From Music to Animal Sounds:
Transfer Learning with the Pretrained YAMNet Model

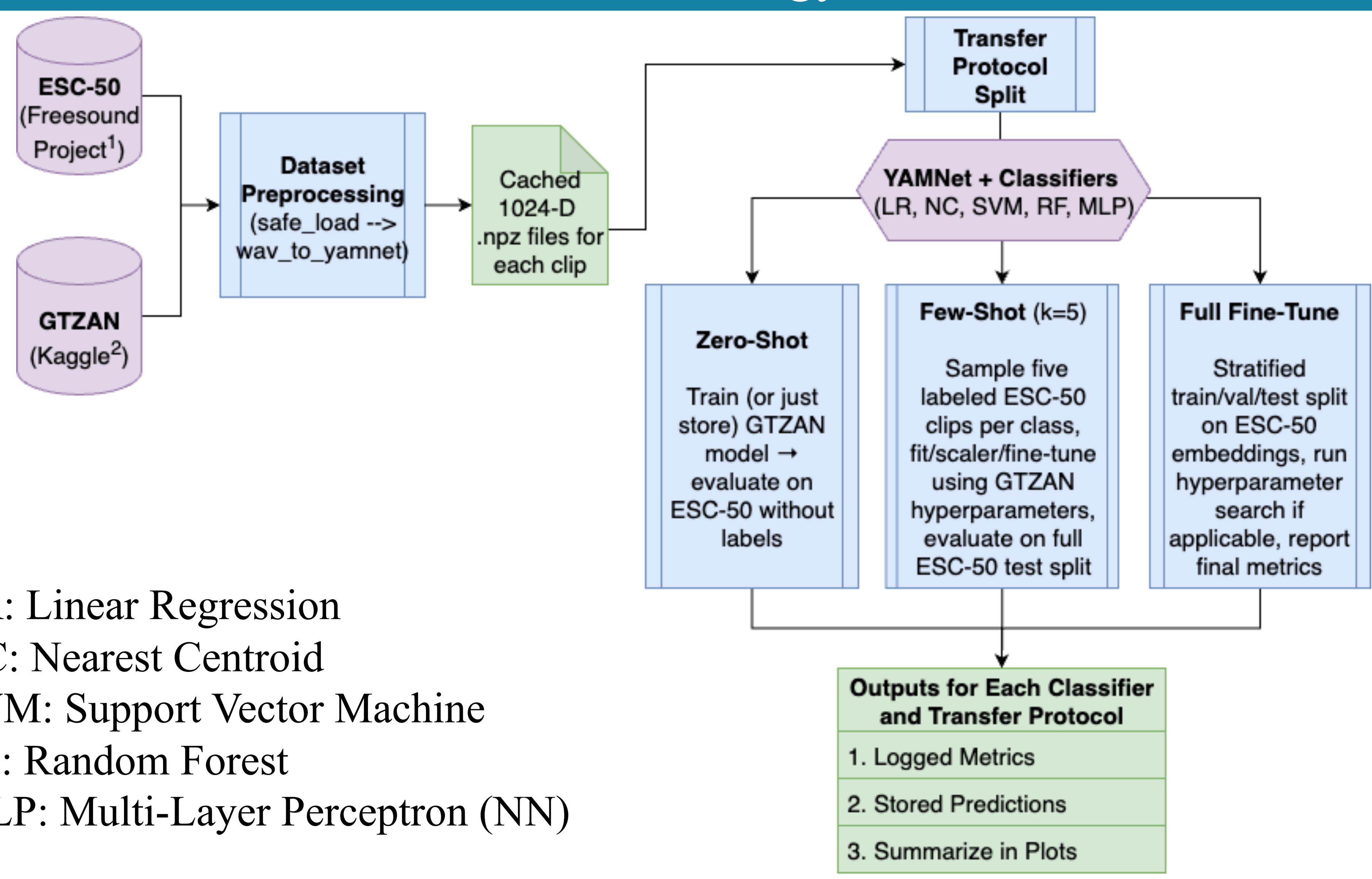
Oyku Cildir and Claudia Martinez
HEC Montréal

Problem & Motivation

Environmental **sound recognition systems still struggle with animal vocalizations**. Building custom deep networks requires more data than most field recordings provide, models either overfit or ignore rare classes.

We freeze **YAMNet**'s AudioSet-trained features, **train lightweight heads on GTZAN**, and **transfer them to ESC-50 animal sounds** across zero-shot, 5-shot, and full fine-tune regimes to see how many labels are needed for reliable wildlife recognition.

