Classifying Lake Trout vs Smallmouth Bass from Wideband Hydroacoustics (45–170 kHz)

ETC5543 — Business Analytics Creative Activity (Single-student project)

Dulitha Perera

Report for

27 October 2025

Classifying Lake Trout vs Smallmouth Bass from Wideband Hydroacoustics (45–170 kHz)
ETC5543 — Business Analytics Creative Activity (Single-student project)

Table of contents

Classifying Lake Trout vs Smallmouth Bass from Wideband Hydroacoustics (45–170 kHz)
ETC5543 — Business Analytics Creative Activity (Single-student project)

1 Abstract

Wideband hydroacoustics enables non-invasive monitoring of fish populations, but reliable species-level identification remains challenging when visual confirmation is impossible. This project investigates whether frequency-only acoustic signals (45-170 kHz) can accurately distinguish Lake Trout (LT) and Smallmouth Bass (SMB). