Honey Bee Image Classification: Predicting Subspecies and Health Status

Aditya Vijay
GA DSI 7/25 2022

The Problem

- Honey Bees provide upwards of $15 billion towards agricultural production in the US each year, and are essential for healthy and abundant crop production, as approximately ⅓ of all food people eat comes from bee pollinated crops [1]

- Beekeepers in the US reported losing over 45% of managed hives per year as of 2021[3] due to stressors including Africanization (blending of subspecies leading to increased aggressiveness), pests (*varroa destructor*, hive beetles, etc), and Queen deficiencies

- **Beekeepers need advanced methods for preemptively identifying and treating weaknesses in the hive**

The Solution

- **Build a web application for beekeepers to predict the subspecies and the health status of bees based on their images**

- To achieve this, Convolutional Neural Networks (CNN) will be trained on the Honey Bee Annotated Image Dataset, hosted on Kaggle, to carry out a multi-class classification with respect to the target variables of subspecies and health

- Models will be compared using the accuracy metric, and will be tuned for robust classification

- The Streamlit framework will be used to deploy the CNNs to a web app, where users can upload their own images for prediction

The Data

- 5172 images, all with matching rows in the annotated dataset

- Annotated data fields include:
  - Health, subspecies, pollen status, bee caste, location
  - Image capture time, image filename

- Image dimensions vary between files
  - Image dimensions for each file are appended to data

- The features for prediction are the numerical pixel data read from each image after RGB conversion

The Targets: Subspecies

- Why does predicting the subspecies matter? Blending of pure-bred stocks with sub-optimal subspecies can negatively influence a hive, and disrupt cohesion.

- The presence of bees from a subspecies that your hive does not belong to can indicate a weak hive or a weak queen

- Sometimes, subspecies blending is benign, but incompatible genetics can lead to sterile male offspring, signaling for the workers to kill their queen and raise a new one[4]

- The subspecies this model is trained on are:
    - ‘1 Mixed Local Stock 2’ (a specific mix of subspecies), Carniolan honey bees, Italian honey bees, Varroa-mite sensitive Italian honey bees, Russian honey bees, Western honey bees, and ‘Unknown’, the grouping for bees without subspecies annotations

The Targets: Subspecies (examples)

Russian Honey Bee

Western Honey Bee

Carniolan Honey Bee

Italian Honey Bee

The Targets: Health Status

- Why does predicting the health status matter? The health status annotations record whether or not a bee or it's hive is afflicted by one of many parasites or conditions.

- The health statuses this model is trained on are:
  - Varroa, Small hive beetles, Ant problems, Few Varroa, Hive Beetles, Hive being robbed, Missing queen, and Healthy.

- The single most destructive parasite to honey bees is the Varroa mite, or *varroa destructor*, which is a parasite that leeches nutrients and is responsible for the most bee death[5]. Hive beetles lay eggs inside hives and destroy comb, brood (developing offspring), and honey cells, while raiding the nutritious pollen stores that are essential for surviving winter

- Additional hive issues include ant problems, which can disrupt functioning and deplete honey stores, missing a queen which can quickly lead to colony collapse if a new queen is not reared immediately, and having the hive be robbed by wasps, or other bugs/animals

The Targets: Health Status (examples)

Varroa Mite

Hive Beetle

Aftermath of a hive Robbery

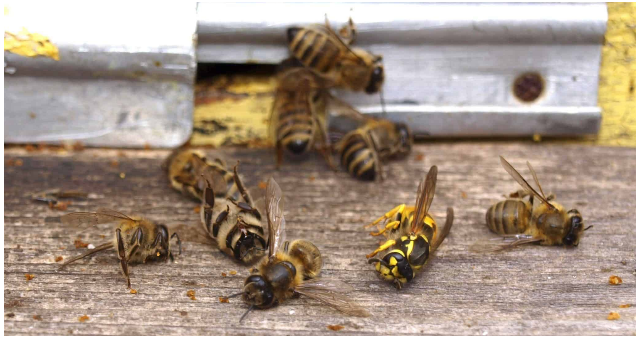
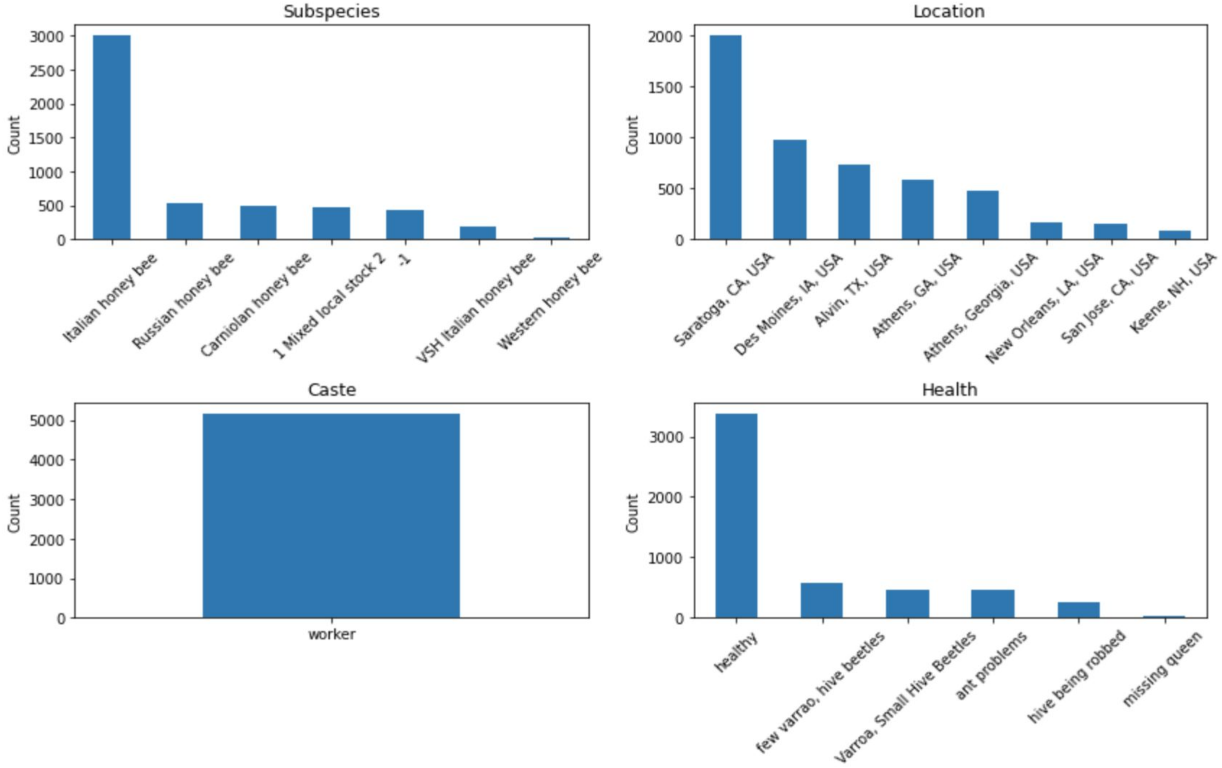
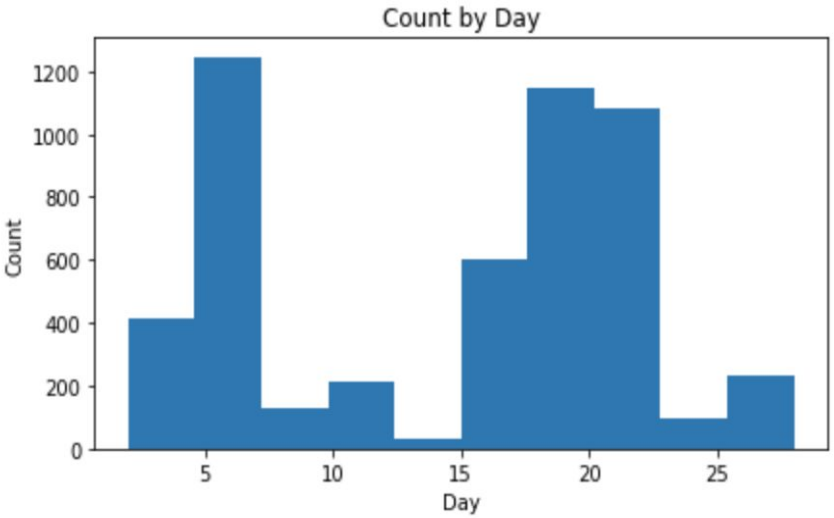
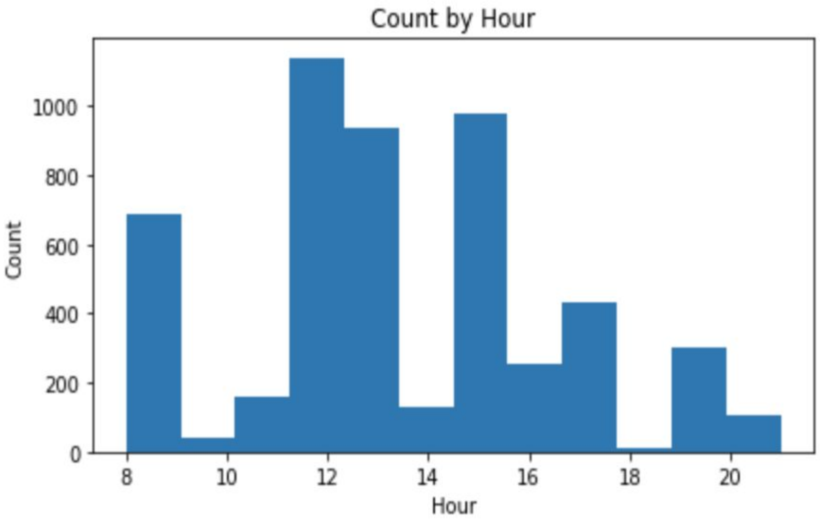


Image Count Distributions



Times of Image Capture





Times of Image Capture

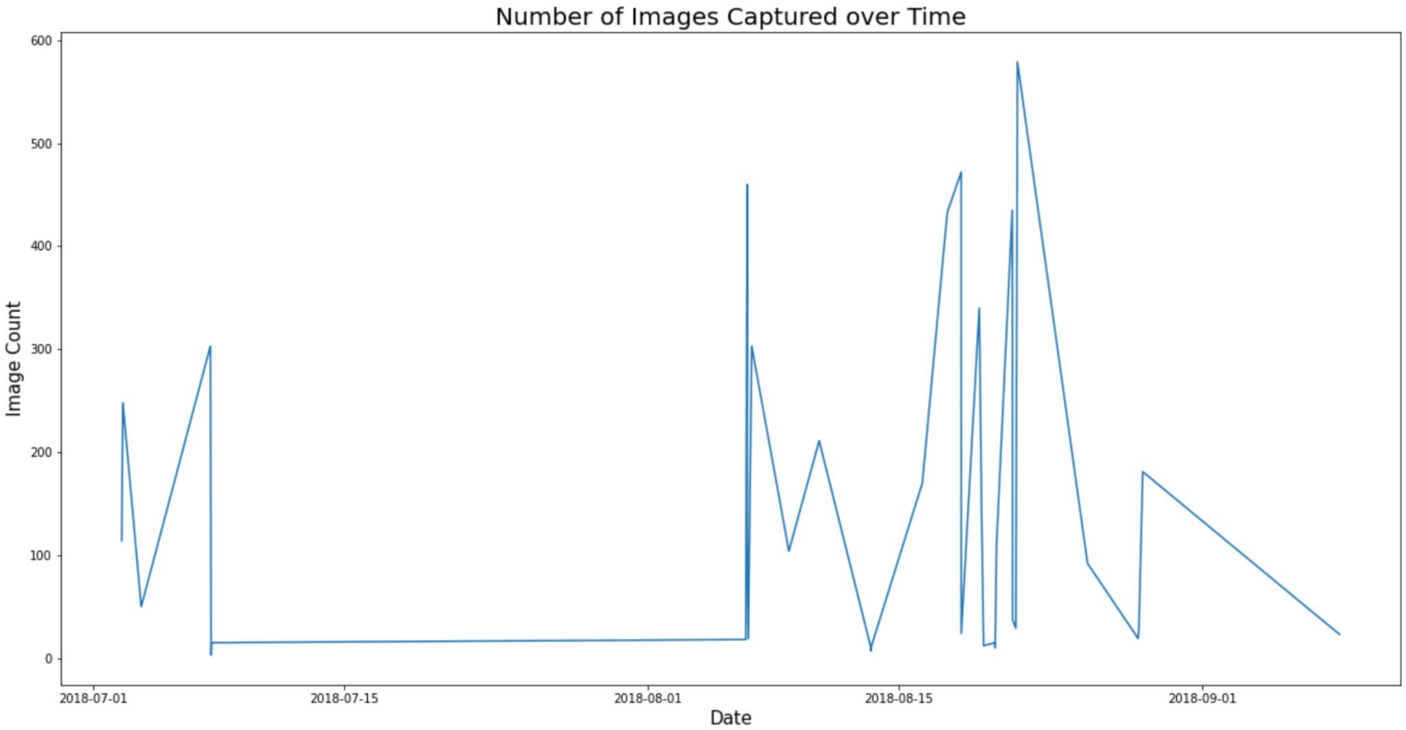
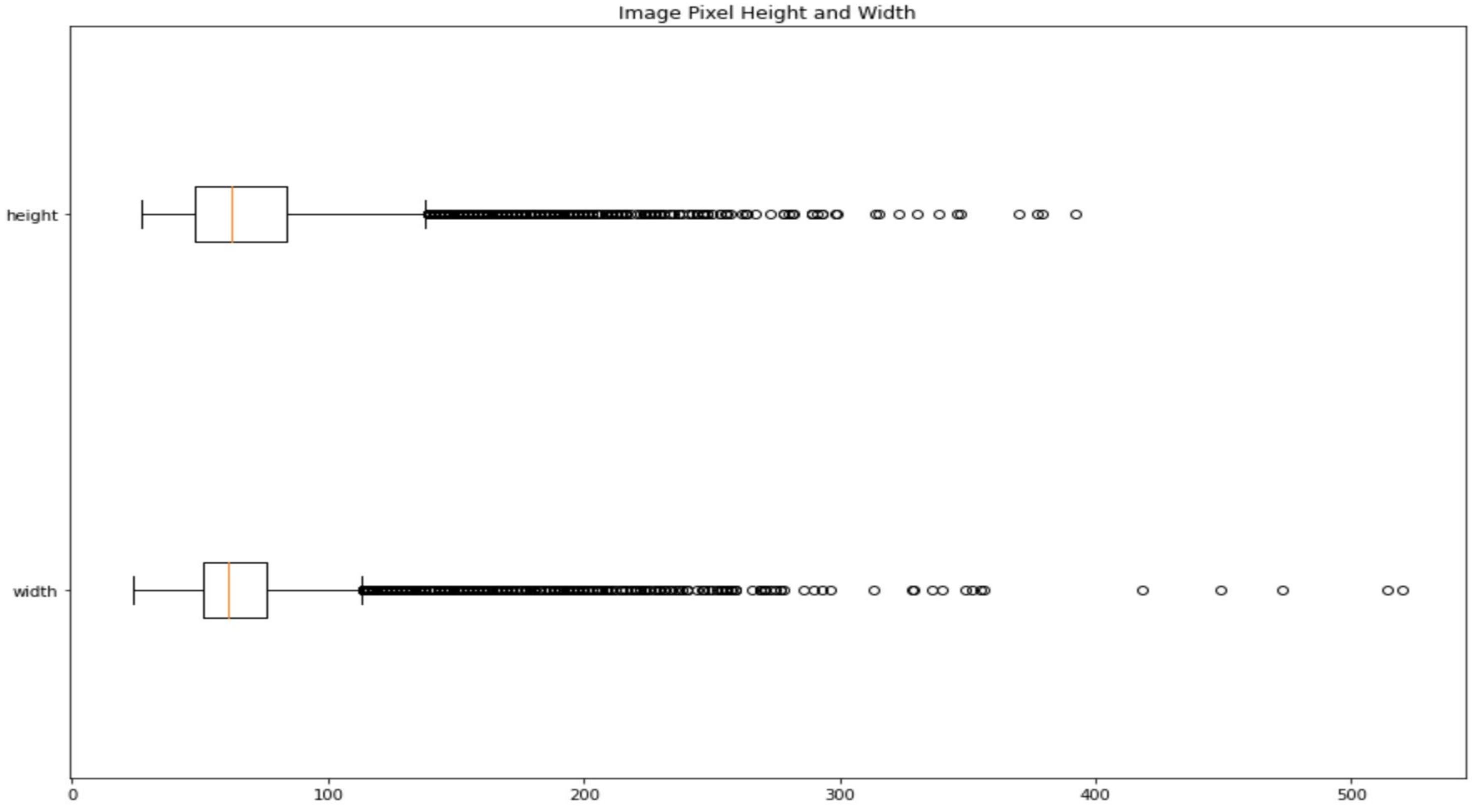
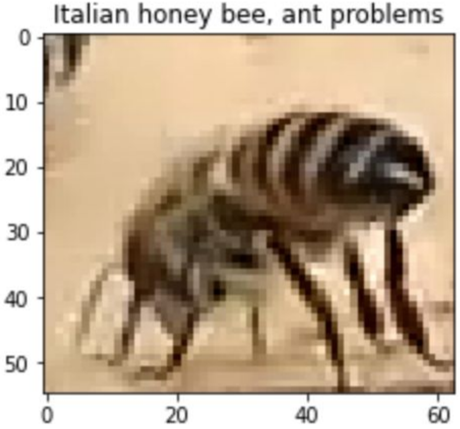


