Exploring Visual Similarities and Genetic Similarities with Machine Learning

Research Project by Niloy Roy

Supervised by Martin Hofmann

The SPIRIT of science

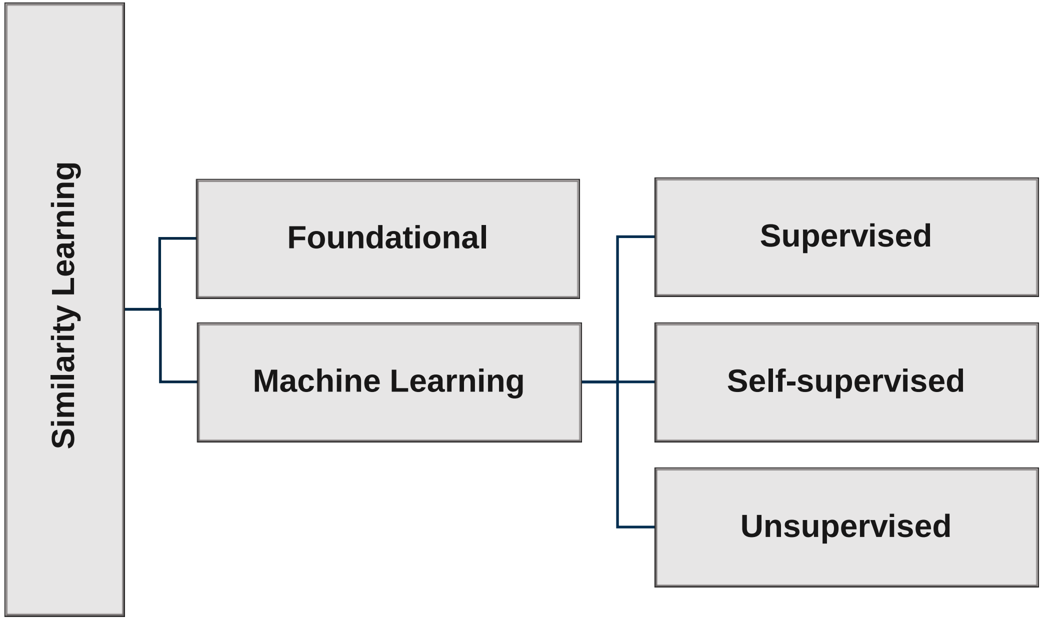
TECHNISCHE UNIVERSITÄT ILMENAU

Contents

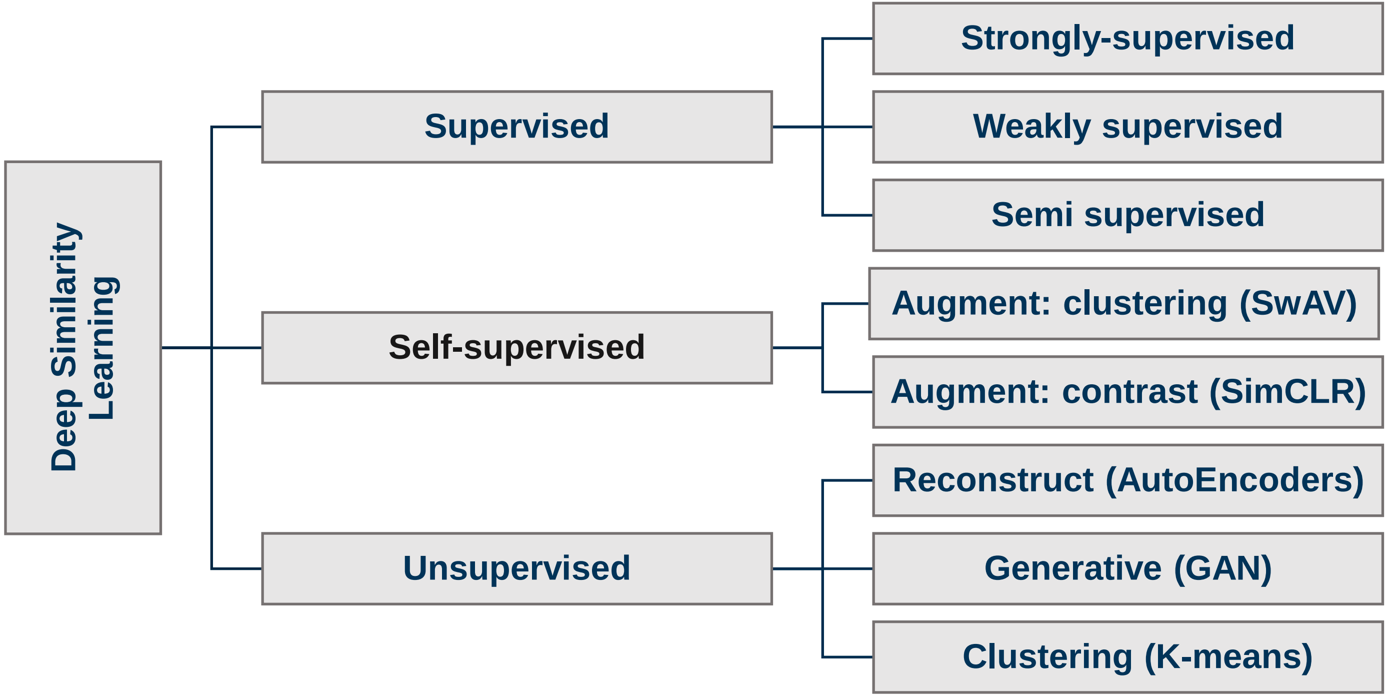
Introduction

- Taxonomic classification: Morphological, Genetic, Ecological similarity and differences
- Challenges in taxonomic study [e.g., complex taxonomic groups (TCGs)]
- Similarity Learning: Foundational Vs Machine-Learning
- Why Machine-Learning? Large-scale image dataset, high dimensionality
- Applications: Automated tools for taxonomy, Meaningful visual latent space
- Range of Impact: Ecologic & Evolutionary Study, Biodiversity Conservation, Medical Study (targeted treatments), Agriculture (new crops)

Similarity Learning Paradigms 1/2



Similarity Learning Paradigms 2/2



