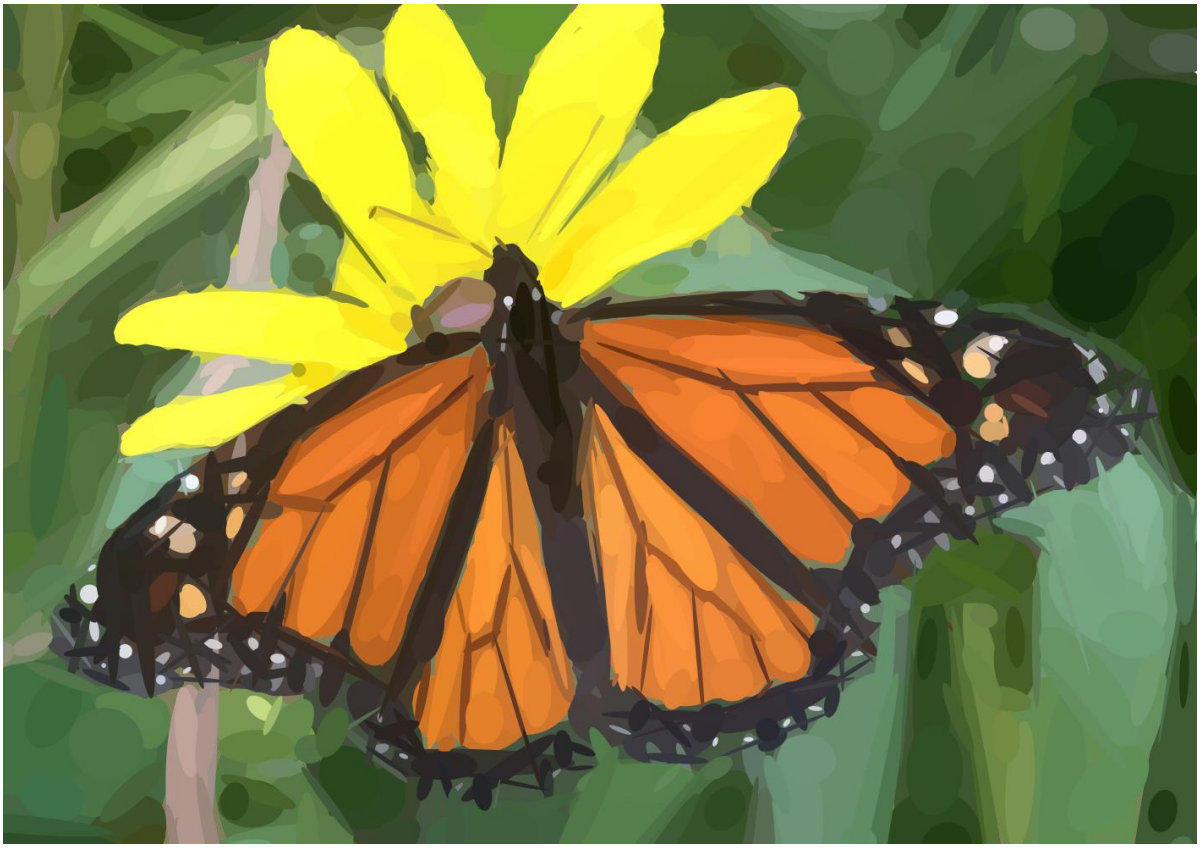


Yum or Yuck – Butterfly Mimics

This is the tiny dataset of two common North American butterflies: the toxic Monarch and its mimic the Viceroy.

In the, eat and be eaten world of the animal kingdom, butterflies use a variety of tactics to stay alive long enough to reproduce. Adult butterflies suffer from predation from birds, ants, wasps, other insects, and mammals including bats, and yes, even rats. To combat this, butterflies in all their stages have developed an array of defensive measures, including use of toxins, repellents, camouflage, and mimicry.

With this dataset we are interested in the defensive use of toxins and mimicry. The question is, can an artificial neural network outperform a bird and distinguish between the toxic butterfly and the mimic?