Methodology



LR: Linear Regression
NC: Nearest Centroid
SVM: Support Vector Machine
RF: Random Forest
MLP: Multi-Layer Perceptron (NN)

Dataset and Embeddings

GTZAN – Music Genres
1,000 audio clips, 10 genres, 30 sec per clip. Used as the **source domain** for zero-shot & few-shot transfer

ESC-50 – Animal Sounds Subset
2,000 clips, 10 animal classes, 5 sec per clip. Used as the **target domain** for few-shot & full fine-tuning

Each clip is resampled, **embedded once with YAMNet** to a cached 1024-D vector. Every transfer experiment reuses that same descriptor so **only the classifier head and number of ESC-50 labels change.**

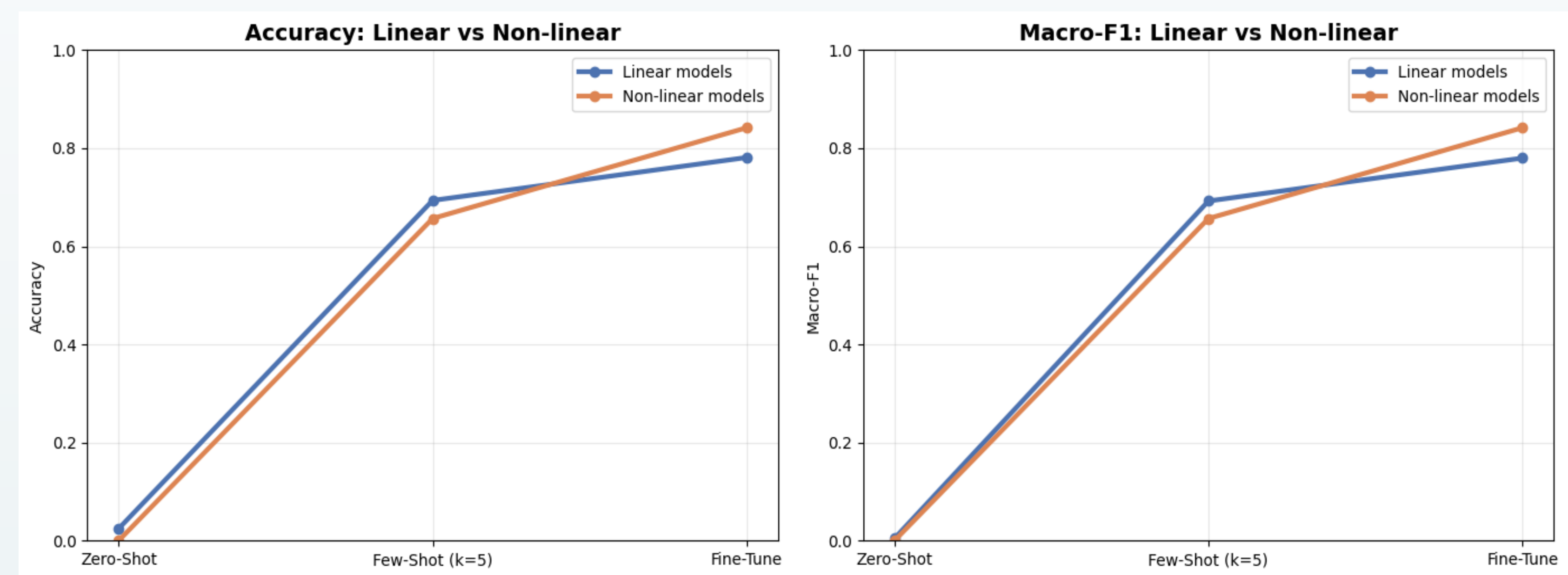
Transfer Protocol

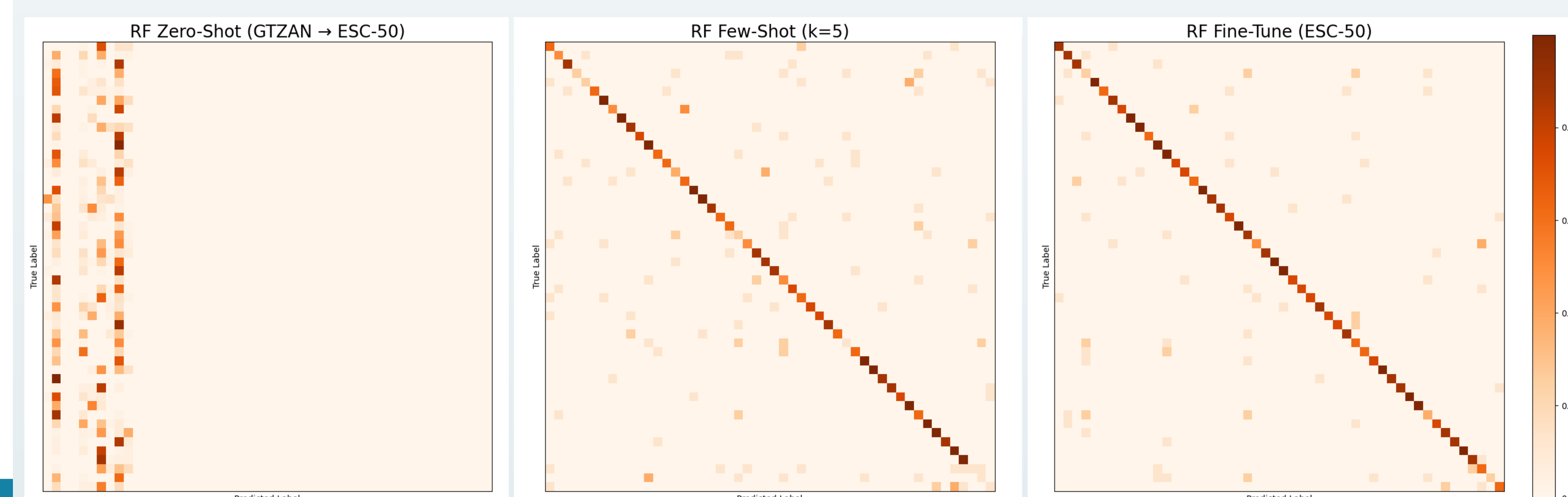
Zero-shot: train each classifier only on GTZAN embeddings, freeze it, and evaluate directly on the ESC-50 animal subset without seeing any target labels to measure domain transfer.

Few-shot (k=5/class): sample 5 labeled ESC-50 clips per class (~250 total), adapt the GTZAN-tuned head using those examples, and test on the full ESC-50 holdout to gauge label efficiency.

Full fine-tune: run a stratified train/val/test split on ESC-50 embeddings, perform limited hyperparameter sweeps, report the best model per head as upper bound when adequate labels are available.

Results





TRAINING EFFICIENCY

Category	Summary	Key Point
Few-shot	65–70% macro-F1; tiny training time (0.003–1.23 s); inference very fast	250 labeled clips = 80% of full-tund performance
Zero-shot	Very fast, but accuracy drops to ~0–5% macro-F1 due to σ gap	Works only when source and target
