Master's Thesis Proposal

Learning Wing Similarity Representations for Taxonomic Clustering

Muhammad Faraz Mazhar
Friedrich-Alexander-Universität Erlangen-Nürnberg

January 2026

---

1 Thesis Title

Main title: "Learning Wing Similarity Representations for Taxonomic Clustering: Validating Deep Learning-Based Insect Wing Classification Against Biological Taxonomy"

2 Core Objective

a) **Learn Wing Similarity:** Train a deep learning model to learn meaningful similarity representations of insect wings across multiple taxonomic groups.

b) **Cluster Wings:** Group similar wings together using the learned similarity representations.

c) **Biological Validation:** Systematically analyse whether the model's clusters align with actual biological relationships (taxonomic families, genera, and orders).

3 Research Question

Can deep learning learn meaningful wing similarity representations that reflect biological taxonomy? Specifically, do wings that the model clusters together actually belong to the same taxonomic family, genus, or order, i.e., do the learned representations capture properties that correlate with evolutionary relatedness?

4 Current Challenge

- Insect wings are used for taxonomic identification, but manual classification is time-consuming.
- Deep learning can learn similarity representations, but it is unclear whether these align with biological taxonomy.
- Need to validate whether AI-learned features capture biologically meaningful relationships.

5 Research Gap

- To my knowledge, most deep learning work on insect wings focuses on classification or detection/segmentation tasks, rather than similarity learning and clustering-based validation against taxonomy.

- Preliminary literature suggests limited work on validating whether learned representations (embeddings, clusters) align with established biological taxonomy using metrics such as ARI or NMI.
- I am not aware of a systematic study examining whether AI-learned wing similarity matches taxonomic similarity across multiple insect orders.
- Quantitative validation frameworks for assessing the biological meaningfulness of learned wing features appear to be lacking.

6 Objectives

6.1 Learn Wing Similarity Representations

- Train a deep learning model to learn meaningful embeddings of wing images.
- Use contrastive learning and/or metric learning approaches (e.g. one of: SimCLR, SupCon, triplet loss, etc.).
- **Concurrently:** Compute wing similarity based on venation structure and landmarks.

6.2 Cluster Wings by Similarity

- Use learned embeddings to cluster wings into groups.
- Apply clustering algorithms (examples: k-means, hierarchical clustering, DBSCAN, spectral clustering, using at least one is sufficient).
- Visualise clusters (e.g. t-SNE or UMAP, one is sufficient).

6.3 Validate Against Biological Taxonomy

- Compare model clusters with known taxonomic relationships at multiple hierarchical levels.
- Measure: Do wings from the same species/genus/family/order cluster together?

7 Methodology

7.1 Phase 1: Data Preparation

7.1.1 Dataset Characteristics

- **Multi-Order Insect Wings:** Dataset comprising wings from multiple insect orders.
- **Temporal Diversity:** Both recent and fossil specimens to assess cross-domain generalisation.
- **Taxonomic Coverage:** Multiple taxonomic levels represented (species, genus, family, order).
- **Taxonomic Labels:** Hierarchical taxonomic information extracted from specimen metadata and verified against established taxonomic databases.

7.1.2 Data Preprocessing

- Extract wing regions using automated segmentation or manual annotation.
- Normalise images (size, orientation, lighting conditions).

- Apply data augmentation (rotation, scaling, brightness, contrast adjustments).
- Create taxonomic label hierarchy (Order → Family → Genus → Species).
- Handle class imbalance and ensure adequate representation across taxonomic groups.

7.2 Phase 2: Similarity Learning

7.2.1 Approach 1: Contrastive Learning & Metric Learning

- **Method:** Contrastive learning and metric learning are closely related. I will use one or more methods from a unified framework: examples include SimCLR, MoCo, SupCon (contrastive), Triplet Loss, Contrastive Loss, ArcFace (metric learning). Both yield embeddings where taxonomic similarity correlates with embedding similarity.
- **Architecture:** One encoder (e.g. ResNet or EfficientNet, pre-trained on ImageNet, fine-tuned on insect wings).
- **Training strategy:**
  - Positive pairs from same taxonomic group (species/genus/family).
  - Negative pairs: wings from different taxonomic groups.
  - Learn embeddings where taxonomic similarity correlates with embedding similarity.
  - Multi-level supervision: incorporate taxonomic hierarchy (order, family, genus, species).
- **Output:** High-dimensional embedding vectors (128 to 512 dimensions) capturing wing morphology.