The Dataset

The dataset consists of 224x224 jpg images each showing a single butterfly in the wild. These images are of 6 species of very common North American butterflies. The images were gathered from the internet and cropped to a square dimension. Images should be used for education or research purposes only.

The dataset is split 80/20 for training and testing – for validation, use k-fold cross validation against the training split or something similar.

The dataset has the following directory structure (a *tiny* dataset is also provided with the same structure).

```
data/
├── butterfly_info.csv
│
└── butterfly_mimics/
    ├── image_holdouts.csv
    ├── images.csv
    ├── image_holdouts/
    │   ├── ggc1e08cbc.jpg
    │   ├── gh20ab0d9c.jpg
    │   ├── gi31f90cd5.jpg
    │   ⋮
    └── images/
        ├── gh150f104b.jpg
        ├── gh2d5c8c79.jpg
        ├── gh6adf74a4.jpg
        ⋮
```

CSV Image Files

The `image_holdouts.csv` and `images.csv` files have the same structure of 4 columns:

- Image
- Name
- Stage
- Side

For example, the `images.csv` begins with:

```
image,name,stage,side
gh150f104b,tiger,adult,both
gh2d5c8c79,monarch,adult,dorsal
gh6adf74a4,pipevine,adult,dorsal
⋮
```

Image Label

The image label, for example `gh150f104b` is the file name less the `.jpg` extension. The label is random and contains no coded information.

Name Label

The `"data"` dataset consists of six common North American butterfly species that we have given a shortened name of: `black`, `monarch`, `pipevine`, `spicebush`, `tiger`, and `viceroy`. The `monarch` and `pipevine` are toxic butterflies and the `black`, `spicebush`, `tiger`, and `viceroy` are the mimics. (This is a somewhat oversimplified description, see the butterfly descriptions that follow for a more accurate picture.)

Name	Common Name	Species	Yum ?	Real ?	Note
black	Black Swallowtail	Papilio polyxenes	yum	mimic	mimics pipevine
monarch	Monarch	Danaus plexippus	yuck	real	cardiac glycoside toxins
pipevine	Pipevine Swallowtail	Battus philenor	yuck	real	sequester aristolochic acid
spicebush	Spicebush Swallowtail	Papilio troilus	yum	mimic	mimics pipevine
tiger	Eastern Tiger Swallowtail	Papilio glaucus	yum	mimic	female mimics pipevine
viceroy	Viceroy	Limenitis archippus	yuck	mimic	sequester salicylic acid

The `"tiny"` dataset contains just the `monarch`, and `viceroy` butterflies.

Stage Label

Of the 4 stages of the butterfly lifecycle we only include images of the winged adults. In other words, all stage labels will be, `adult`.

If you have ever touched a butterfly and had colored dust come off on your fingers, that is not dust but the butterfly's scales. Butterfly wings are covered in scales making it possible for them to have very different patterns on either side of the wing. Butterfly wings have an independent forewing and hindwing allowing some of the dataset images to have a very different wing outline than others. For the purpose of identification we consider the fore- and hindwing to be parts of a single wing. So instead of 4 wings, we would say that a butterfly has two wings: a left and right. Additionally, each wing has two sides a dorsal and ventral.

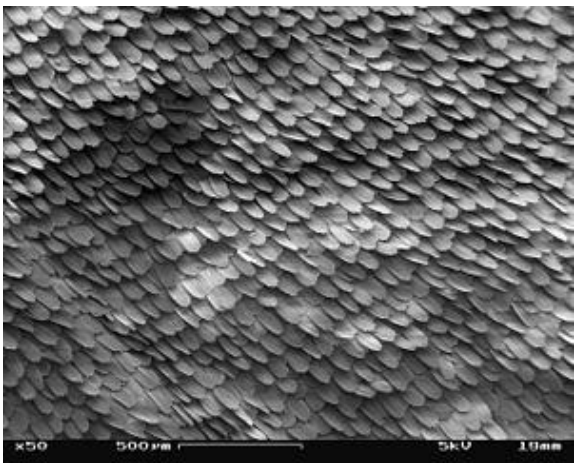


Figure 1 Electron microscope image of scales of wing, see https://en.wikipedia.org/wiki/Lepidoptera

Side Label

We have labeled the images as showing wing sides of either:

- `dorsal`
- `ventral`
- or, `both`

If you think of a butterfly being like a book where the body of the butterfly is the book's spine, then the `ventral` side of the wing is the like the book cover and the `dorsal` is like the inside of the book. So to avoid confusion it may help to think of the dorsal not being the back of the wing but instead the inside of the wings.

In this dataset most images labeled `ventral` are of only one wing. Most images labeled `dorsal` are of the left and right wings. Most images labeled `both` are of a ventral wing in the foreground and a dorsal wing obscured behind it.

We use the following rules for labeling:

Label side as `both` if:

- portions of ventral and dorsal wing sides are clearly visible
- or, If some pattern of the background wing is visible
- or, if more than a thin edge of the background wing is visible
- of, if it is uncertain if the image is of a dorsal side or ventral side or both

Label side as `ventral` if:

- only ventral side of wings are visible
- or, if swallowtail and tail of background wing is visible so that two tails are visible but no other part of the background wing
- or, if a thin edge of the background wing is visible but with no visible pattern

Label side as `dorsal` if:

- only dorsal side of wings are visible
- or, if only a paper thin edge of the foreground wing is visible and one dorsal wing in back

Butterfly Descriptions

Monarch *(Danaus plexippus)*

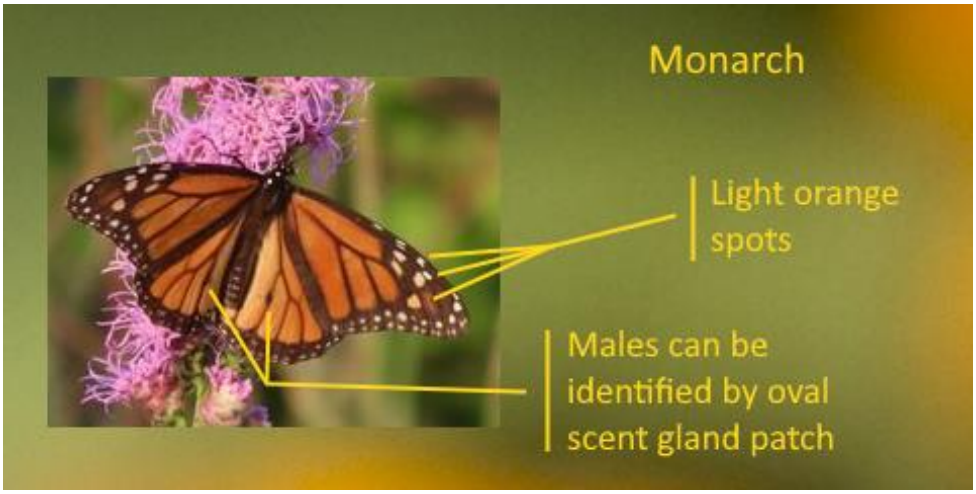
Monarchs are strong fliers and can fly 1000km without stopping. Some monarchs cross 600km of open ocean migrating to their overwintering sites in Mexico. Some Monarch populations migrate from Canada to Mexico and back.

Monarch caterpillars consume milkweed, sequestering cardiac glycoside toxins. Mowing and human intrusion is negatively impacting milkweed availability and may be affecting Monarch populations.

Caterpillar Host Plants

Milkweeds and other plants of the Dogbane family.

Reliable Identifiers



Viceroy *(Limenitis archippus)*

Viceroys are not related to Monarch butterflies. They effectively mimic Monarchs and it is not unusual to find publications about Monarchs incorrectly showing a picture of a Viceroy. Viceroys are from the genus of Admiral butterflies and Monarchs are from the genus of Milkweed butterflies. Viceroys use many defenses during their lifecycle to avoid being eaten: caterpillars mimic bird droppings and eat willows which make them distasteful, adult butterflies mimic Monarchs.

Viceroys are categorized as a "mimic" and at one time it was not known they sequestered salicylic acid and tasted bad.

Caterpillar Host Plants

Willows, Cottonwood

Reliable Identifiers