Project Title: Honeybee Image Classification to Predict Beehive Health

(Members: Shilpika Banerjee, Swetha Govindu, Fiona Senchyna)

1. Source code repository and instructions for code execution

Source Code on Github. Repository link **https://github.com/fsenchyna/CSC869_Term_Project**
Steps to run the code are detailed in the ReadMe document on GitHub.

2. Problem Definition

Bees are an essential part of our ecosystem. They are vital to plant pollination on which humans rely for their food. Bee populations have been in steep decline over recent decades. Many factors are thought to have contributed to their decline like climate change, mite infestations, etc. However, in recent times personal beekeeping has gained popularity due to the increased awareness of bees' decline. As for beekeepers, they must periodically inspect their beehive to ensure its health. Not only does it require manual intervention but also means disturbing the internal environment of the hive which could have unintended consequences. We found a Kaggle dataset[1] which proposes that the health of a hive can be inferred from the images of bees taken when they leave the hive (e.g., wings intact, bees with pollen). A model that could predict the hive health from bee images would reduce the work of beekeepers and also make the regular health checkup less invasive for the hives.

3. Dataset Description

Our dataset contains 5100+ bee images annotated with location, date, time, subspecies, health condition, caste, and pollen.

Data Visualization:

Early on in this project we performed some data visualization on the metadata of the images. We saw how the hive health, depicted by the six class labels (Fig 1), vary based on geographical locations, time of the day, bee subspecies, and caste of the bees (e.g., drones, workers, queen bees). For the 'caste' we found that all data points, regardless of the hive health, are images of worker bees. Therefore, we conclude that the column 'caste' is most likely insignificant in contributing to the prediction and can most likely be dropped. Also, since the worker bees are the ones who fly out of the hive to fetch nectar, it was a reasonable assumption that the 'caste' attribute was irrelevant.

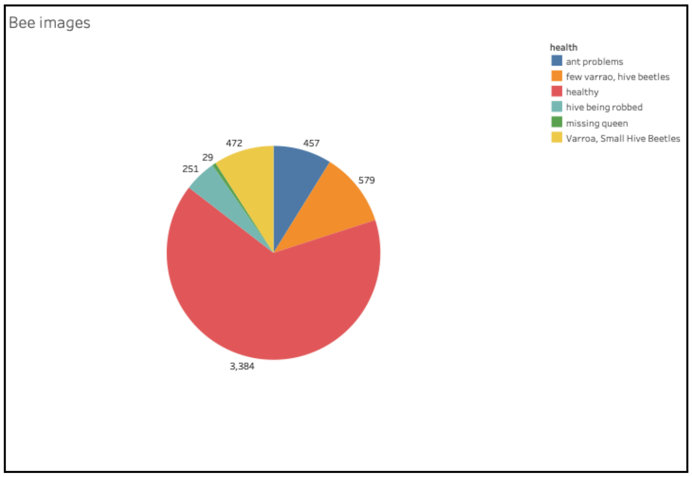


Figure 1. Division of labels for the bee image dataset.

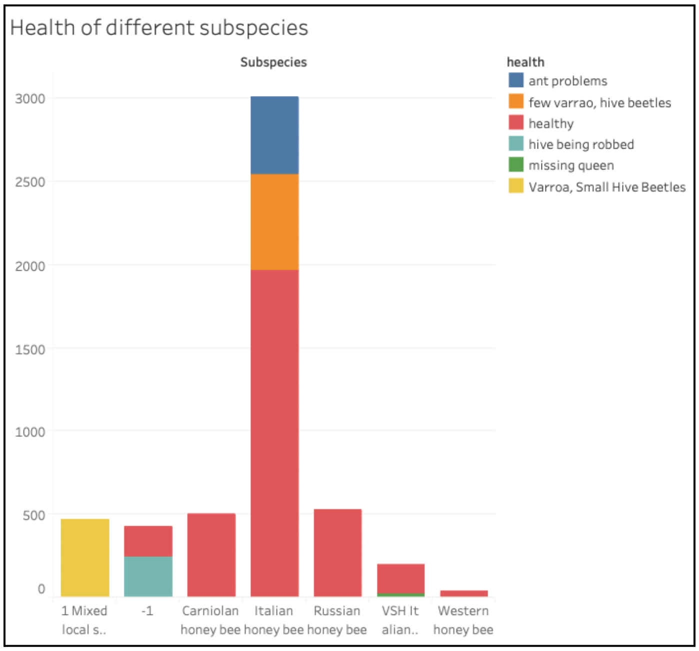


Figure 2. Different subspecies of honeybees in the dataset and their hive health. There is a subset of images where there is a mixed local stock or the subspecies is not known (-1).

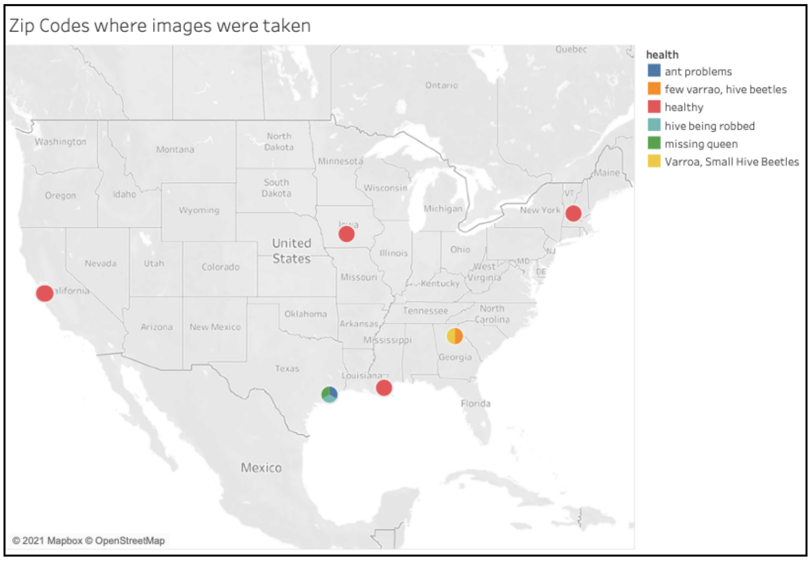

Zip Codes where images were taken
health
🟦 ant problems
🟧 few varrao, hive beetles
🟥 healthy
▨ hive being robbed
🟩 missing queen
🟨 Varroa, Small Hive Beetles
United States
Mexico
© 2021 Mapbox © OpenStreetMap

Figure 3. Location where the images were taken and the health of the hives in those images.

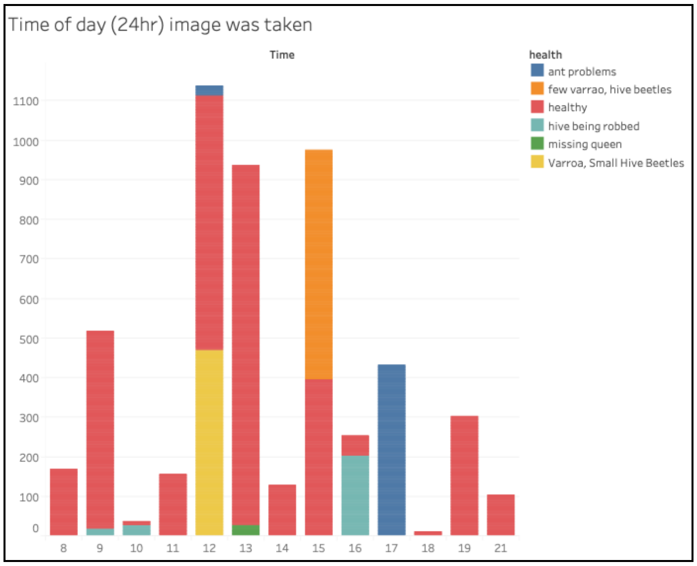


Figure 4. Distribution of the times when the images were taken on a 24-hour clock.

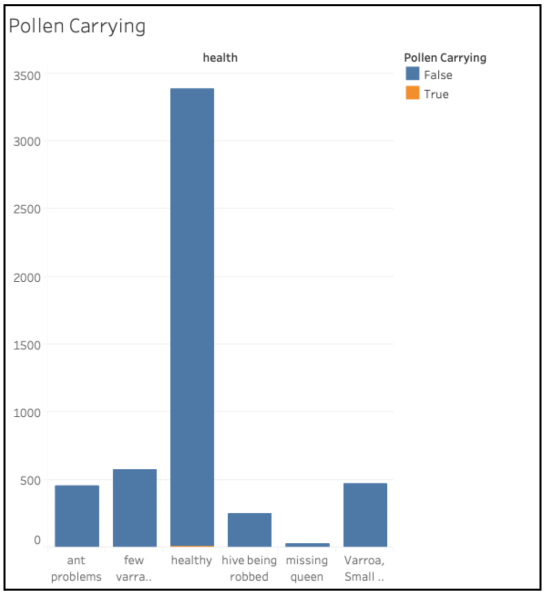


Figure 5. Bees of different health conditions who carry pollen. In the vast majority of images, the bees are not carrying pollen (except for 18 images in healthy).

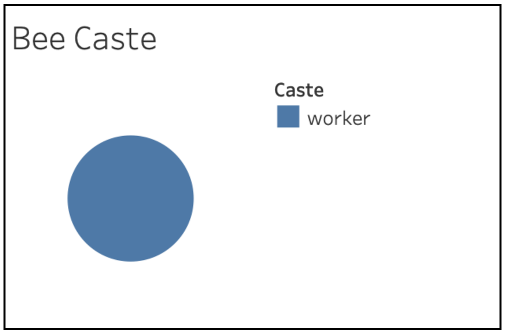


Figure 6. Bee caste

4. Main strategies of our project

To classify our images we use a convolutional neural network (CNN) as it is the most popular deep learning model for image classification, and past research has shown that it can far outperform other models. The TensorFlow library was used to develop the network. We tested our data using two CNN models.

The first prototype is a CNN written from scratch. It contains two convolutional layers. Each layer is followed by a max pooling layer. The first convolutional layer has 32 filters with kernel sizes of 3x3 and a rectified linear unit activation. The second convolutional layer is similar to the first, except it has 64 filters. Each max pooling layer reduces the dimensionality of the previous layer by two. The output layer has six units, which corresponds to the six labels for beehive health. Softmax activation is used to output prediction probability for each label. The model contained 1,223,622 trainable parameters.

For the second CNN, we use a pretrained model. TensorFlow's mobile net is chosen as it is a relatively small and efficient CNN[2]. To make the model suitable (as the model was originally trained for 1,000 different classes) the last six layers were modified so that there were six output classes. In total the model has 3,213,126 trainable parameters.

The entire dataset was divided into training, validation, and testing roughly in the ratio of 70:20:10. Parameters were kept the same across both models for consistency in comparison (Table 1).

Parameter	**Value**
Batch Size	10
Epochs	10
Learning Rate	0.0001
Loss function	Categorical cross-entropy

Table 1. CNN parameters.

The code was run from Google Colaboratory using GPU mode. For each epoch, the training time varied from 5 to 39 seconds for the first model and from 22 to 24 seconds for the pretrained CNN. Since this was our first attempt towards developing a model using CNN, we referenced a tutorial[3] as guidance.

