Recording session leakage inflates accuracy in marine mammal call classification

Nov 11, 2026 · @Erzsebet Frey

Erzsebet Frey, Independent researcher. ORCID 0009-0002-7482-7986. Correspondence: info@simplyecologist.com.

Abstract

Splitting clips at random overstated the accuracy of a marine mammal call classifier by up to 23 percentage points, because clips from the same recording tape landed in both training and test data. I classified 4,761 recordings of nine Arctic and subarctic marine mammal species from the Watkins Marine Mammal Sound Database, recorded on 148 tapes between 1949 and 1997, at two levels: species and taxonomic group (baleen whales, toothed whales, pinnipeds). A random forest on 90 acoustic features a small transformer trained from scratch on log mel spectrograms and linear classifiers on embeddings from the pretrained BirdNET and Perch 2.0 models were evaluated with 5 fold cross validation grouped by tape, and the random forest and both embedding classifiers were also scored with random splits.

With random splits the random forest reached 97% accuracy for both tasks. With whole tapes held out, species accuracy fell to 74% (balanced accuracy 49%, against 89% with random splits), while group accuracy fell to 91% (balanced accuracy 71%). The loss was concentrated in species recorded on few tapes: F1 for narwhal fell from 0.91 to 0.02 and for ringed seal from 0.73 to 0.00, with most of their clips predicted as killer whale. The transformer matched the random forest on balanced accuracy for groups (71%) but trailed it for species (42%). Embeddings from the pretrained BirdNET and Perch 2.0 models did best on unseen tapes (species balanced accuracy 66% and 69%) but were still inflated by 28 and 26 points under random splits. Unsupervised clustering recovered neither species nor groups.

Random splits of clip level bioacoustic data therefore measure partly how well a model recognises recording conditions, not animals, even when the model is pretrained. I recommend grouping cross validation folds by recording session, reporting balanced accuracy and per class scores, and reporting the random split score alongside as a measure of leakage.

Keywords: bioacoustics, passive acoustic monitoring, data leakage, cross validation, marine mammals, Arctic, BirdNET, transfer learning, random forest, transformer, Watkins Marine Mammal Sound Database

1. Introduction

Passive acoustic monitoring now produces far more audio than people can annotate, and machine learning classifiers are increasingly used to detect and identify marine mammals in it (Gibb et al. 2019; Shiu et al. 2020; Stowell 2022). How well such a classifier will work on new deployments is usually estimated by cross validation on a labelled archive.

The estimate is only honest if training and test data are independent. Clips cut from one recording share the hydrophone, the recording chain, the background noise and often the same individual animals. When clips are assigned to folds at random, clips from one recording can sit on both sides of the split, and a model can score well by recognising the recording rather than the species. Ecologists know this problem as non independence in structured data (Roberts et al. 2017), and the wider machine learning literature calls it data leakage (Kaufman et al. 2012; Kapoor and Narayanan 2023).

The Watkins Marine Mammal Sound Database is one of the most widely used open archives of marine mammal sounds (Sayigh et al. 2016) and appears in general bioacoustic benchmarks such as BEANS (Hagiwara et al. 2023). Its file names encode the year and tape of each recording, which makes it possible to measure the effect of leakage directly.

Published classifiers on this archive have mostly been evaluated with random splits of clips. BEANS split its Watkins subset at random, with stratification, into training, validation and test sets (Hagiwara et al. 2023), and other studies used stratified random holdouts or random k fold cross validation (Cai et al. 2022; Licciardi and Carbone 2024), reporting accuracies between 87% and 98%. Aguiar et al. (2021) showed on the "best cuts" subset that accuracy fell from 78% under random ten fold cross validation to 24% when recording locations were kept apart, but did not examine which species drive the loss or whether pretrained models are affected.

This study asks three questions on 4,761 Watkins recordings of nine Arctic and subarctic species: (1) how much do random splits overstate accuracy compared with splits that hold out whole tapes; (2) which species are most affected; (3) does a small transformer trained from scratch on spectrograms outperform a random forest on hand crafted acoustic features under the stricter evaluation; and (4) are embeddings from a pretrained bioacoustic model also inflated by random splits.

2. Methods

2.1 Data

I used every recording of nine species in the "full cuts" collection of the Watkins Marine Mammal Sound Database (Sayigh et al. 2016), accessed through its public Internet Archive mirror because the Woods Hole site was offline. Six species are Arctic or

subarctic residents (bowhead whale *Balaena mysticetus*, narwhal *Monodon monoceros*, beluga *Delphinapterus leucas*, bearded seal *Erignathus barbatus*, ringed seal *Pusa hispida*, walrus *Odobenus rosmarus*); three wider ranging species were added for comparison (humpback whale *Megaptera novaeangliae*, North Atlantic right whale *Eubalaena glacialis*, killer whale *Orcinus orca*). Species were also grouped into baleen whales, toothed whales and pinnipeds.

