 252, 2020, 108821, ISSN 0006-3207, https://doi.org/10.1016/j.biocon.2020.108821.

Yang, Zheng Rong, Biological applications of support vector machines, *Briefings in Bioinformatics*, Volume 5, Issue 4, December 2004, Pages 328–338, https://doi.org/10.1093/bib/5.4.328

This thesis represents my own work in accordance with University regulations.

- Jahir Morris