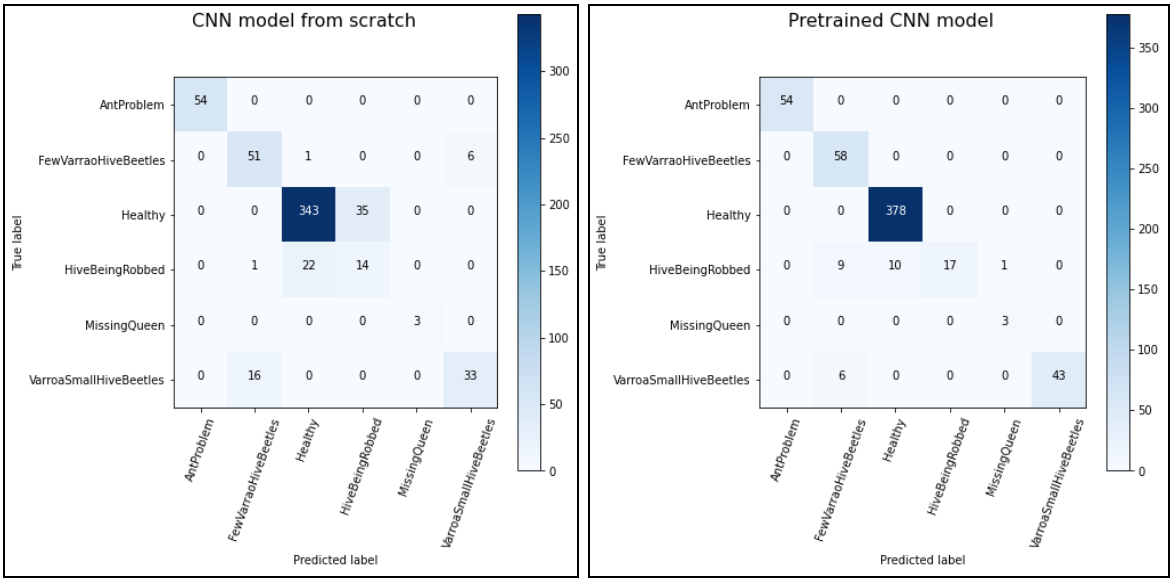
5. Evaluation strategy and results

Based on a given bee (input) image, our classifiers would predict the status of the hive. There were six predicted labels defined earlier. Some 3000+ images were used for training and an accuracy of ~99% was achieved for both the CNN models. On the validation set we were able to reach an accuracy of ~86% and 88% for the initial model and on the pretrained model, respectively. We then derived the confusion matrix for these multilabel classifiers and used the matrix to calculate precision, recall, f1 measure, receiver operating characteristic (ROC) curve, and area under curve (AUC) on the test dataset. We use the one-vs-the-rest(OvR) multiclass strategy to derive the ROC and AUC.

On the test dataset, our initial model achieved an accuracy of 86% whereas after using the pre-trained model our classifier had 96% accuracy. In the initial model, we observe a unique pattern from the confusion matrix (Figure 7). The vast majority of incorrect predictions for the label '*HiveBeingRobbed*' belong to the class label *'Healthy'* and vice-versa. This indicates that data for these two class labels may have some overlapping attributes as a result of which the model is unable to differentiate the two labels. Looking back into the definitions of a robber bee (which cater to *'HiveBeingRobbed'* status) and healthy bee, we make the below hypothesis.

Hypothesis: Robber bees usually fly towards a hive to destroy the hive and steal any stored nectar. These robber bees have shiny bodies with no pollen. On the other hand, healthy bees when leaving the hive also have shiny bodies and do not have any pollen. They will only have pollen on their bodies when they return to the hive after collecting nectar. Therefore, it is likely for our model to incorrectly distinguish between robber bees and healthy bees since the input data set does not contain any information about the direction with respect to the hive. That is, we cannot conclude whether a bee is flying into or out of the hive which would be essential to differentiate between robber bees flying towards a hive and healthy bees flying away from the hive.

The first model performs poorly for the *'HiveBeingRobbed'* label. The same is reflected in the ROC and AUC for this label where we see that *'HiveBeingRobbed'* has the lowest AUC (Fig 9). After using the pretrained model we do not see any significant improvement for this label (refer to the recall, f1-score values from Fig 8). However, the performance for the other class labels improves drastically. This implies that the pretrained model was able to improve the overall image classification problem and also strengthens our hypothesis that the issue with *'HiveBeingRobbed'* is at the input data level and not with the model.



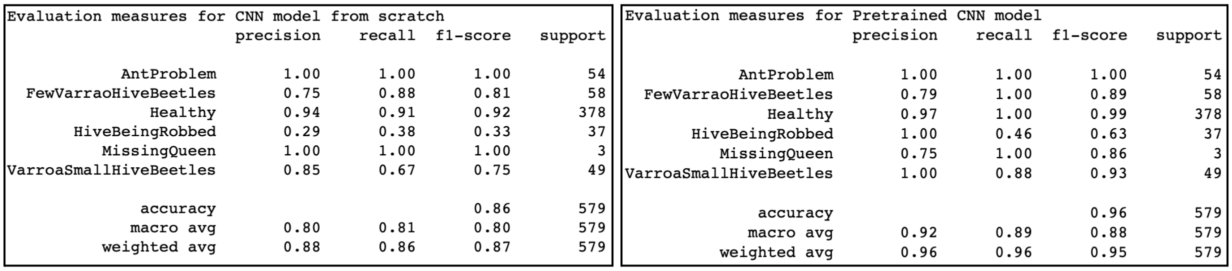
CNN model from scratch

True label \ Predicted label	AntProblem	FewVarraoHiveBeetles	Healthy	HiveBeingRobbed	MissingQueen	VarroaSmallHiveBeetles
AntProblem	54	0	0	0	0	0
FewVarraoHiveBeetles	0	51	1	0	0	6
Healthy	0	0	343	35	0	0
HiveBeingRobbed	0	1	22	14	0	0
MissingQueen	0	0	0	0	3	0
VarroaSmallHiveBeetles	0	16	0	0	0	33

Pretrained CNN model

True label \ Predicted label	AntProblem	FewVarraoHiveBeetles	Healthy	HiveBeingRobbed	MissingQueen	VarroaSmallHiveBeetles
AntProblem	54	0	0	0	0	0
FewVarraoHiveBeetles	0	58	0	0	0	0
Healthy	0	0	378	0	0	0
HiveBeingRobbed	0	9	10	17	1	0
MissingQueen	0	0	0	0	3	0
VarroaSmallHiveBeetles	0	6	0	0	0	43

Figure 7. Confusion Matrices



```
Evaluation measures for CNN model from scratch
                        precision    recall  f1-score   support

            AntProblem       1.00      1.00      1.00        54
  FewVarraoHiveBeetles       0.75      0.88      0.81        58
               Healthy       0.94      0.91      0.92       378
       HiveBeingRobbed       0.29      0.38      0.33        37
          MissingQueen       1.00      1.00      1.00         3
VarroaSmallHiveBeetles       0.85      0.67      0.75        49

              accuracy                           0.86       579
             macro avg       0.80      0.81      0.80       579
          weighted avg       0.88      0.86      0.87       579
```

```
Evaluation measures for Pretrained CNN model
                        precision    recall  f1-score   support

            AntProblem       1.00      1.00      1.00        54
  FewVarraoHiveBeetles       0.79      1.00      0.89        58
               Healthy       0.97      1.00      0.99       378
       HiveBeingRobbed       1.00      0.46      0.63        37
          MissingQueen       0.75      1.00      0.86         3
VarroaSmallHiveBeetles       1.00      0.88      0.93        49

              accuracy                           0.96       579
             macro avg       0.92      0.89      0.88       579
          weighted avg       0.96      0.96      0.95       579
```

Figure 8. Calculated Precision, Recall, F1-score for each class label. Accuracy on the test data was calculated for the entire model.

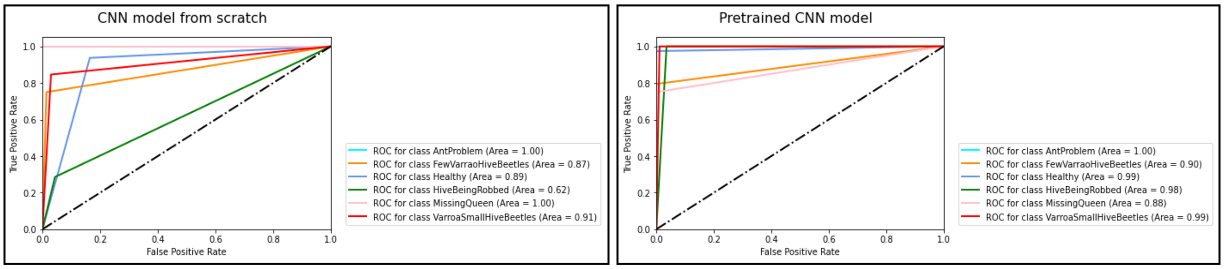


Figure 9. One-vs-the-rest strategy to derive ROC and AUC for each class label

6. Pros

- This model improves the ability to understand the hive health without checking the hive manually.
- Using the pretrained CNN, the final model has 96% accuracy on test data.

7. Cons

- Dataset currently has no way to identify the direction of flight of the bees - towards or away from the hive. The direction of flight of bees helps to identify the difference between the robber and healthy bees.
- The original batch of images was extracted from still time-lapse videos of bees and each frame of the video was subtracted against that background to bring out the bees in the forefront. As we are not bee experts, we were unable to tell the difference in the photos as image quality is inconsistent. It would be beneficial to collaborate with apiarists to help us in this.
- As CNN's are black-box models, the impact of different attributes on classification probabilities is unknown. Therefore, it is not known whether our models are making predictions based on the bee itself or some other attributes in the background.
- The dataset is biased because the source of varroa images is from the same location. Since they are from the same location, the model might train based on the other common feature of the image from the background rather than the bee.
- MissingQueen data has very few images, and all are most likely for the same hive, therefore the model could be learning based on the image background.

8. Conclusion

The first prototype of our CNN model presented during the presentation fell short of achieving the desired results and was giving an accuracy of 86% on the test data set. During our demonstration, we proposed to use additional techniques to try and improve the model's performance. Following that route, we tested our classifier by adding a pre-trained model for comparison. This bumped up the overall model's performance by improving both the validation and test data results. We were able to achieve an accuracy of 96% on the test data.

However, we are still struggling to achieve proper results for the '*HiveBeingRobbed*' class label. This would probably require gathering more information on how to identify robber bees and enhance our input data set to accurately capture/identify those features.

In conclusion, while there are limitations to the dataset and machine learning model used in this project (as specified in the cons section above), it does have the potential to have a large impact in the apiarist community. An accurate prediction model could greatly reduce the hive maintenance time for beekeepers and constant automatic monitoring of the hive would ensure the beekeeper could act quickly when there is a probability that the hive is unhealthy. Additionally, the collection of this data could lead to further research and discovery in the life of bees. For example, the effect of climate and seasonality could also be monitored and analysed.

9. Future Directions

Currently, we manually divided the dataset into the train/validate/test data sets. In the future, we would like to shuffle dynamically to ensure our model renders the same level of performance and also investigate wrongly classified images and their metadata to find possible correlations.

It would be beneficial to gather more images from a variety of hives across the United States, this would decrease the potential bias in our dataset for classification based on the image background.

10. A summary of contributions from each team member

- Repository setup and data visualization: (Fiona+Swetha)
- Data Setup: (Shilpika)
    - Segregation of images based on metadata
    - Create train, validate, and test set
- Classifier model design: (Fiona)
- Generating Confusion Matrix for Multilabel classification: (Swetha)
    - Identifying TP, FP, TN, FN
- Deriving other measurements (Shilpika)
    - Precision, Recall, F1 measure, Accuracy
    - OvR multiclass strategy for ROC and AUC

11. References

1. https://www.kaggle.com/jenny18/honey-bee-annotated-images
2. https://arxiv.org/abs/1704.04861
3. https://deeplizard.com/learn/video/RznKVRTFkBY

Proposal:

Team Members:
What is the main problem that your team is going to address in this project?
Why are you interested in this problem?
Where or how will you obtain the data?

How do you plan to work as a team?

As a team, we will work on

As a team, we will work on

- Collecting raw data from sources, then filter, organize effectively as possible.
- Exploring different methods and using different data mining tools to process the data and get outcomes.
- Major tasks we do as a team are Data Collection, Data preprocessing, Data Analysis, Data Visualization, Data Modeling and Data Analytics.

Title (Tentative**):** Given a location in SF how suitable it will be for Urban Beekeeping

- Available stats show that urban beekeeping does well over rural beekeeping
    - Could be due to warmer temperatures
    - Or other modes of pollination
    - Need to collect data as to what facilitates urban beekeeping over rural's.
- When a user enters a specific location our system would calculate the followings to determine the area's prospects for urban beekeeping:
    - Area available (larger the better)
    - Height from sea-level (need to figure out how this impacts)
    - Flower plants or trees nearby (advantageous) - nearby gardens, parks, vineyards
    - Nearby school area (may be a deterrence)
    - Noise decibel (need to figure out how this impacts)
    - Humidity levels, rainfall (need to figure out how this impacts)
    - Min, Max and Average temperatures
    - Wind velocity
    - Sunlight
    - Probable periods for overwintering (longer periods would be a deterrence)
    - Probability to making new plantations nearby
    - Volunteers to manage and administer the beekeeping process, their age, experience could matter. We could take this field as input from users in case they want to do personal beekeeping.
    - Bee predators would be a deterrence
    - Existing nearby beehives (natural or man-made)
    - Government-aided/Corporation-owned locations or fundings. Guidelines for beekeeping and audits. Tax benefits or other benefits if any.
    - Available trainings/bootcamps for beekeeping
    - Food consumption patterns of the locality might impact what plantation is present and whether it is bee friendly or not. Ex: if people consume more fish

and meat than fruits and vegetables then chances of plant/tree plantations near by would be slim.

Available Datasets:

1. https://beescape.org
   Description: Bees fly great distances - sometimes several kilometers - to collect nectar and pollen to feed their young. Beescape allows a user to select a specific location (an apiary site, a garden, a farm) and get information about the number of floral resources, overall toxic load of applied insecticides, and the availability of nesting habitat for wild bees in the landscape surrounding the selected location.
2. https://www.gbif.org/dataset/c4a2c617-91a7-4d4f-90dd-a78b899f8545
   Description: USBombus database, bee numbers collected in the United States (one location in San Francisco) from 2007-2010
3. Weather data for San Francisco over 30 years:
   https://www.meteoblue.com/en/weather/archive/export/san-francisco_united-states-of-america_5391959 -> Must pay for data
   https://www.wunderground.com/history/weekly/us/ca/san-francisco/KSFO -> no file download
   https://w2.weather.gov/climate/index.php?wfo=mtr-> no file download
4. GIS datasets
   https://guides.lib.berkeley.edu/gis

Relevant papers:

1. "Urban areas as hotspots for bees and pollination but not a panacea for all insects"
   https://www.nature.com/articles/s41467-020-14496-6
   Gives reasons why insect conservation is important and evidence that bees (but not all insects) do better in urban vs. rural.
2. "Summer weather conditions influence winter survival of honey bees (*Apis mellifera*) in the northeastern United States"
   https://www.nature.com/articles/s41598-021-81051-8
   Used 3 years of Pennsylvania beekeepers' survey data to assess the importance of weather, topography, land use, and management factors on overwintering mortality at both apiary and colony levels, and to predict survival given current weather conditions and projected climate changes.

Issues:

1. Difficult to assess relative impact of factors on bee population.

Possible Term Project:
Proposal: Optimizing San Francisco for urban beekeeping

Team Members: Shilpika Banerjee, Swetha Govindu, Fiona Senchyna

What is the main problem that your team is going to address in this project?
We are going to address the issue of optimizing urban beekeeping in San Francisco. Many factors affect the survival of beehives, including temperatures, rainfall, availability of flowering plants etc. We wish to collect data on all of these factors as it relates to the city of San Francisco and use it to pick out the best regions in the city for beekeeping. We also wish to provide it as a service for users, allowing them to enter in their address in the city and we will generate a report on its potential for beekeeping.

Why are you interested in this problem?

We are interested in this problem because bees (along with many other species) have been in great decline over recent decades, in large part due to human activities. Urbanization can reduce biodiversity; however, some studies have shown that honeybees survive better in urban areas compared to rural[1]. Additionally, urban beekeeping has risen in popularity due to the increasing awareness of their decline. Bees are vital to plant pollination, which humans rely on for food. We wanted to help the urban beekeepers of San Francisco by analyzing the full potential of the city for urban beekeeping.

[1] Theodorou, P., Radzevičiūtė, R., Lentendu, G. *et al.* Urban areas as hotspots for bees and pollination but not a panacea for all insects. *Nat Commun* 11, 576 (2020). https://doi.org/10.1038/s41467-020-14496-6

Where or how will you obtain the data?

We found a project (https://data.world/finley/honey-bees-and-apiculture) that contains the following datasets on San Francisco:

1. Community Resiliency Indicator System for the city of San Francisco contains environmental data for 37 regions in the city.

2. SF CA Land Use contains information on the land use in the city (i.e., residential, business, open space)

3. SF CA Park and Open space contains information on the parks and gardens in the city.

(still searching for datasets…)

How do you plan to work as a team?