The 4,761 clips were recorded between 1949 and 1997. Their median length was 2.0 s (interquartile range 1.3 to 3.2 s) and they had been digitised at 20 different sample rates between 1.28 and 81.92 kHz. Killer whales made up 56% of clips, and ringed seal only 1% (Figure 1). Labels were taken from the database's species folders and were not reverified.

2.2 Recording sessions

Watkins file names begin with two digits for the year and three for the tape, so I defined a recording session as species plus the first five characters of the file name. This gave 148 sessions, from 5 (ringed seal) to 65 (killer whale) per species.

2.3 Features

Each clip was mixed to mono and resampled to 22,050 Hz. For the random forest I computed 90 features with librosa (McFee et al. 2015): the mean and standard deviation over time of 20 mel frequency cepstral coefficients and their deltas, and of spectral centroid, bandwidth, rolloff, zero crossing rate and RMS energy. For the transformer I computed a 64 band log mel spectrogram (FFT size 1,024, hop 256) of the first 128 frames (about 1.5 s), padded where shorter, standardised within each clip and clipped to plus or minus 4 standard deviations. About 34% of clips were shorter than 1.5 s.

As the first pretrained model I used BirdNET version 2.4 (Kahl et al. 2021), a convolutional network trained mainly on bird sounds whose embeddings transfer well to other taxa, including marine mammals (Ghani et al. 2023). Each clip was resampled to 48 kHz and cut into 3 s windows, the last one padded with silence, and the 1,024 dimensional embeddings of the first ten windows were averaged per clip.

I also used Perch 2.0 (van Merriënboer et al. 2025), a newer bioacoustic model trained on about 15,000 species, mostly birds, that transfers well to marine mammal tasks despite having almost no marine training data (Burns et al. 2025). Each clip was resampled to 32 kHz and cut into 5 s windows, the last one padded with silence, and the 1,536 dimensional embeddings of the first six windows were averaged per clip, so both models saw up to the first 30 s of each clip.

2.4 Models

Unsupervised check. Standardised features were reduced to 10 principal components and clustered with k means for k from 2 to 12; k was chosen by silhouette score and agreement with labels measured with the adjusted Rand index (Hubert and Arabie 1985).

Random forest. 400 trees with balanced class weights (Breiman 2001; scikit-learn, Pedregosa et al. 2011).

Transformer. A vision style transformer (Dosovitskiy et al. 2021) trained from scratch in PyTorch (Paszke et al. 2019): 8 by 8 spectrogram patches, a learned class token and position embeddings, 64 dimensional embeddings, 2 encoder layers with 4 attention heads, and dropout 0.3. It was trained for 25 epochs with class weighted cross entropy, AdamW (learning rate 5e-4, weight decay 0.01), a cosine learning rate schedule, batch size 64, and one random frequency mask and one time mask per clip (Park et al. 2019). The model after the last epoch was evaluated, so no test data was used for model selection.

Pretrained embeddings. A logistic regression (L2 penalty, C = 0.1, balanced class weights) on standardised BirdNET embeddings, the linear probe usually used to evaluate such embeddings (Ghani et al. 2023). C was fixed beforehand, not tuned. A random forest with the settings above was also fitted to the embeddings for comparison.

The same logistic regression, with the same settings, was fitted to the Perch 2.0 embeddings.

2.5 Evaluation

All models were evaluated with 5 fold cross validation. In the grouped scheme (StratifiedGroupKFold) every session falls entirely in one fold, so test clips always come from tapes the model has not seen. In the random scheme (StratifiedKFold) clips were assigned to folds regardless of session. The random forest and the pretrained embedding models were scored under both schemes on both tasks; the transformer, which took about 33 minutes per task on two CPU cores, was scored under the grouped scheme only.

I report accuracy, balanced accuracy (mean recall per class) and macro F1 as mean and standard deviation over folds, and per species F1 from out of fold predictions. Because one species dominates, balanced accuracy is the main metric; the majority class baseline is 56% accuracy for species and 61% for groups, and chance balanced accuracy is 11% and 33%. All code is available (Section 6).

3. Results

3.1 Random splits inflated accuracy by up to 23 points

Under random splits the random forest classified species with 97% accuracy and 89% balanced accuracy. Holding out whole tapes lowered accuracy to 74% and balanced accuracy to 49%, a loss of 23 and 41 percentage points (Table 1, Figure 2). Group classification was more robust but still lost 7 points of accuracy (97% to 91%) and 21 points of balanced accuracy (92% to 71%). Standard deviations across folds were about 3 to 20 times larger under grouped splits, because each fold then contains a different set of tapes.