Image Dimensions



Example Images



Pre-Processing

- Images are converted to RGB and re-sized to 100x100 px, with 3 color channels.

- Using Keras's ImageDataGenerator, images are augmented in real-time through horizontal and vertical flips, random rotations, and random zooms

- The images are originally extracted from stills of a video, with the average backgrounds subtracted to accentuate individual bees

Model Architecture

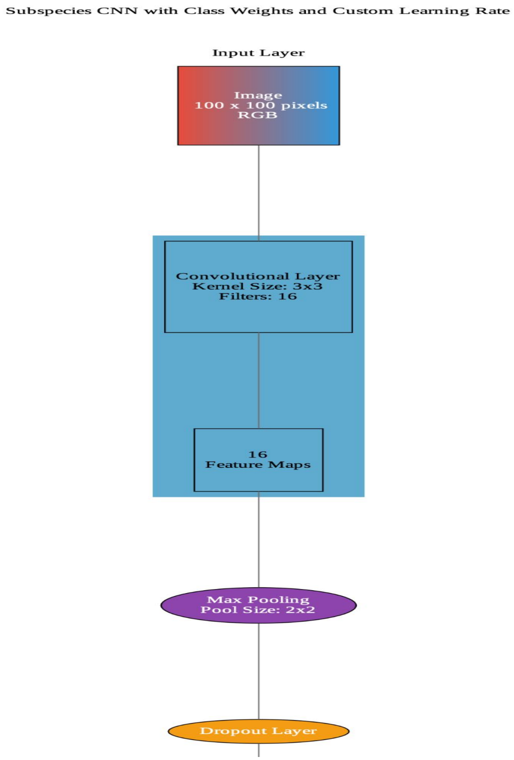
- Multiple CNNs were built, starting from a simple architecture and incorporating image augmentation, dropout layers and learning rate modification. Three models were saved per target for 6 in total.