As a team, we will work on the followings:

- Collecting raw data from sources, filtering the data, and organizing it effectively.
- Exploring different ML models and using various data mining tools to process the data to achieve the required outcome.
- Major tasks will be data collection, data preprocessing, data analysis, data visualization, data modeling and data analytics.

<u>Point of Contact for Beekeeping Insights:</u>

Prof. John Hafernik: Work on Zombie Bees https://news.sfsu.edu/releases/sf-state-researchers-confirm-first-zombie-bee-sightings-southern-us

Prof. John Hafernik: http://biology.sfsu.edu/people/john-hafernik
Retired

Prof. Gretchen LeBuhn:
https://news.sfsu.edu/biologists-design-method-monitor-global-bee-decline

Action Items:

Fiona - to draft emails for professors. Swetha and Shilpika to review the email and we can send that out on Monday 15th March.
Team - Look into Kaggle, biology journals and google to find relevant data points.

Reconnect on Wednesday, March18th (11am PST, daylight saving is on) to check if any response was received from the professors and plan the steps ahead.

Draft e-mail to Dr. Gretchen LeBuhn:

Dear Dr. LeBuhn,

I hope you are well.

My name is Fiona. I, along with Swetha and Shilpika (cc'd here), are Computer/Data Science SFSU graduate students who are doing a group term project in our Data Mining class (CSC869).

After reading about global bee decline and research suggesting bees have the potential to do well in urban areas (link), we are interested in doing a data mining project related to urban beekeeping in San Francisco. Specifically, we are interested in developing a program that looks at the layout and weather conditions of San Francisco, and finds the best spots to place hives (e.g., proximity to community gardens, weather etc.) However, we are new to the subject and are unsure which environmental factors are most important when determining whether a bee colony will thrive. For example, is there a certain temperature and/or humidity bees do best in? How close must flowering plants be to the hive for the bees to be able to find them, and how much plants are needed to feed a colony?

We found that you have a lot of expertise in this area and were wondering if you could give us some information on the subject or point us to important sources that would help us in taking up this project. If you deem suitable we could set up some time to connect with you over zoom as per your availability.

Thank you.

Regards,
Fiona

3/18:

Links: https://www.nature.com/search?journal=srep&q=bee

- Cost of urban beekeeping could be between $300-$500 for the first time
- The beekeeping model should be 16inches from the ground to help with ventilation and getting out the moisture.

Dear Prof. Yang,

Upon further research on data for our project and a reply to our email from Professor John Hafernik, we have all decided that the scope of our project is too large and has issues beyond lack of data (e.g., there is no direct data as to what conditions bees survive best in and we do not see a way we could validate our model since there are no datasets on where honeybees are in San Francisco and how much their location influences how healthy they are).
Given this, we wish to change the focus of our project slightly. Instead of predicting where bees survive best, we wish to use a Kaggle dataset (https://www.kaggle.com/jenny18/honey-bee-annotated-images) to predict whether the hive is healthy or not based on an image of a bee leaving the hive. Like our previous idea, this is a problem that could be used to help beekeepers maintain a healthy hive, and thus help prevent their decline. The narrower focus of this project would allow us to spend our time focusing on how best to solve the problem, rather than what data to use and how to validate it.

Thank you,

Written update 1 page (Due: 4/2/2021, 11:59pm):

1. How has the team been communicating with each other (e.g., email, discord, video chat)?
   - Discord
   - E-mail
   - Zoom
2. What have you completed so far and by which team member(s)?
   - Researched datasets related to our problem formulation.
   - Communicated with the professor and experts in the field.
   - Learned more about the domain of our problem.
   - Narrowed the focus of our problem.
   - Visualized and analyzed the dataset.
3. What tasks are you currently working on and by which team member(s)?
   - Set up a shared google co-laboratory notebook, where we will write our code.
   - Became (more) familiar with convolutional neural networks for image classification.
4. What do you plan to do next?
   - Preprocess the data – handle missing values (Fig. 2) and unbalanced dataset (Fig. 1) and remove columns that do not give any information on the dataset (Fig. 6).
5. Any questions/challenges you have encountered or foresee?
   - Unbalanced dataset (Fig. 1)
   - Effect of attributes, such as different subspecies (Fig. 2), locations and times the images were taken (Fig. 3 and 4, respectively) on image classification.
   - Effect of a small minority of bee images carrying pollen (Fig. 5) – should these images be included? should we manipulate the images (e.g., crop, rotate) so that we increase the number?
6. If needed, how would you like the instructor to help your team?

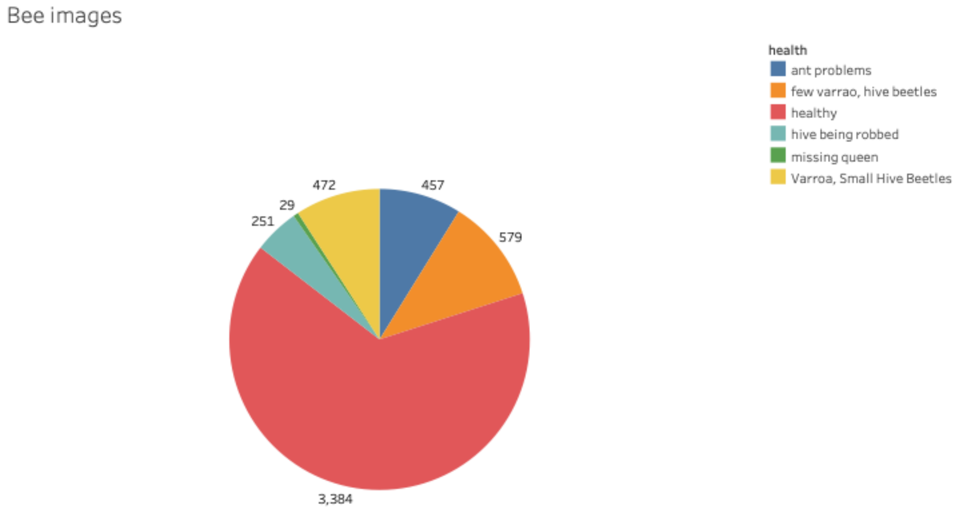


Figure 1. Division of labels for the bee image dataset.