7.2.2 Approach 2: Similarity Metric Based on Venation Structure

- **Rationale:** This is more specific than feature space alone, and in principle this approach may be more promising, as venation structure and landmarks are biologically interpretable and directly relevant to taxonomy.
- **Method:** Compute wing similarity based on venation structure and landmarks. Examples: segment veins via Fast Marching Method or deep learning, extract landmark positions, define a similarity/distance metric on landmark configurations or venation graphs. One or more of these steps suffice.
- **Implementation options** (examples, using one is sufficient):
  - **Landmark-based:** Use positions of venation landmarks (e.g. from pose/keypoint models or manual annotation) and a geometric similarity metric (e.g. Procrustes or landmark distances).
  - **Venation-graph-based:** Extract venation graphs (nodes = junctions, edges = vein segments), then use graph matching or graph embeddings (e.g. GCN/GAT) to obtain a similarity metric.
- **Output:** A similarity (or distance) measure between wings derived from venation structure. This can be used as an alternative or complementary "embedding" for clustering and validation.
- **Combination:** If the feature-space approach (1) and the venation-structure approach (2) are both helpful, they can be combined later (e.g. concatenated features, ensemble clustering, or late fusion).

7.3 Phase 3: Clustering

Clustering and validation (Phase 4) can be applied to either (or both): feature-space embeddings from contrastive/metric learning, and similarity/distance from the venation-structure-based approach.

7.3.1 Clustering Methods

Examples (using at least one is sufficient):

- **K-Means:** Cluster embeddings into k groups.
- **Hierarchical clustering:** Build dendrogram of wing relationships.
- **DBSCAN:** Density-based clustering (handles outliers).
- **Spectral clustering:** Use graph-based clustering on similarity matrix.

7.3.2 Visualisation

Examples: t-SNE or UMAP (one is sufficient) to visualise 2D embedding space, colour-code by taxonomic labels (genus/family), and visualise clusters and their taxonomic composition.

7.4 Phase 4: Biological Validation

7.4.1 Validation Metrics

Examples (using at least one or a subset is sufficient):

- **Adjusted Rand Index (ARI):** Measures agreement between clusters and taxonomic labels, range -1 to 1 (1 = perfect match). Compares model clusters vs. genus labels, model clusters vs. family labels.
- **Normalised Mutual Information (NMI):** Measures shared information between clusters and taxonomy, range 0 to 1 (1 = perfect match).
- **Taxonomic accuracy:** Genus-level accuracy (% of wings in a cluster that belong to the same genus), family-level accuracy, purity (dominant genus/family in each cluster).
- **Hierarchical validation:** Multi-level analysis at each taxonomic level (species, genus, family, order), hierarchical metrics, cross-level consistency (lower-level clusters nested within higher-level), accuracy, precision, recall at each level.

7.4.2 Statistical Analysis

- Analyse failure cases (why some wings do not cluster correctly).

8 Data Contingencies

If taxonomic labels are not provided or are incomplete, the thesis will: (1) use only the subset of images with verified labels, and (2) prioritise the Cicadidae case study using existing labelled data. Manual annotation of thousands of images is out of scope for this thesis.

9 Timeline (6 Months)

9.1 Month 1: Data Preparation & Literature Review

- Extract and verify taxonomic labels, organise dataset by taxonomic hierarchy.

- Create stratified train/validation/test splits.
- Preprocess images (normalise, augment), handle class imbalance.
- Literature review on similarity learning and taxonomic classification.
- Set up evaluation framework (e.g. ARI, NMI, or taxonomic accuracy).
- Implement one baseline (e.g. ImageNet features).

9.2 Month 2: Similarity Learning

- Implement one contrastive/metric learning approach (e.g. SupCon or triplet loss).
- Train encoder on insect wing dataset with taxonomic supervision.
- Evaluate embedding quality (visualise with t-SNE or UMAP, colour-code by taxonomy).

9.3 Month 3: Clustering & Validation

- Implement one clustering method (e.g. k-means or hierarchical).
- Cluster wings using learned embeddings, determine optimal number of clusters.
- Calculate validation metrics, compare with baseline.
- Visualise clusters and analyse initial results.

9.4 Month 4: Refinement & Fossil Evaluation

- Refine model if needed, test on fossil images.
- Analyse failure cases and key patterns.
- Optional: feature analysis or attention visualisation.

9.5 Month 5: Final Evaluation & Results

- Final model training and evaluation.
- Create visualisations (t-SNE plots, cluster dendrograms), prepare results tables.
- Write results section, begin writing other chapters.

9.6 Month 6: Writing & Finalisation

- Complete thesis (all chapters).
- Address supervisor feedback, proofreading and submission.