BugBot: Understanding the Dataset

Shirley Fong, Keegan Veazey, Le Ju

January 21, 2025

The main objective of our project is to deliver a user-friendly model that classifies common New England insect images uploaded by New England residents. This objective is met through our thorough dataset collection and filtering process, which are discussed below. The data collection process is key to achieving our goal and provides a strong and consistent foundation for our project because of its diverse and evenly represented classes.

The majority of the BugBot dataset was scraped using a publicly available API for scraping images from Bing search queries. A manual review and filtering process was then applied to the Bing-based dataset and was subject to the following standards:

1. The insect must be present in the image
2. The image must show at least 75% of the insect's body
3. The photo is not a drawing, cartoon, or an AI generated image
4. The image depicts the adult/matured insect

For all collected images, it is also essential to eliminate duplicates or near-duplicates to prevent data leakage. Visual inspection techniques will be used to verify this in addition to image hashing as needed. Since no temporal or spatial dependencies exist in the dataset, random splitting for training, validation, and test subsets is appropriate. The validation set will guide hyper-parameter tuning, while the test set will provide an unbiased assessment of the model's performance.

Furthermore, it is important to note the unique nature of many pests that usually infest home environments in groups or colonies, including ants, termites, and bed bugs. Therefore, a direct effort was made to collect additional group images of insect infestations to be included in the dataset. After filtering, a range of approximately ˜ 16%-76%, depending on the insect class, were kept from the original scraped data based on the above criteria.

After web scraping, additional data was collected by performing manual Google searches for image results using insect keywords, and by taking advantage of community postings on websites such as Reddit to collect home-environment-specific images. After combining the scraped images and manual supplement, the resulting dataset includes 160 raw images across each of the 11 insect classes.

Since our dataset consists of RGB images with x and y coordinates, the only features are the image pixels themselves. However, additional hidden features will be extracted through preprocessing and modeling the data and includes visual characteristics, such as texture, RBG value, and pixel sequence. Furthermore, when the complete image matrix

is flattened into a 1-dimensional feature vector, each pixel element becomes a feature of the overall image.

The target variable in this supervised learning project is the insect class, representing one of the 11 common household pests. It is a well-defined categorical variable with clear labels derived from the dataset. Since the dataset is designed with a balanced distribution of 160 images per class (100 for training, 40 for validation, and 20 for testing), there is no inherent class imbalance, and resampling techniques are not necessary. This ensures that the model will not favor any single class during training or evaluation.

Due to the dataset design and manual filtering criteria, we have curated a dataset that adequately represents every insect class to ensure equal distribution of training, validation, and test data to reduce bias. We are also aware of potential bias in web scraping data due to the prevalence of scientific images which may not reflect the targeted end user – an everyday homeowner classifying an insect in their home. To mitigate this we have further supplemented the dataset with manually selected images with diverse backgrounds and contexts including images from Reddit and YouTube thumbnails.

In terms of resources required, the dataset, consisting of 1,760 images is relatively small and is manageable to process with standard resources. The dataset size, even after augmentation, is unlikely to exceed the limits of local storage or memory. Therefore, the dataset does not require advanced techniques such as distributed processing (e.g., using Apache Spark or Dask) since the computational and storage demands are minimal. Although no distributed processing will be required, the data will require multiple preprocessing steps which include:

1. Standardizing the data to a fixed image size (e.g., 224x224)
2. Normalizing the pixel values
3. Applying data augmentation techniques including rotation, height and width shift, brightness adjustments, and zoom, to expand the training dataset

It is important to acknowledge the limitations of this dataset which could include a lack of sufficient data, potentially leading to over-fitting during training and data imbalance. When this case occurs, it could skew model predictions and affect the model's accuracy. This will be addressed and resolved through data augmentation, mentioned above, through techniques including image rotation, horizontal and vertical flips, and cropping of images. By ensuring that each insect class contains 160 unique images and by expanding the dataset to include varied versions of all images, we effectively diversify the data set to avoid over-fitting and enhance the model performance.

The final outcome of BugBot will provide interpretable insights to end users by providing accurate classifications of pests in their homes. We will utilize Streamlit to deliver a seamless user-friendly interface that takes in a user's uploaded insect image and communicates the model output – the predicted classification – through a text display on the screen. By using Streamlit as a platform for our model deployment, users will be able to receive their results in real time.