Table 1. 5 fold cross validation, mean (SD) over folds. Grouped = whole recording sessions held out.

Task	Model	Split	Accuracy	Balanced accuracy	Macro F1
9 species	Majority class	none	56%	11%	
9 species	Random forest	Random	97% (0.4)	89% (2.6)	0.92
9 species	Random forest	Grouped	74% (8.9)	49% (6.9)	0.43
9 species	Transformer	Grouped	59% (10.7)	42% (5.6)	0.35
9 species	BirdNET + logistic regression	Random	98% (0.4)	95% (1.2)	0.95
9 species	BirdNET + logistic regression	Grouped	80% (3.7)	66% (11.1)	0.51
9 species	Perch 2.0 + logistic regression	Random	98% (0.5)	96% (0.6)	0.95
9 species	Perch 2.0 + logistic regression	Grouped	79% (4.1)	69% (7.0)	0.56
3 groups	Majority class	none	61%	33%	
3 groups	Random forest	Random	97% (0.6)	92% (2.8)	0.94
3 groups	Random forest	Grouped	91% (4.5)	71% (7.9)	0.70
3 groups	Transformer	Grouped	82% (7.2)	71% (7.3)	0.68
3 groups	BirdNET + logistic regression	Random	99% (0.4)	97% (1.5)	0.97
3 groups	BirdNET + logistic regression	Grouped	90% (4.6)	81% (8.1)	0.77
3 groups	Perch 2.0 + logistic regression	Random	99% (0.4)	98% (1.3)	0.98
3 groups	Perch 2.0 + logistic regression	Grouped	91% (4.4)	83% (6.7)	0.78

3.2 Species recorded on few tapes lost the most

The drop was uneven across species (Figure 3). Killer whale (65 tapes) kept an F1 of 0.93 and bowhead whale (6 tapes) 0.74. Narwhal fell from 0.91 to 0.02 and ringed seal from 0.73 to 0.00; both were recorded on 5 or 6 tapes. On unseen tapes, 70% of narwhal clips and 63% of ringed seal clips were predicted as killer whale, and a further 20% of narwhal clips

were predicted as ringed seal and 35% of ringed seal clips as narwhal. Humpback whale (14 tapes) fell from 0.97 to 0.28, with 54% of its clips predicted as North Atlantic right whale, and 43% of walrus clips were predicted as bowhead whale.

The most important features for both tasks were the mean spectral bandwidth, centroid and rolloff and the lower order MFCC means, which summarise where the energy of a call sits in frequency.

3.3 The transformer matched the random forest only for groups

Under grouped cross validation the transformer reached the same balanced accuracy as the random forest for groups (71% each) with lower accuracy (82% vs 91%). For species it trailed on every metric: 59% accuracy against a 56% majority baseline, and 42% balanced accuracy against 49% for the random forest (Table 1, Figure 4). Its fold to fold spread was similar to the random forest's.

3.4 Pretrained embeddings narrowed the gap but did not close it

The linear classifier on BirdNET embeddings was the best model on unseen tapes: 66% balanced accuracy for species and 81% for groups, against 49% and 71% for the random forest on hand crafted features (Table 1). It was still inflated by random splits, which gave 95% and 97%, a gap of 28 and 16 points. On unseen tapes narwhal reached an F1 of 0.36 and ringed seal 0.50, against 0.92 and 0.89 with random splits, while killer whale stayed at 0.95. A random forest on the same embeddings did worse under both schemes (species balanced accuracy 43% grouped and 58% random).

Perch 2.0 embeddings did slightly better than BirdNET on unseen tapes, with 69% balanced accuracy for species and 83% for groups, and showed almost the same inflation: random splits gave 96% and 98%, a gap of 26 and 15 points. Ringed seal improved to an F1 of 0.80 on unseen tapes, but narwhal fell to 0.22.

3.5 No unsupervised structure

k means on the first 10 principal components chose k = 2 by silhouette score. Agreement with labels was at or below chance (adjusted Rand index −0.07 for species and −0.02 for groups), and the three groups overlapped heavily in the first two principal components (Figure 5).

Figures

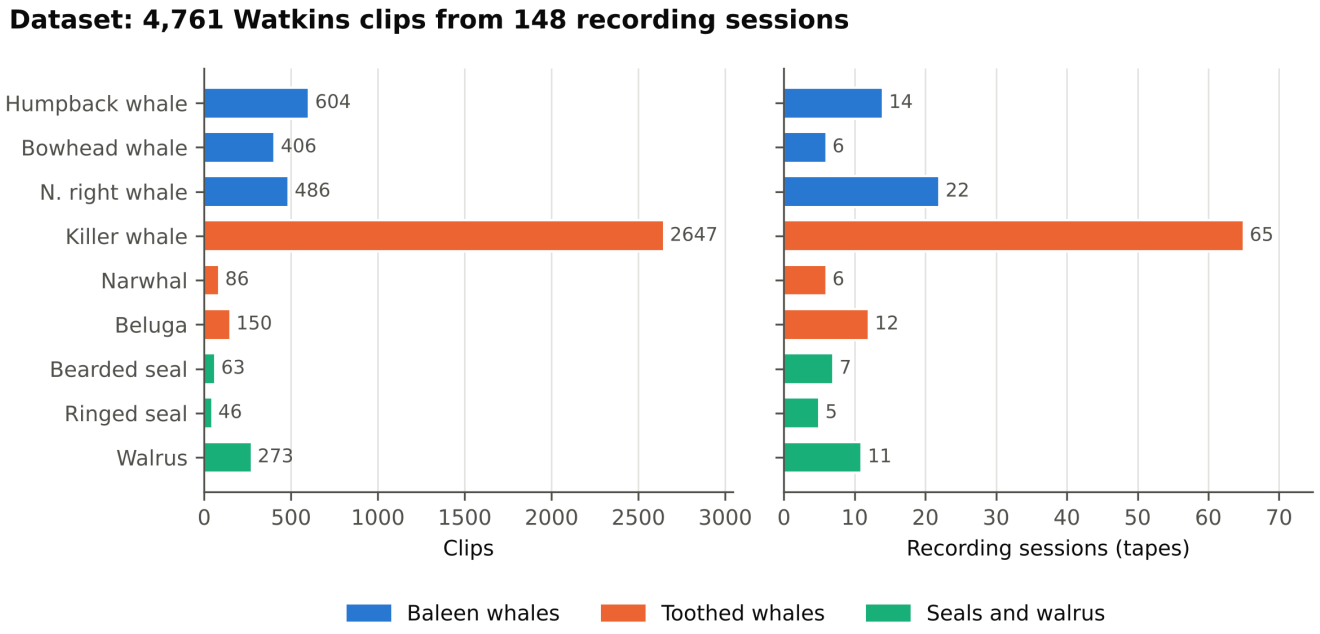


Figure 1. Clips (left) and recording sessions (right) per species, coloured by taxonomic group.

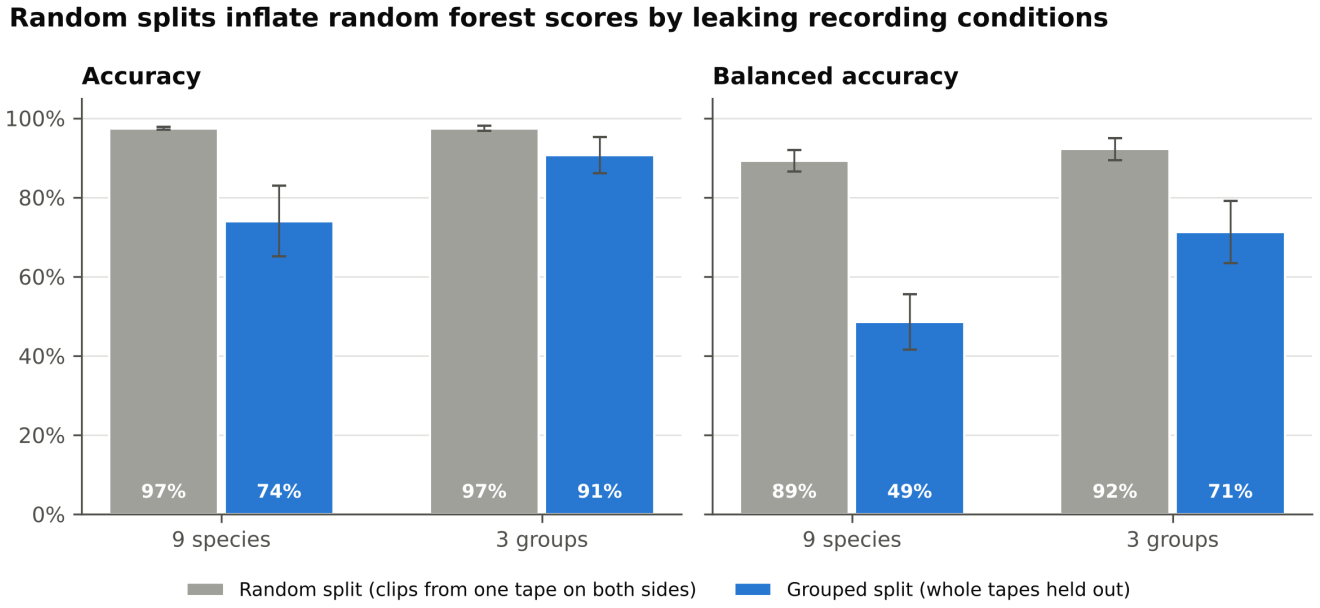


Figure 2. Random forest accuracy and balanced accuracy under random splits (grey) and with whole recording sessions held out (blue). Error bars show the standard deviation over 5 folds.

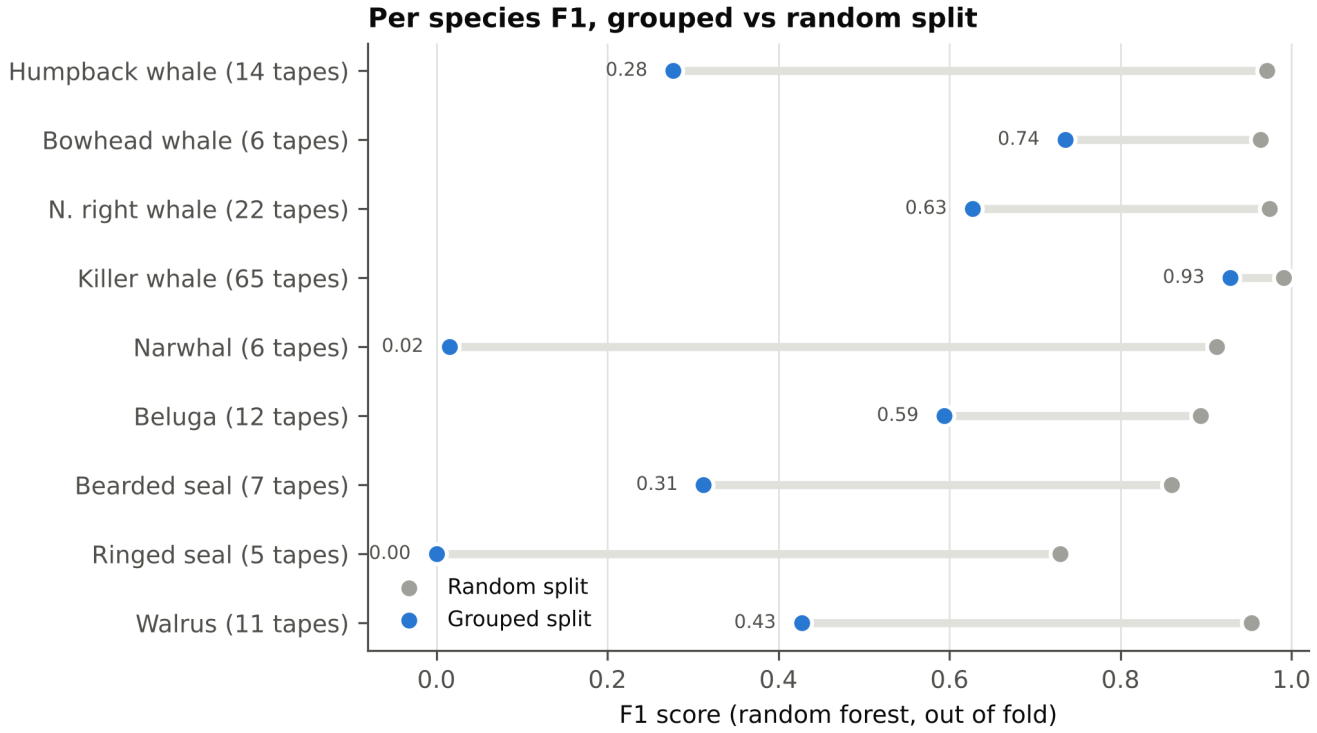


Figure 3. Per species F1 of the random forest from out of fold predictions, random (grey) and grouped (blue) splits. Numbers of recording sessions in brackets.