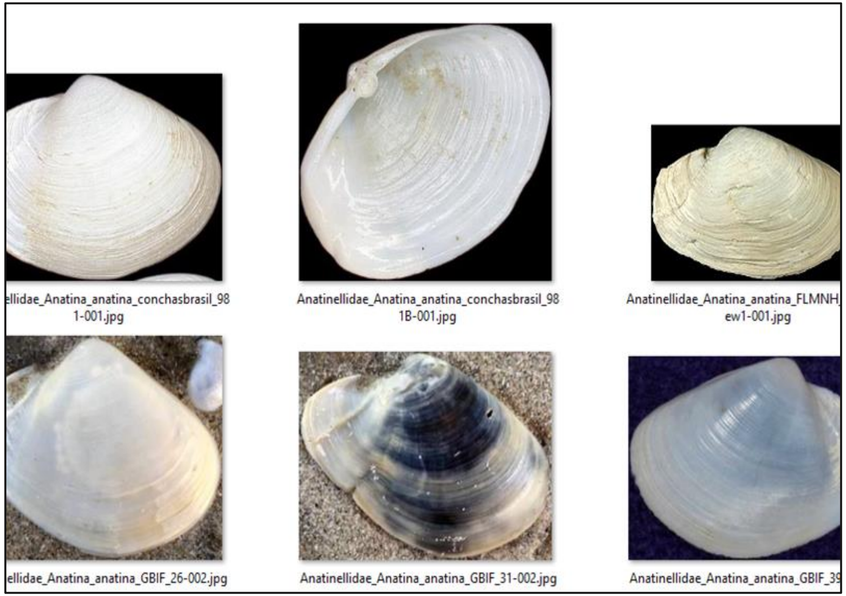
Dataset: Image

- **Phylum:** Mollusca, **Class**: Bivalve
- **Labels taxa**: Subclass (6), Order(26), Family(74), Genera(884), Species(4144)

- 71,888 2D images (.jpg)
- Varying dimension in size (height*width)
- Class imbalance:

  Paleoheterodonta (67 samples)

  Imparidentia (33,852 samples)
- Filename to Taxon label mapping (Meta.csv)
- Source: Hofmann et al. (2024)



Dataset: Phylogenetic Distance

- Family-level pairwise phylogenetic distance
- Distance from phylogenetic tree :
    1. shortest path between taxa
    2. sum of branch lengths,
    3. same order pair < different order pair
- Square and symmetric (74*74)
- Self-distance along diagonal (0)
- Source: Hofmann et al. (2024)

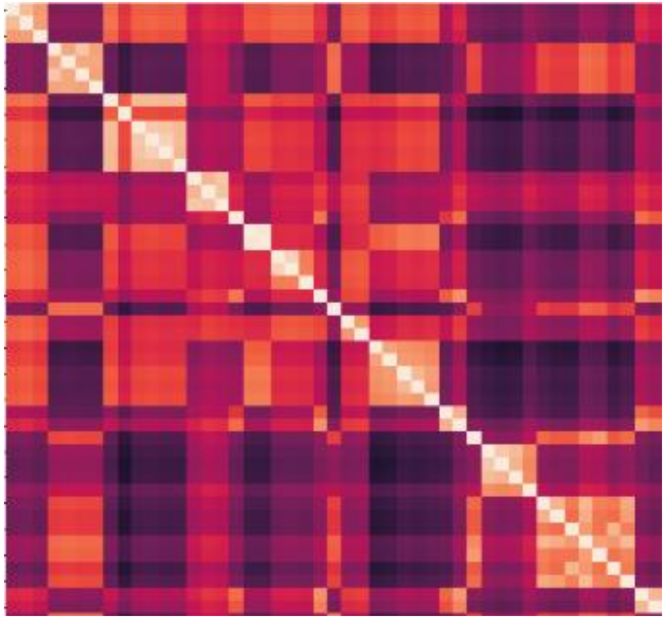


Figure: Heatmap visualization of phylogenetic pair-wise distances of 74 families arranged by order

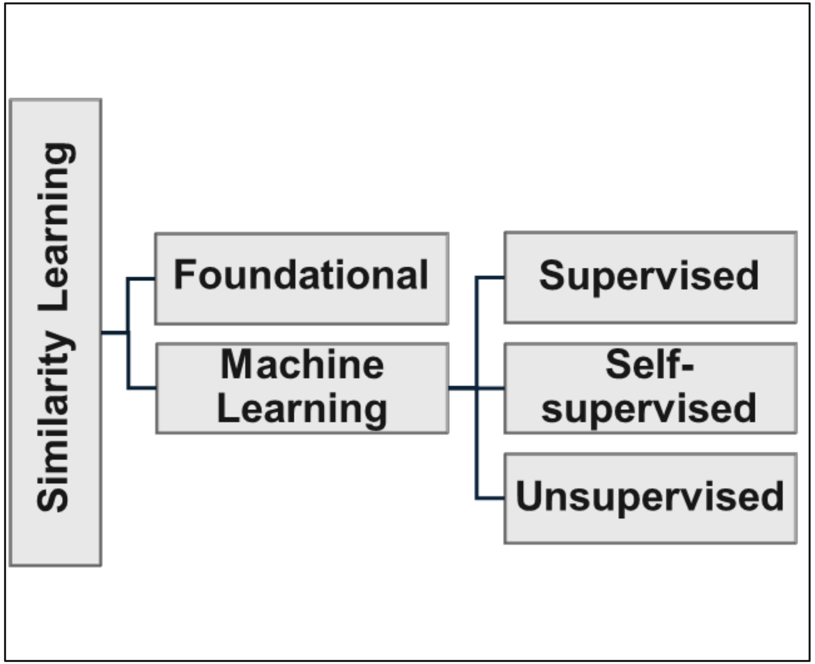
Research Question (1/2)

- **Problem:** Genetic data unavailable for empty shells, fossil records.

- **Question:** With how much confidence can visual similarity serve as a **proxy** for genetic similarity in **taxonomic classification**?

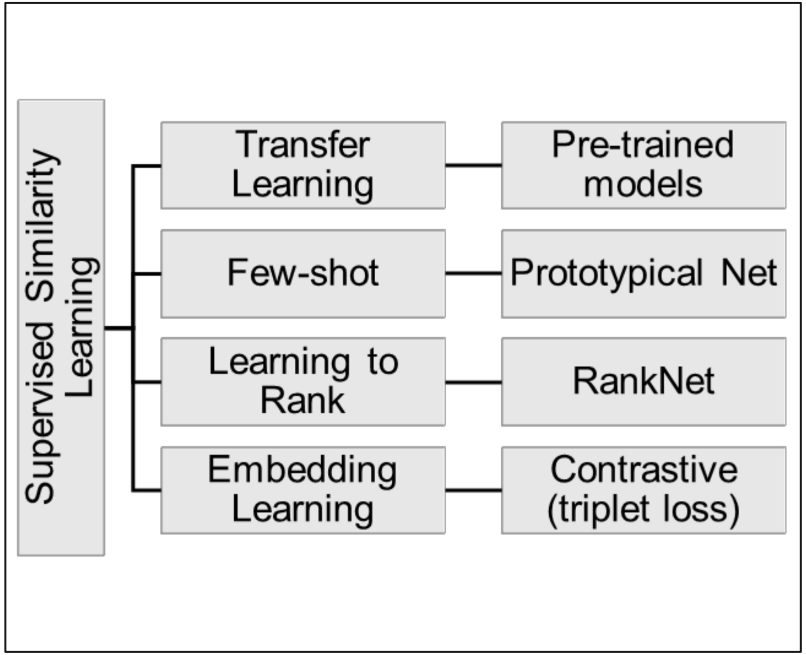
Experiment (1/2): Paradigm selection

- **Foundational or Deep Methods?**
  - ➢ Fine-grained, high-dimensional
- **Decision: Deep Learning**
  - ➢ Taxonomic labels: instance-level, highly-reliable
  - ➢ Semantic supervision preferred
    Convergent evolution (*M. edulis, M. galloprovincialis, and M. trossulus*)
  - ➢ Class imbalance**:** Data-based, model-based, hybrid methods (Das et al., 2022)
- **Decision: Supervised**



Experiment (1/2): Paradigm selection

- **Supervised: Contrastive, Ranking or Downstream Classifier ?**
  - ➢ Objective, Training efficiency
  - ➢ Feature transfer with pre-trained model
- **Decision: Strongly Supervised Classifier with Feature Transfer**
- **Performance**:
  Random guess < Feature Transfer
  < Fine-Tuning (parameter training)



Experiment (1/2): Feature Transfer

Downstream Classifier

- Flexibility: neural network / traditional classifiers (SVM, Random Forest)

CNN Feature Extractor

- Pre-trained model for reduced training effort
- Parameter sharing, local invariance, hierarchical feature extraction

Random Forest

- Robust in noisy high-dimensional settings
- Interpretability: Easier feature ranking
- Weighted Random Forest for imbalance

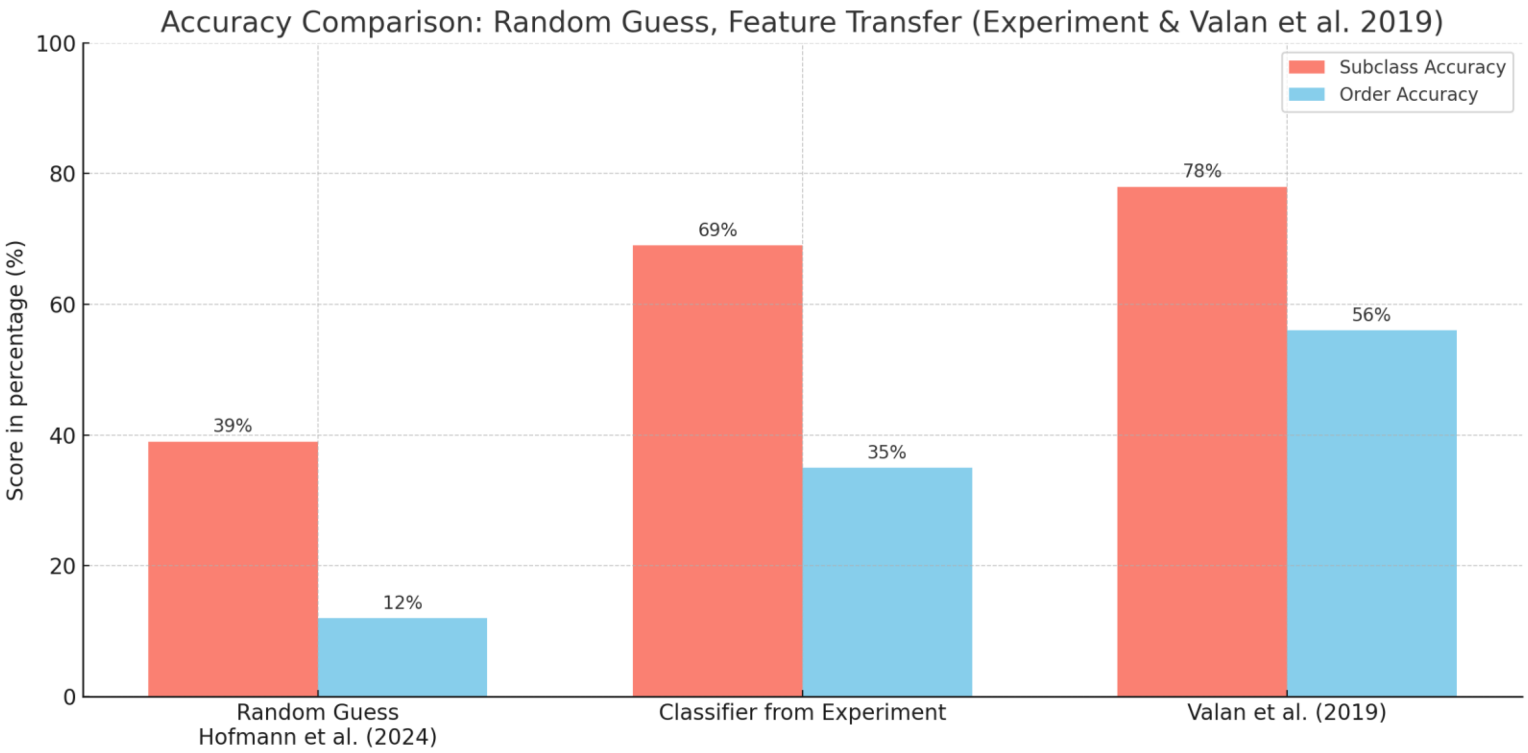
Experiment (1/2): Feature Transfer

Implementation: Feature transfer based taxonomic classifier for subclass and order