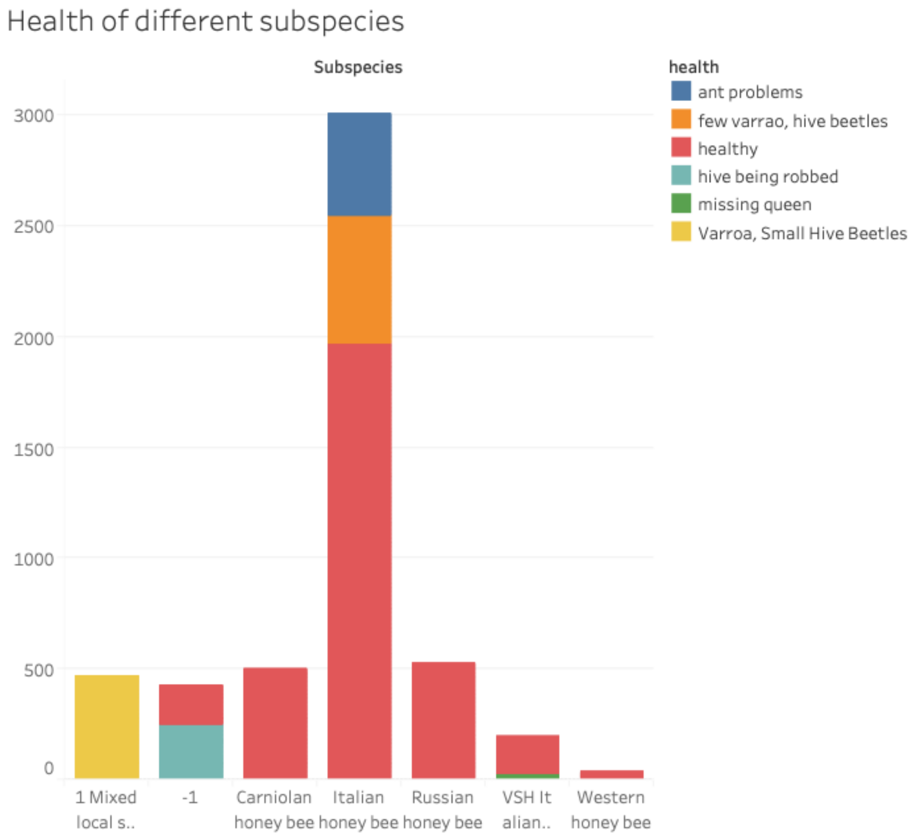


Figure 2. Different subspecies of honeybees in the dataset and their hive health. There is a subset of images where there is a mixed local stock or the subspecies is not known (-1).

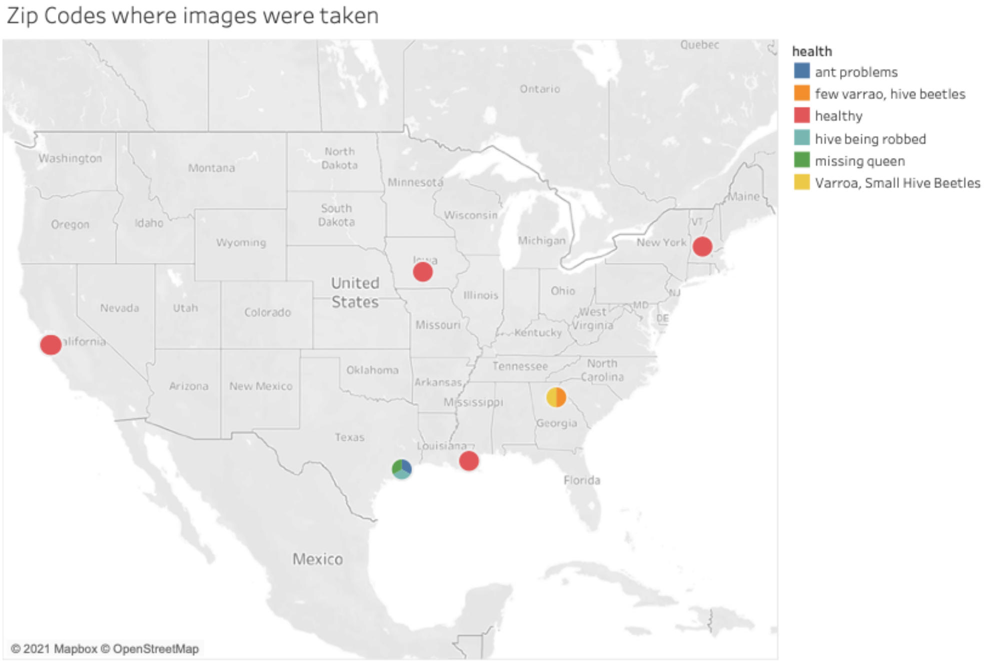


Figure 3. Location where the images were taken and the health of the hives in those images. Unknown if this will affect image classification.