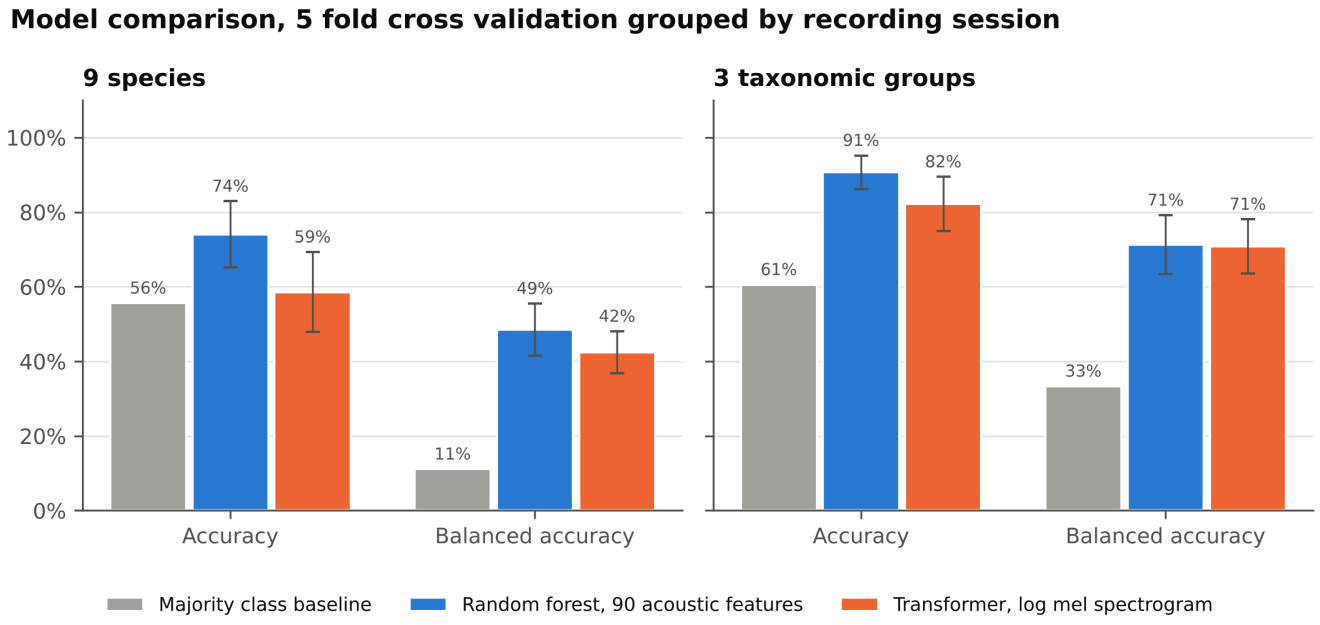


Figure 4. Majority class baseline, random forest and transformer under 5 fold cross validation grouped by recording session.