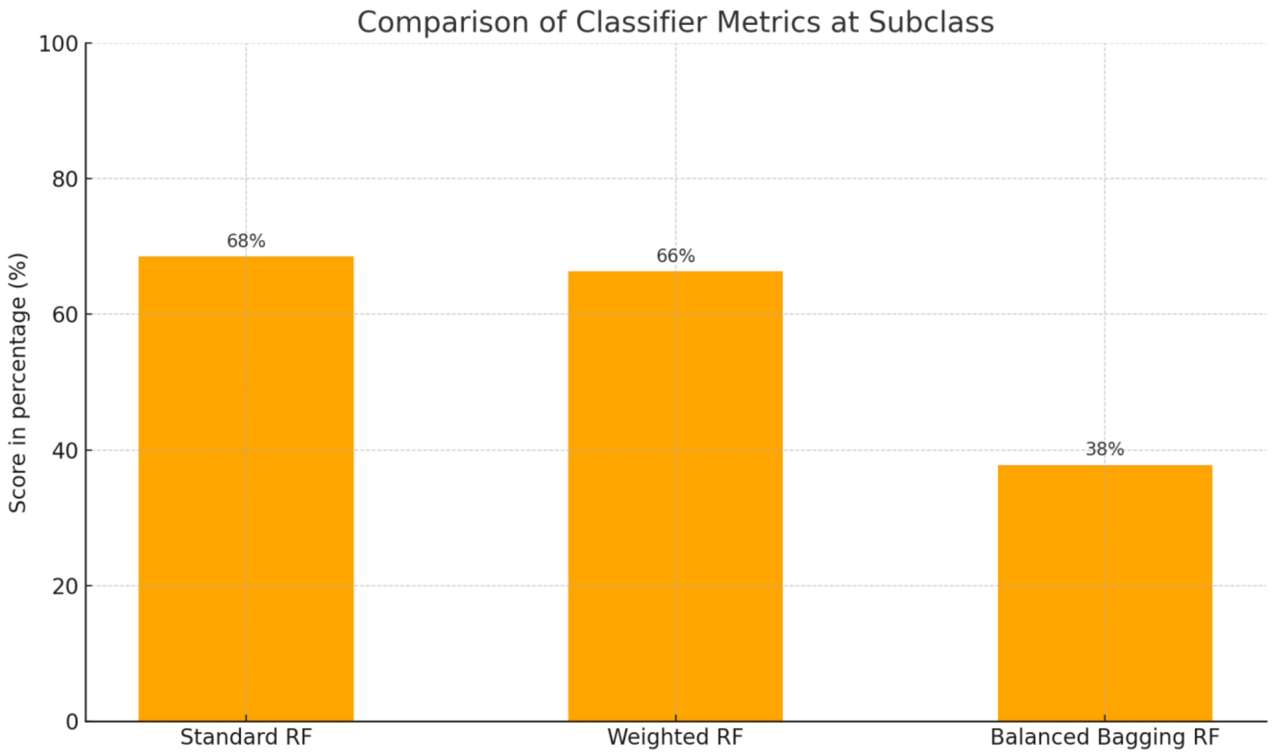
Dataset
↓
Stratified train-test (80-20)
↓
Augmentation of training data
↓
Feature extraction
↓
Random forest
↓
Final taxonomic prediction
↓
Classification report

- Up-sampling minority classes (gaussian noise)
- Augmentation inside feature extractor
- **Pre-trained Model:** ConvNexT (Tiny)
  ⟶ Top performance on ImageNet
- **Feature matrix**: (768, number of samples)
- **Classifiers for taxon**: independent, parallel
- **Random forest**:
  Bootstrap Sampling, Feature Bagging
  100 decision trees

Feature Transfer: Result



Feature Transfer: Result



Research Question (2/2)

- **Problem:** Convergent Evolution, Morphological Disparity
- **Question:** To what extent can a biologically meaningful embedding space be learned from visual morphology data, when **guided** by phylogenetic distance?

Experiment (2/2): Embedding Learning

Implementation: To learn a visual embedding space aligned with phylogenetic distance.

Dataset pre-processing

↓

Feature extraction

↓

Attention-based pooling on latent space

↓

Cosine similarity for pairwise distance of the family feature vectors

↓

HeatMap Comparison

Dataset pre-processing

- Down-sampling majority classes
- Up-sampling minority classes with augmentation
- Normalization of pixel values
- Re-sizing (224*224)

Feature extractor: Resnet50

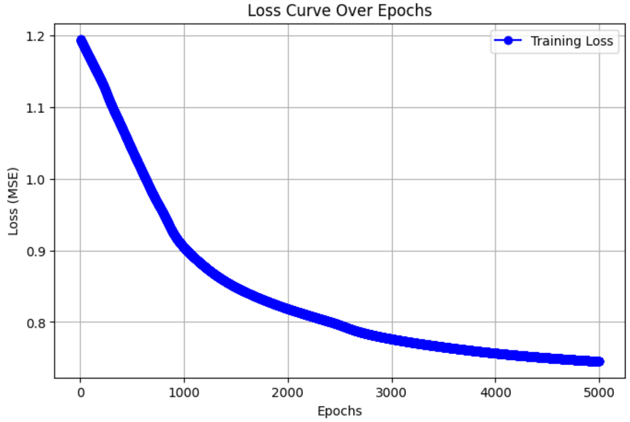
Feature matrix: (number of samples, 2048)

Experiment (2/2): Embedding Learning

Weighted Attention-based Pooling: Importance to discriminative features.

For each family:

n feature vectors → Linear (learns **weight** for each vector) → Softmax → Weighted sum → 1 Representative vector



Pair-wise Visual Distance Matrix:

- 74 family vectors normalized
- Pair-wise cosine distance (74*74)

Training:

- Genetic distance normalization ([0,1.7] to [-1,1])
- Upper-triangle MSE loss

Embedding Learning: Result

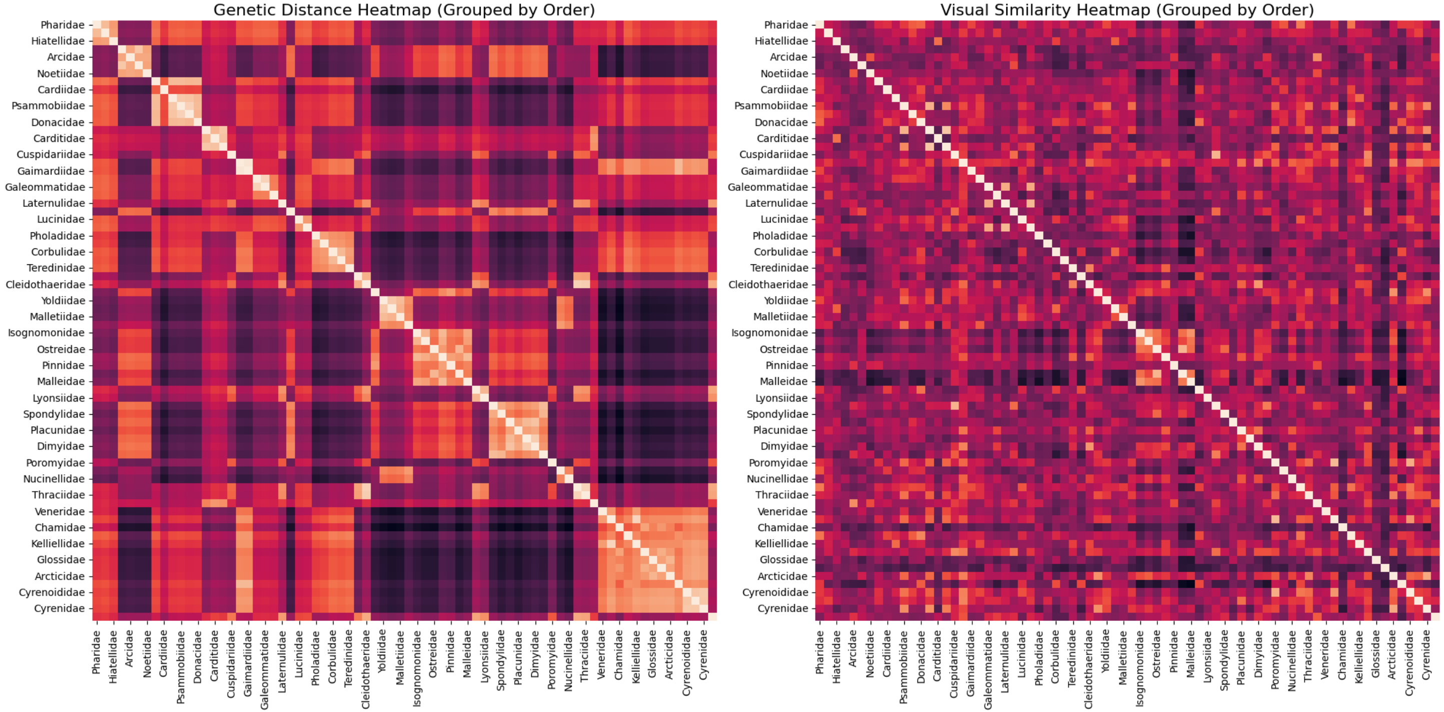


Figure: Comparison of Genetic Distance Heatmap (left) and Visual Similarity Heatmap(right).

Embedding Learning: Result

Correlation	Value	p-value
Experiment	0.4032	5.7419**e**-109
Hofmann et al. (2024)	0.78	< 0.00001

Table: Correlation values comparison from the literature

- **Moderate positive Pearson correlation** between genetic and visual similarity matrices indicates partial linear alignment.
- **Extremely small p-values** confirm correlations are statistically significant and not due to chance.

Conclusion

Taxonomic classification: Visual similarity as proxy (RQ1):

- ➢ Average accuracy significantly better than random guess
- ➢ Aligned with feature transfer performances from the literature
- ➢ Baseline performance: Feature transfer with no fine-tuning,
- ➢ Random Forest with no hyper-parameter tuning.

Distance-guided embedding Learning: Alignment between visual and genetic distances (RQ2):

- ➢ Moderate positive Pearson correlation with very low p-value
- ➢ Encouraging linear alignment between genetic similarity and visual similarity

Technical Tools:

- ➢ Google colab Pro, A100 GPU, Tensorflow (1st experiment), PyTorch (2nd experiment)

Future Work

Fine-tuning:

- Classification: Feature engineering (ranking, fusion); Training neural network
- Embedding: Better representation

Handling Class Imbalance:

- Taxonomic aware re-sampling,
- Neural network training with weighted loss function
- Few-shot learning

Hierarchy-Guided Neural Network (HGNN):

- Taxonomic classification with biological hierarchy (genus and species)
- Dual-ResNet, joint loss function
- Improved accuracy with small, imbalanced data (Elhamod and Tung, 2020)

References

- Hofmann, M., Kiel, S., Kösters, L. M., Wäldchen, J., & Mäder, P. (2024). Inferring taxonomic affinities and genetic distances using morphological features extracted from specimen images: A case study with a bivalve data set. Systematic Biology, XX(XX), 1–22. https://doi.org/10.1093/sysbio/syae042

- Das, S., Mullick, S. S., & Zelinka, I. (2022). On supervised class-imbalanced learning: An updated perspective and some key challenges. IEEE Transactions on Artificial Intelligence, 3(6), 973–991.

- Elhamod, M., & Tung, F. (2020). Hierarchy-guided neural network for species classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3957–3966). IEEE.

- Valan, M., Mokonyi, K., Maki, A., Vondráček, D., & Ronquist, F. (2019). Automated taxonomic identification of insects with expert-level accuracy using effective feature transfer from convolutional networks. Systematic Biology, 68(6), 876–895.

Thank you so much!

The SPIRIT of science TECHNISCHE UNIVERSITÄT ILMENAU

Feature Transfer Classifier: Limitations

Limitations	Implication	Improvement
Class imbalance	Minority class performance drops the overall average accuracy performance	- Hyper-parameter tuning of Random Forest - Support Vector Machine
Class imbalance	- Without down-sampling the majority classes, imbalance skew persists	- Resampling - Repetitions in train_features extractions
High-dimensional feature space	- Weak, noisy and non-discriminative features remaining	- Fusion - Ranking - Principal Component Analysis

Embedding Learning: Limitations

Limitations	Implication	Improvement
Random down-sampling	Loss in feature representation	- Better down-sampling alternatives (Taxonomically aware)
Feature extraction with pre-trained weights only	- No training with our dataset	- **Fine-Tuning**: Unfreeze and train the intermediate to last convolutional layers to balance better feature extraction with reduced over-fitting risk.

Downstream classifier: Results

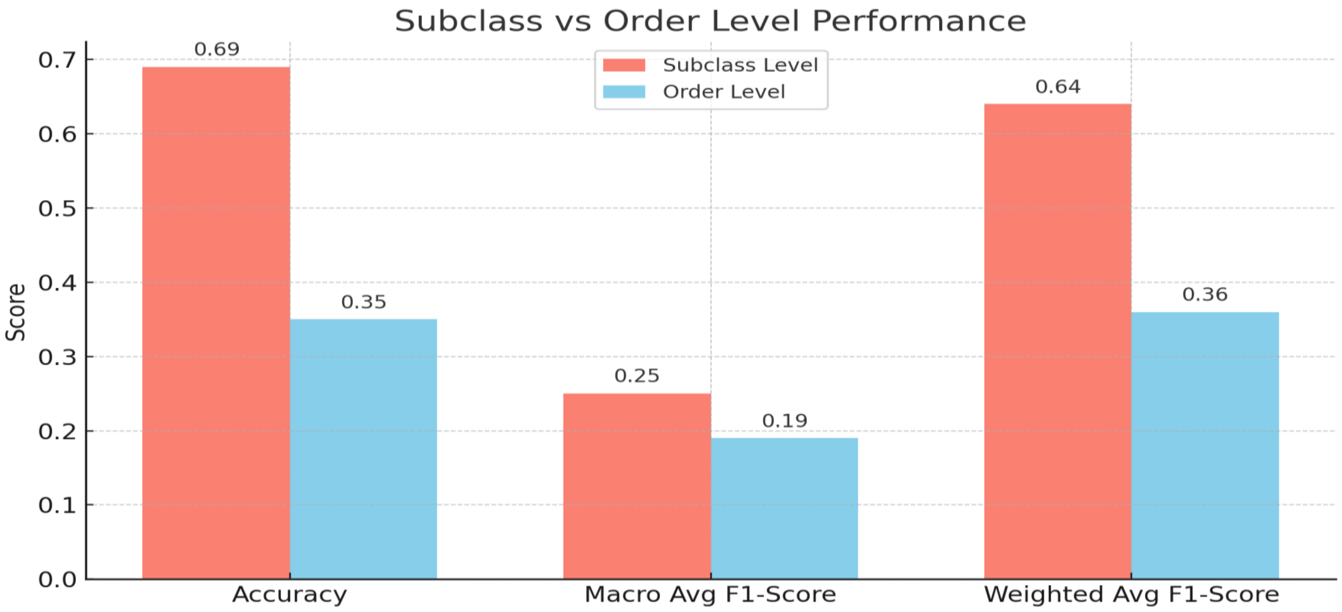


Figure 7: Comparison of Accuracy with macro and weighted average of F1 score

Downstream classifier: Results

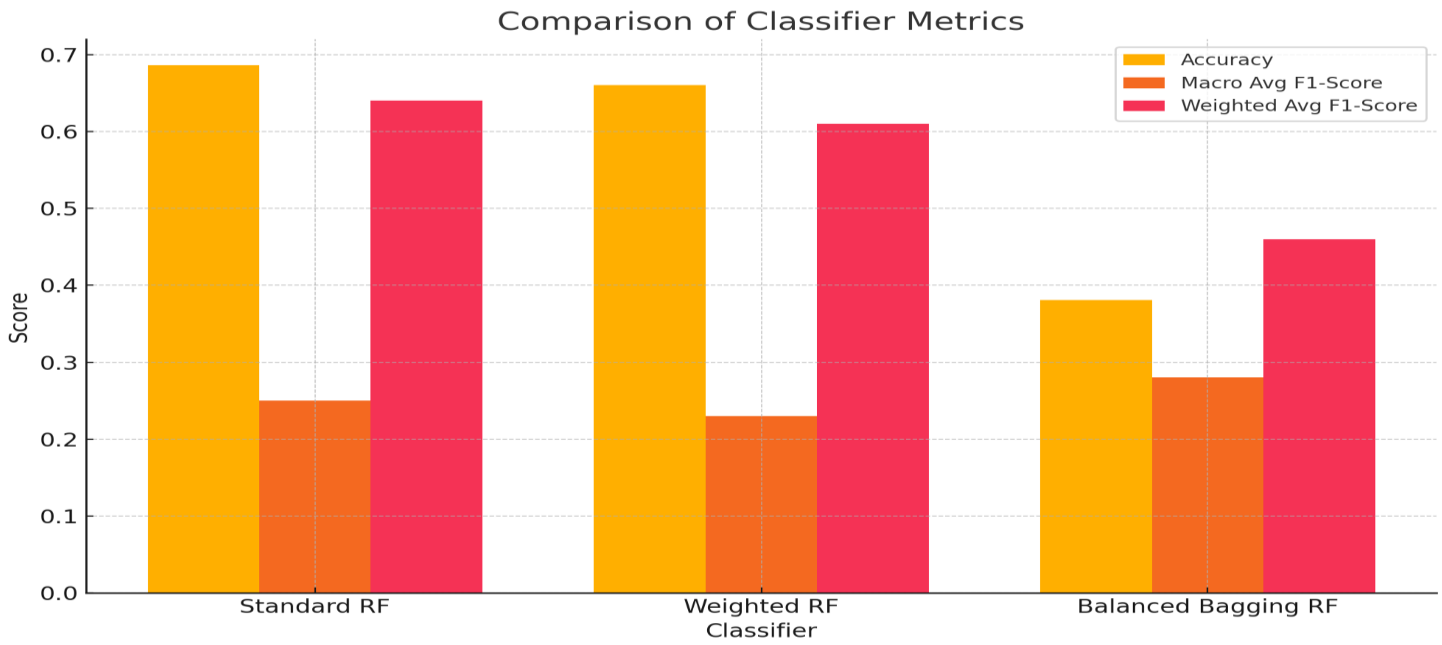


Figure 8: Classification Report or the 6 subclasses (0-5) with different classifiers

Downstram classifier: Results

At taxonomic SubClass:

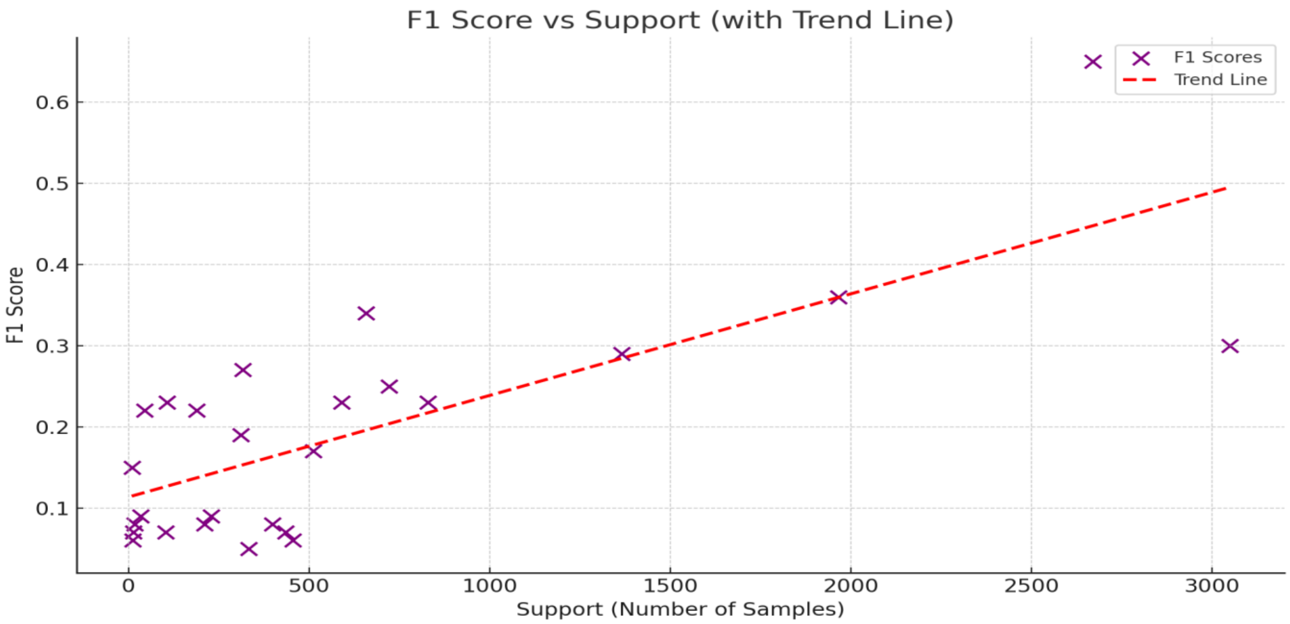


Figure 9: Impact of class imbalance on 26 orders classification performance