Questify Data Collection

Tyler Venner

December 2025

1 Dataset Class Mapping and Taxonomy

The dataset consists of 12 classes targeting medically significant arthropods in North Carolina. Images are sourced from iNaturalist with a "Research Grade" quality filter. The taxonomy strategy varies by class, utilizing Family-level, Genus-level, or Order-level groupings.

Table 1: Target Classes and Taxonomic Mapping for North Carolina Arthropod Detection

ID	Class Name	Target Taxon	Rank	Notes & Exclusions
0	Venomous Spiders	*Theridiidae*, *Sicariidae*	Family	Includes *Latrodectus* (Widows) and *Loxosceles* (Recluses).
1	Ticks	*Ixodidae*	Family	Hard ticks only. Excludes *Argasidae* (soft ticks).
2	Mosquitoes	*Culicidae*	Family	Includes all major vectors (*Aedes, Anopheles, Culex*).
3	Stinging Wasps	*Vespidae*	Family	Includes Yellowjackets, Hornets, and Paper Wasps.
4	Bees	*Apidae*	Family	Includes Honeybees, Bumblebees, and Carpenter bees.
5	Fire Ants	*Solenopsis*	Genus	Strict genus-level filtering to distinguish from harmless Formicidae.
6	Assassin Bugs	*Reduviidae*	Family	Includes Kissing Bugs (*Triatoma*).
7	Venomous Caterpillars	*Megalopygidae*, *Limacodidae*	Family	Focus on larval stages; excludes adult moths.
8	Blister Beetles	*Meloidae*	Family	Includes *Epicauta* spp.
9	Scorpions	*Vaejovidae*	Family	Primary NC target: *Vaejovis carolinianus.*
10	Horse & Deer Flies	*Tabanidae*	Family	Includes *Tabanus* and *Chrysops.*
11	Centipedes	*Chilopoda*	Order	Broadest category; includes *Scolopocryptops.*

2 Biological Context and Taxonomic Challenges

To effectively train the classification model, it is necessary to understand the biological hierarchy (taxonomy) from which the classes are derived. The dataset operates at varying levels of taxonomic specificity, primarily Family, but occasionally Genus or Order.

2.1 Taxonomic Levels of Specificity

Taxonomy classifies organisms into a hierarchy: *Kingdom* → *Phylum* → *Class* → *Order* → *Family* → *Genus* → *Species.* This project utilizes three distinct levels:

- **Family Level:** Most classes (e.g., *Culicidae* for Mosquitoes, *Vespidae* for Wasps) are defined at the Family level. Organisms within a Family typically share strong visual characteristics (morphology) that a Convolutional Neural Network (CNN) can easily generalize.
- **Genus Level (High Precision):** Class 5 (Fire Ants) targets the specific genus *Solenopsis*. This requires the model to distinguish minute features (such as head shape and coloration) to differentiate dangerous fire ants from harmless ants in the broader *Formicidae* family.
- **Order Level (Broad Categorization):** Class 11 (Centipedes) targets the Order *Chilopoda*. This is a high-level grouping containing diverse species, but they share a unique body plan (elongated, multi-segmented) distinct from all other classes.

2.2 Morphological and Biological Challenges

2.2.1 The "Stinging Trio" (Order Hymenoptera)

Classes 3 (Wasps), 4 (Bees), and 5 (Fire Ants) belong to the same Order, *Hymenoptera*, sharing genetic and visual similarities. The model must learn subtle distinctions:

- **Wasps (*Vespidae*):** Characterized by smooth, shiny bodies and a distinct "pinched" waist.
- **Bees (*Apidae*):** Distinguished by a hairy/fuzzy texture (for pollen collection) and stockier body shape.
- **Fire Ants (*Solenopsis*):** Must be distinguished from other ants by specific coloration (copper-brown head, darker abdomen) and lack of wings (in most castes).

2.2.2 Metamorphosis and Life Stages (Class 7)

Class 7 (Venomous Caterpillars) presents a unique challenge due to metamorphosis. The families *Megalopygidae* and *Limacodidae* include both the venomous larval stage (caterpillar) and the harmless adult stage (moth).

> **Data Requirement:** Training data must be strictly filtered for **"Life Stage: Larva"** to prevent the model from learning to flag harmless moths as dangerous.

2.2.3 Intra-Class Variance (Class 0)

Class 0 (Venomous Spiders) aggregates two visually distinct families:

1. **Theridiidae (We only use Latrodectus):** Bulbous, round abdomens with glossy textures.
2. **Sicariidae (Recluses):** Flatter bodies, longer legs, and matte textures.

The model must learn a multi-modal representation for this single class, as the two subgroups share few visual features despite both being "venomous spiders."

3 Data Collection and Balancing Strategy

An initial audit of the iNaturalist database revealed significant disparities in species prevalence within North Carolina (NC). We use the *pyinaturalist* api wrapper. While classes such as *Apidae* (Bees) and *Vespidae* (Wasps) yielded over 15,000 local research-grade observations, medically critical classes like *Vaejovidae* (Scorpions) were virtually absent in the local dataset ($N = 26$).

To address this volume disparity while maintaining ecological relevance, we adopted a **Hybrid Geographic Strategy**:

- **Tier 1: NC-Sourced Classes.** For classes with $N > 1,000$ local observations, we restricted data collection to North Carolina (Place ID: 30). This preserves the local visual context (background flora) and ensures the model learns to distinguish targets from local look-alikes.

- **Tier 2: USA-Sourced Classes.** For classes with insufficient local data ($N < 800$), we expanded the geographic filter to the United States (Place ID: 1). This ensures sufficient training volume for rare medical threats.

Table 2: Final Data Source Configuration by Class

Class	**NC Count**	**Source Geography**	**Rationale**
Scorpions	26	**USA**	Insufficient local data.
Blister Beetles	755	**USA**	Below 1k threshold.
Horse/Deer Flies	652	**USA**	Below 1k threshold.
Venomous Spiders	1,187	**NC**	Sufficient local volume.
Ticks	1,291	**NC**	Sufficient local volume.
Mosquitoes	1,483	**NC**	Sufficient local volume.
Fire Ants	1,535	**NC**	Sufficient local volume.
Venomous Cats.	1,977	**NC**	Sufficient local volume.
Centipedes	2,135	**NC**	Sufficient local volume.
Assassin Bugs	8,209	**NC**	Abundant local data.
Stinging Wasps	16,266	**NC**	Abundant (Requires Downsampling).
Bees	38,242	**NC**	Abundant (Requires Downsampling).

3.1 Class Balancing

To prevent the model from biasing towards majority classes (e.g., Bees), we implemented a **strict cap of 2,000 images per class**. This is subject to change later. For abundant classes, images were randomly sampled. For scarcity classes expanded to the USA, we collected up to the cap to ensure a balanced training distribution.