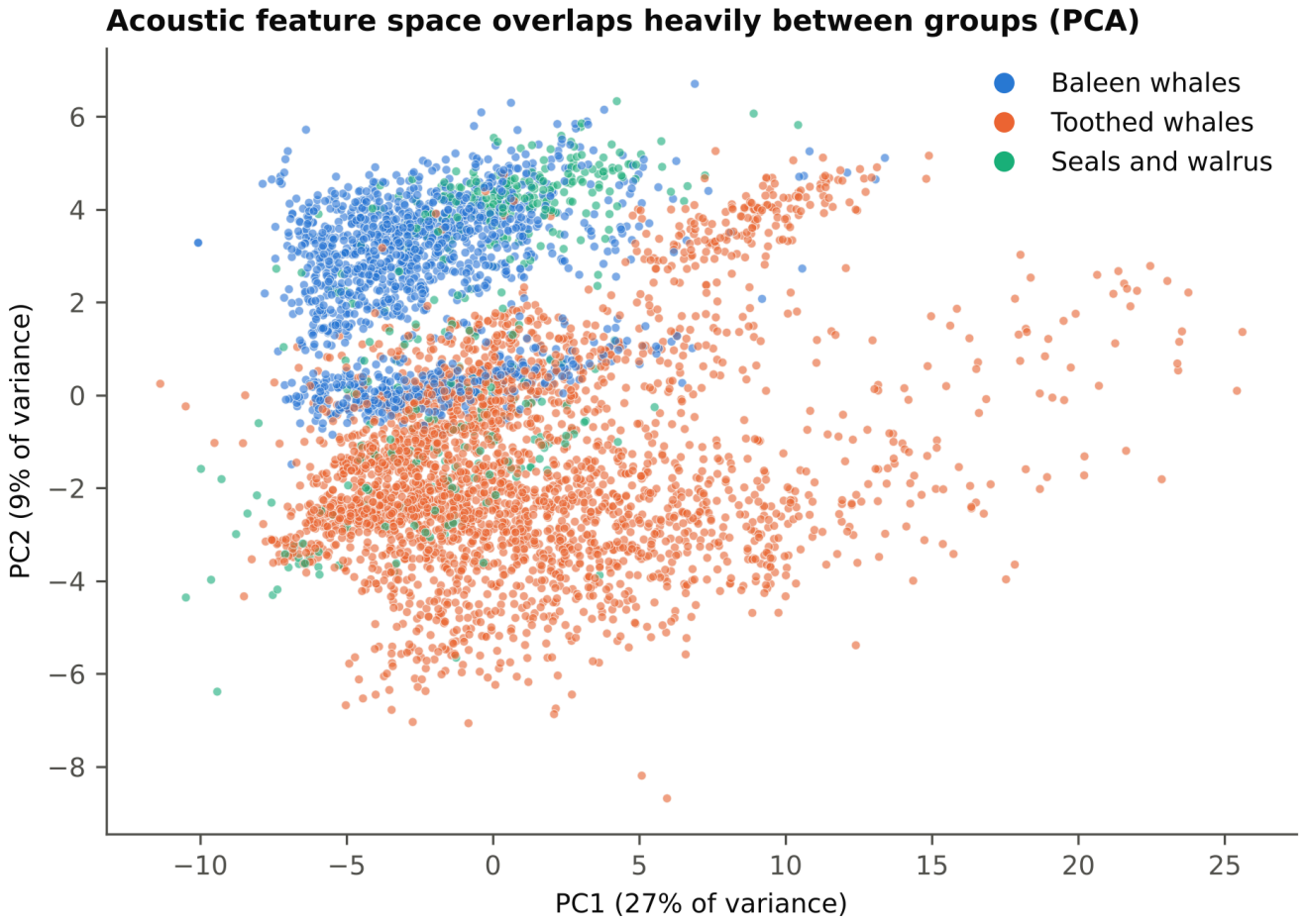


Figure 5. First two principal components of the 90 standardised acoustic features for all 4,761 clips, coloured by taxonomic group.

4. Discussion

A random split of these clips mostly measured how well the classifier recognised tapes. The same random forest, on the same features, looked nearly perfect at the species level under random splits and mediocre under grouped splits, and the gap was largest for exactly the species an Arctic monitoring program would care about: narwhal and ringed seal, each represented by a handful of tapes.

The pattern of errors points to recording conditions rather than biology. Under grouped splits narwhal and ringed seal were confused with each other and with killer whale, species from different families whose calls differ widely in frequency and structure. Ringed seal clips were also very short (median 0.28 s), so duration and the noise floor of a few tapes may carry more of the signal than the calls themselves. Other confusions were more plausible acoustically, such as humpback whale with North Atlantic right whale, two baleen whales with low frequency calls.

The taxonomic group task held up better because each group spans many tapes and broad frequency differences survive a change of recording chain. Even so, its balanced accuracy fell by 21 points, so leakage affects coarse tasks too.

The transformer result is a caution against reading too much into small, clip level comparisons. In an earlier version of this analysis on 225 clips (Frey 2026), a single held out split put the transformer slightly ahead of the random forest. With all clips and grouped cross validation the transformer only matched the random forest for groups and lost for species. A transformer trained from scratch on a few thousand 1.5 s spectrograms is a weak baseline; pretrained models did better here, with BirdNET and Perch 2.0 embeddings and a linear classifier reaching 66% and 69% species balanced accuracy on unseen tapes.

Pretraining did not remove the leakage effect. This held for Perch 2.0 as well, a model that outperforms specialised marine models on underwater tasks (Burns et al. 2025), whose species score was inflated by 26 points. Even with embeddings learned from large bird sound collections, random splits overstated species balanced accuracy by 28 points: features that describe the acoustic texture of a recording describe its tape as well as its caller. This matters for the Watkins literature, where most reported accuracies come from random splits of clips (Cai et al. 2022; Hagiwara et al. 2023; Licciardi and Carbone 2024). Those numbers are likely to overstate performance on new recordings, as Aguiar et al. (2021) found when they held out recording locations.

4.1 Limitations

Tapes are an imperfect proxy for independent recordings: tapes from the same expedition, year or region may still share conditions, so the grouped scores may themselves be optimistic. The archive is historical, unevenly sampled and labelled only at the folder level, and clips vary from under 0.1 s to several minutes. Only one feature set, one forest configuration, one small transformer and two pretrained models were tested, without tuning. BirdNET and Perch 2.0 were trained mainly on bird and other terrestrial sounds, so a model trained on marine mammals may describe very low frequency baleen whale calls and very high frequency toothed whale clicks better. The results describe this archive and should not be read as the accuracy expected on modern passive acoustic monitoring data.

4.2 Recommendations

1. Assign cross validation folds by recording session, deployment or site, never by clip, whenever clips share a recording.
2. Report the random split score alongside the grouped score; the gap is a direct measure of how much the model depends on recording conditions.

3. Report balanced accuracy and per class scores, because overall accuracy hides the collapse of rare classes behind a dominant one.
4. Report the number of independent recordings per class, not only the number of clips.

5. Conclusion

On 4,761 Watkins recordings, holding out whole tapes cut species level balanced accuracy from 89% to 49% and erased the apparent ability to recognise narwhal and ringed seal. Pretrained BirdNET and Perch 2.0 embeddings did best on unseen tapes but were still inflated by 26 to 28 points under random splits. Grouped evaluation by recording session is a simple change that gives a far more honest estimate of how a classifier will perform on new recordings.

6. Data and code availability

The recordings are from the Watkins Marine Mammal Sound Database (Woods Hole Oceanographic Institution and the New Bedford Whaling Museum), retrieved from the Internet Archive mirror. The clip list, download, feature extraction, pretrained embedding, analysis and figure code, all results and all figures are archived on Zenodo as version 1.2.0 (doi:10.5281/zenodo.24158120; all versions doi:10.5281/zenodo.22859088) and on GitHub. Code is under the MIT licence and derived results under CC BY 4.0.

Acknowledgements

I thank the curators of the Watkins Marine Mammal Sound Database and the Internet Archive for keeping the recordings openly available.

References

Aguiar R, Maguolo G, Nanni L, Costa Y, Silla C. 2021. On the importance of passive acoustic monitoring filters. *Journal of Marine Science and Engineering* 9: 685. doi:10.3390/jmse9070685

Breiman L. 2001. Random forests. *Machine Learning* 45: 5–32. doi:10.1023/A:1010933404324

Burns A, Harrell L, van Merriënboer B, Dumoulin V, Hamer J, Denton T. 2025. Perch 2.0 transfers "whale" to underwater tasks. *NeurIPS 2025 Workshop on AI for Non Human Animal Communication*. arXiv:2512.03219

Cai W, Zhu J, Zhang M, Yang Y. 2022. A parallel classification model for marine mammal sounds based on multi dimensional feature extraction and data augmentation. *Sensors* 22: 7443. doi:10.3390/s22197443

Dosovitskiy A, et al. 2021. An image is worth 16x16 words: transformers for image recognition at scale. *International Conference on Learning Representations (ICLR 2021)*.

Frey E. 2026. Arctic and subarctic marine mammal call classification (v1.0.0). Zenodo. doi:10.5281/zenodo.22859089

Ghani B, Denton T, Kahl S, Klinck H. 2023. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports* 13: 22876. doi:10.1038/s41598-023-49989-z

Gibb R, Browning E, Glover-Kapfer P, Jones KE. 2019. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution* 10: 169–185. doi:10.1111/2041-210X.13101

Hagiwara M, Hoffman B, Liu JY, Cusimano M, Effenberger F, Zacarian K. 2023. BEANS: the benchmark of animal sounds. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*. arXiv:2210.12300

Hubert L, Arabie P. 1985. Comparing partitions. *Journal of Classification* 2: 193–218. doi:10.1007/BF01908075

Kahl S, Wood CM, Eibl M, Klinck H. 2021. BirdNET: a deep learning solution for avian diversity monitoring. *Ecological Informatics* 61: 101236. doi:10.1016/j.ecoinf.2021.101236

Kapoor S, Narayanan A. 2023. Leakage and the reproducibility crisis in machine learning based science. *Patterns* 4: 100804. doi:10.1016/j.patter.2023.100804

Kaufman S, Rosset S, Perlich C, Stitelman O. 2012. Leakage in data mining: formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data* 6: 15. doi:10.1145/2382577.2382579

Licciardi A, Carbone D. 2024. WhaleNet: a novel deep learning architecture for marine mammals vocalizations on Watkins Marine Mammal Sound Database. *IEEE Access* 12: 154182–154194. doi:10.1109/ACCESS.2024.3482117

McFee B, et al. 2015. librosa: audio and music signal analysis in Python. *Proceedings of the 14th Python in Science Conference*: 18–25.

Park DS, Chan W, Zhang Y, Chiu CC, Zoph B, Cubuk ED, Le QV. 2019. SpecAugment: a simple data augmentation method for automatic speech recognition. *Interspeech 2019*: 2613–2617. doi:10.21437/Interspeech.2019-2680

Paszke A, et al. 2019. PyTorch: an imperative style, high performance deep learning library. *Advances in Neural Information Processing Systems* 32.

Pedregosa F, et al. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Roberts DR, et al. 2017. Cross validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40: 913–929. doi:10.1111/ecog.02881

Sayigh L, Daher MA, Allen J, Gordon H, Joyce K, Stuhlmann C, Tyack P. 2016. The Watkins Marine Mammal Sound Database: an online, freely accessible resource. *Proceedings of Meetings on Acoustics* 27: 040013. doi:10.1121/2.0000358

Shiu Y, et al. 2020. Deep neural networks for automated detection of marine mammal species. *Scientific Reports* 10: 607. doi:10.1038/s41598-020-57549-y

Stowell D. 2022. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ* 10: e13152. doi:10.7717/peerj.13152

van Merriënboer B, Dumoulin V, Hamer J, Harrell L, Burns A, Denton T. 2025. Perch 2.0: the Bittern lesson for bioacoustics. arXiv:2508.04665