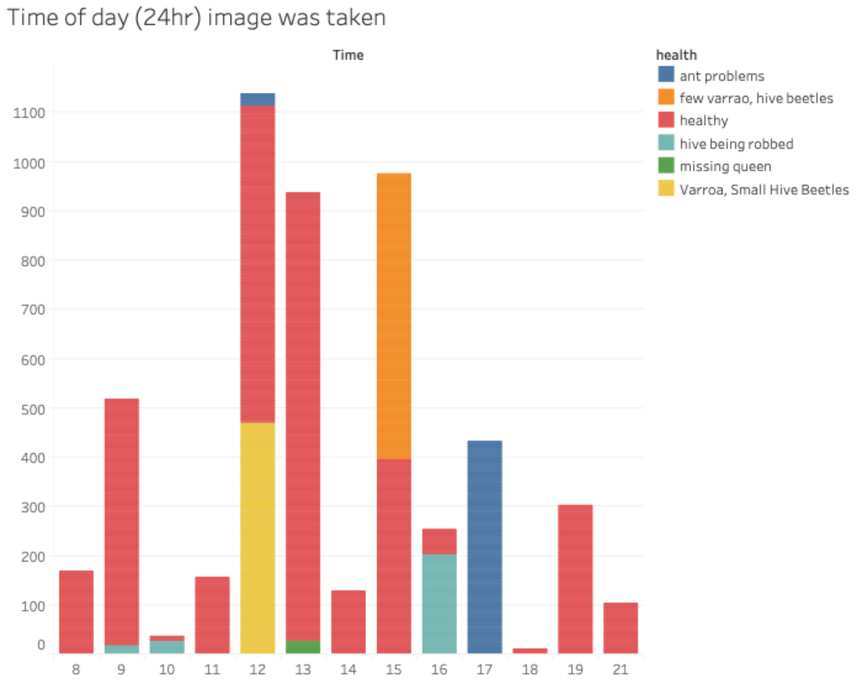


Figure 4. Distribution of the times the images were taken on a 24-hour clock. Unknown if this will affect image classification.

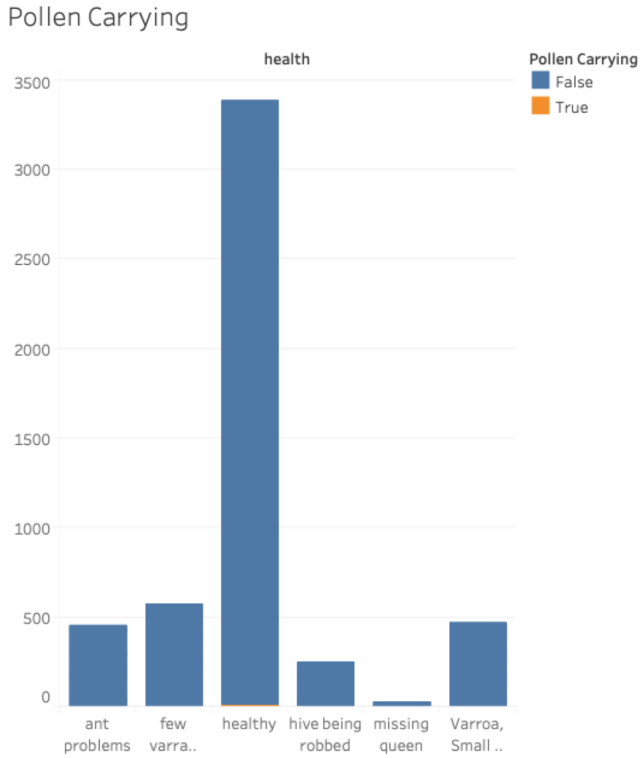


Figure 5. Bees of different health carrying pollen. Vast majority of images the bees are not carrying pollen (except for 18 images in healthy).

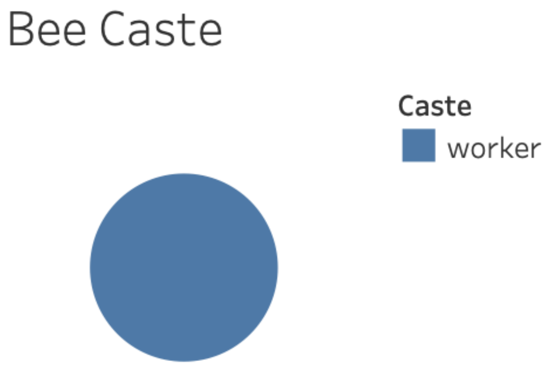


Figure 6. Bee caste. All images in this dataset, regardless of hive health, are images of worker bees. Therefore, this column can most likely be dropped.

Google colaboratory notebook
https://colab.research.google.com/drive/17yUAFvvEagGdmhvWeiVmQb0cmxFkRZFw?usp=sharing

4/12 Meeting

Overall aim:

Predict entry in the ‘health’ column of ‘bee_data.csv’ based on the corresponding image (given in ‘file’ column) using a convolutional neural network (CNN).

First iteration aim:

Follow basic image classification as seen in the tutorial (https://www.youtube.com/watch?app=desktop&v=qFJeN9V1ZsI, ~1:01:00).

Steps:

1. Image processing (a bit different to tutorial as have to use csv file to find label):
   a. Create train, validate and test folders.
   b. Add images into subfolders in training/validation/testing folders, where subfolder names correspond to classification names. Images should be added to training/validation/testing randomly at a ratio ~70/20/10 for each classification.

2. Use Keras with TensorFlow to build a convolutional neural network (CNN). Try default parameters as per the tutorial.

3. Analyze results (confusion matrix?)

   a. Does the network classify certain label(s) better than others?
   b. Are misclassified images evenly represented over different subspecies, pollen carrying etc.

Expectations for next iteration:

Based on results of previous iteration, images could be manipulated so the number of images in underrepresented classification labels are increased, parameters of CNN model could be changed, could add other attributes to input etc...

Fiona: Set up github repository and initial google colab notebook. (today/tomorrow)(
Shilpika to start with step 1 of the first iteration (Apr Wed14th-Sat 17th) tentatively by first half
Fiona to work on step 2 (April 14th - Sun 18th)
Swetha to work on step 3 (April 18th - April 22nd)

Next Meeting on 19th April, Monday 5:30 PM