Full fine-tuning	78–84% macro-F1; SVM ~37 s NC/KNN in milliseconds	NC/KNN reach SVM-like accuracy with much lower training cost

Random Forest and **Logistic Regression** heads reach ~0.86–0.88 macro-F1 after full ESC-50 fine-tuning, showing that **frozen YAMNet embeddings + lightweight classifiers are sufficient for strong performance.**

Conclusion

1. **Lighter Classifiers Outperform:** Frozen YAMNet embeddings plus lightweight heads can deliver strong animal-sound recognition: zero-shot is weak, but even 5 labeled clips/class push macro-F1 above 0.6 and full ESC-50 fine-tuning exceeds 0.85.

2. **Label-efficiency Insight:** with only ~250 labeled samples the models already become useful, making this pipeline attractive for wildlife projects with scarce annotations.

3. **Benefits of Pretrained Models:** Using pretrained audio embeddings and simple classifiers gives a fast, reproducible starting point, and future work can swap in richer source domains or semi-supervised techniques to push accuracy further.

Limitations & Future Work

All experiments rely on a single frozen feature extractor (YAMNet) and two datasets, so **results may not generalize to other animal sound distributions** or benefit from end-to-end adaptation.

Improvements could include **testing other source corpora** (AudioSet animal subsets, UrbanSound8K) for closer domain alignment, **exploring semi-supervised label propagation** on ESC-50, and **benchmarking efficient heads** (e.g., prototypical networks) for faster adaptation.

Acknowledgements: YAMNet authors for pretrained weights; ESC-50 (Piczak, 2015), GTZAN (Tzanetakis & Cook, 2002), and Sound Understanding in Google AI Perception for their work on AudioSet/YAMNet authors.