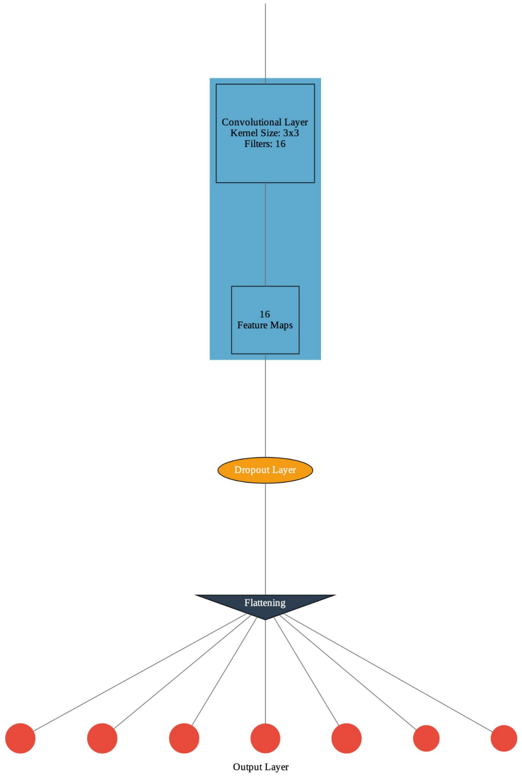
- Class imbalances are resolved by passing a dictionary with class weights to the class_weight attribute during fitting.

- Network topography:

  [(1 Conv + 1 MaxPool) + 1 Dropout + (1 Conv + 1 MaxPool) + 1 Dense]

Visualizing Model Architecture





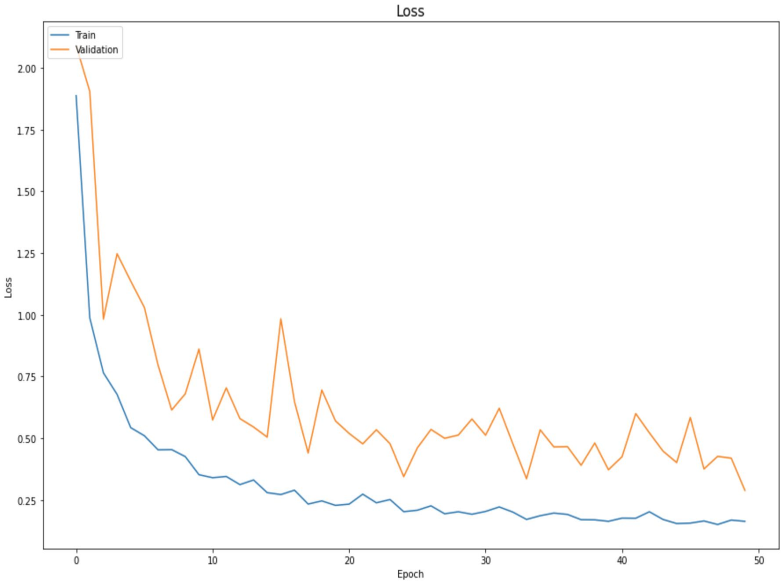
Model Evaluation

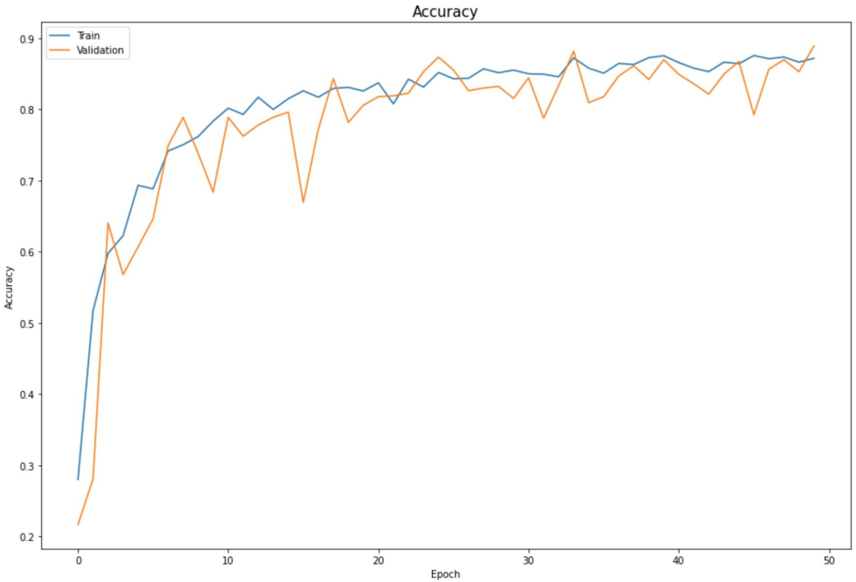
- Using models with class weighting and learning rate modification, the CNNs achieved:
  - 86.09% accuracy when predicting health, an improvement from the baseline of 65.43%
  - 89.27% accuracy when predicting subspecies, an improvement from the baseline of 58.16%

- Dropout layers regularize the data to prevent overfitting

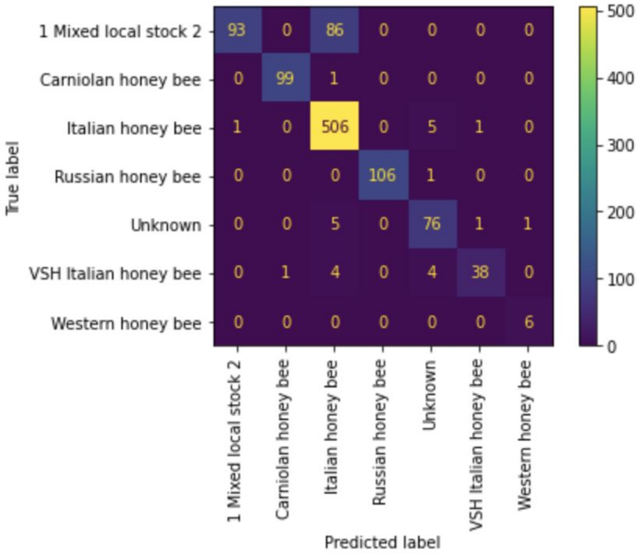
- The learning rate is gradually decreased to allow for finer tuning as the model progresses through each epoch

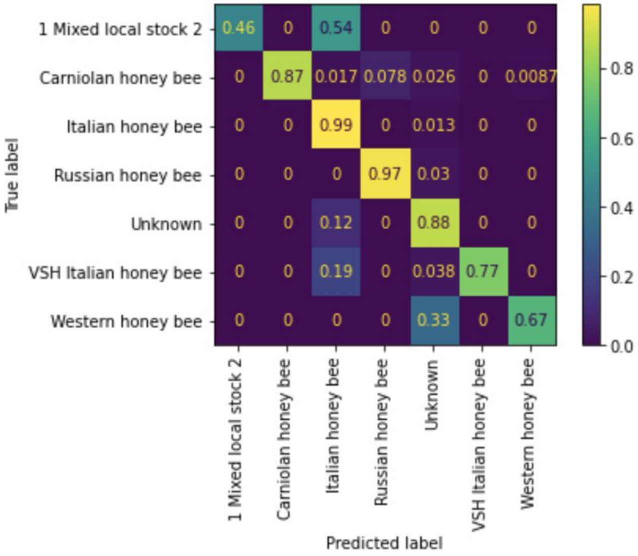
Model Evaluation: Subspecies



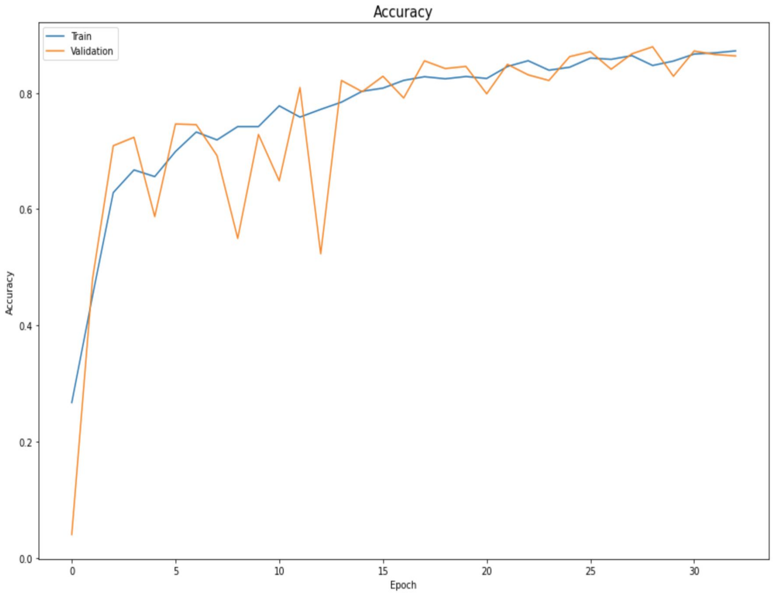


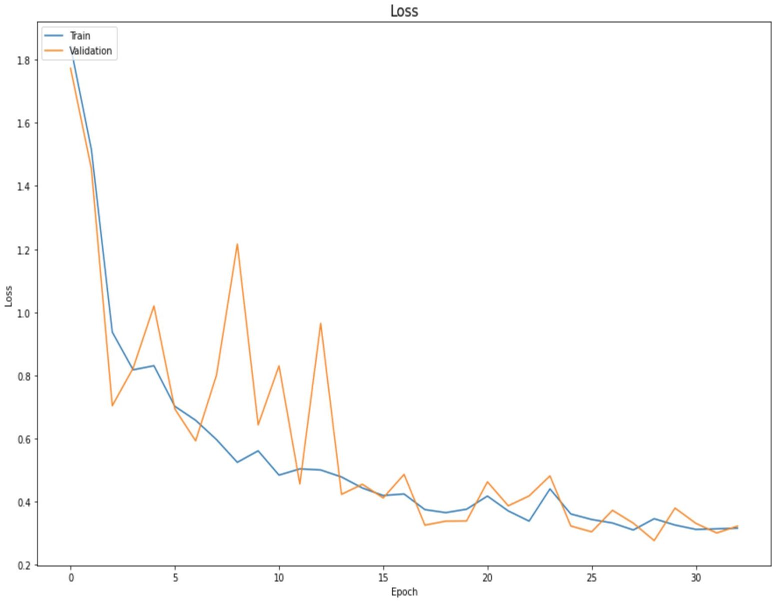
Model Evaluation: Subspecies (cont)



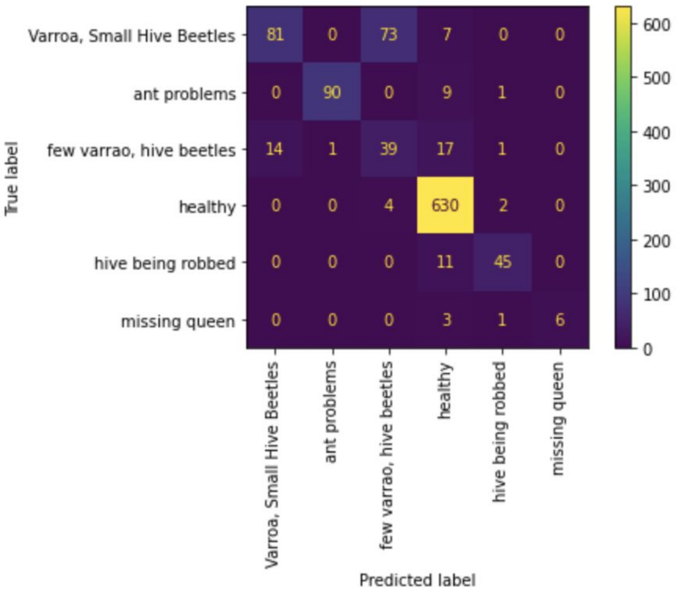


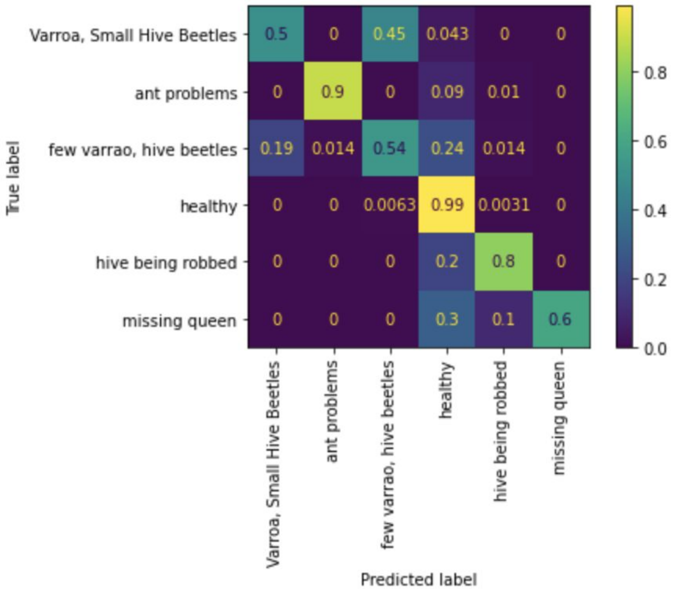
Model Evaluation: Health



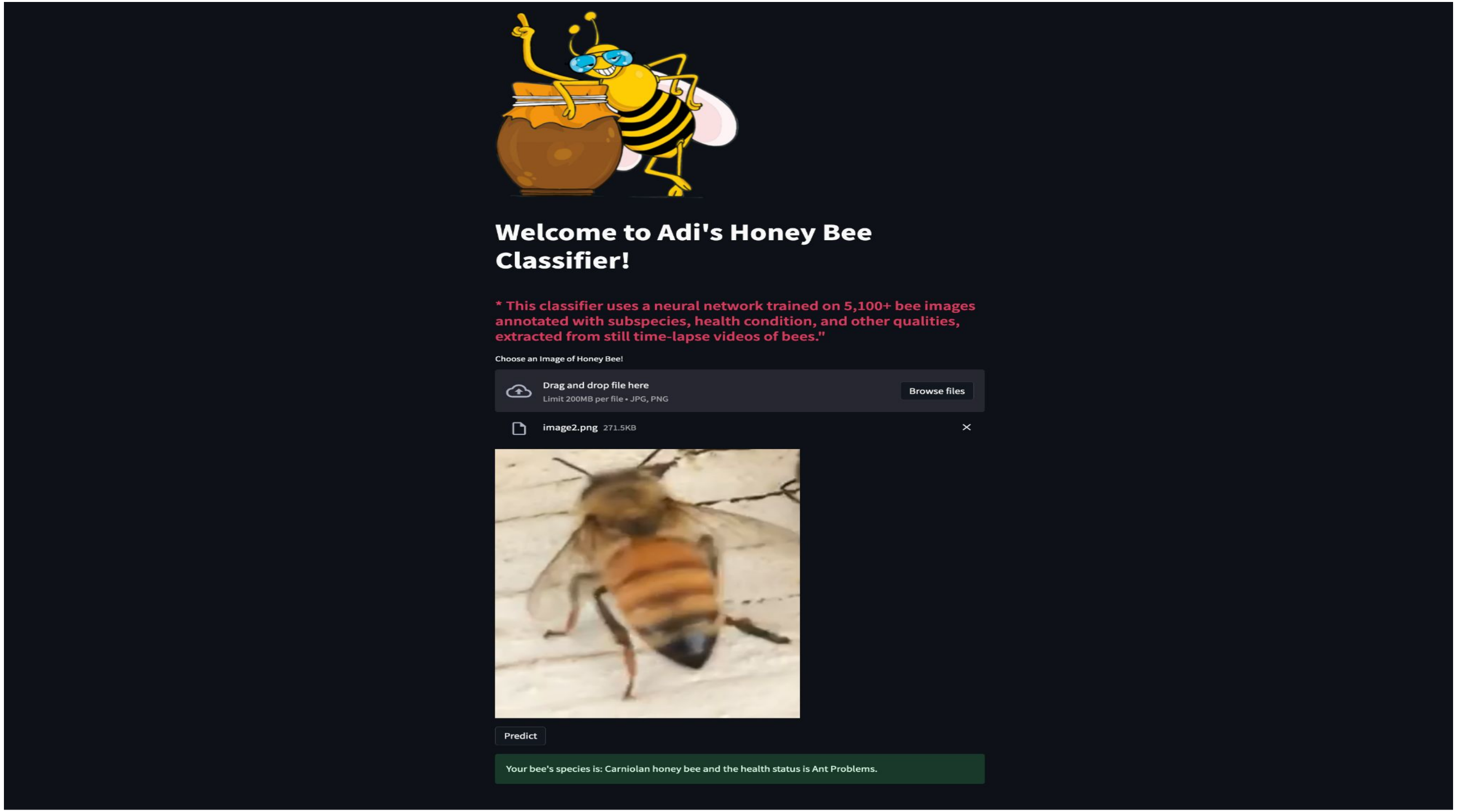


Model Evaluation: Health (cont)





Web App Implementation



Conclusions

- **The CNNs were successful in training to perform better than the baseline accuracy**
  - **86.09% testing accuracy when predicting health, 89.27% testing accuracy when predicting subspecies**

- Weakness include misclassification of images with backgrounds that are similar in color to the bee in the image

- The implementation of these models in Streamlit provides a simple interface for generating predicted subspecies and health status from an image

Recommendations

- Training on a more robust dataset can increase model performance
- Transfer Learning can be used to modify a pre-existing model to take advantage of highly accurate image classification models
- Developing an additional CNN for image segmentation, or using a computer vision library can further automate predictions by processing video to extract images

Sources

1. https://www.usda.gov/media/blog/2017/06/20/being-serious-about-saving-bees
2. https://www.fda.gov/animal-veterinary/animal-health-literacy/helping-agricultures-helpful-honey-bees
3. https://ocm.auburn.edu/newsroom/news_articles/2021/06/241121-honey-bee-annual-loss-survey-results.php
4. https://theconversation.com/a-game-of-drones-why-some-bees-kill-their-queens-83624
5. https://hal.archives-ouvertes.fr/hal-00892055/file/hal-00892055.pdf

Demo Time!

- The Streamlit app is run entirely through Google Colab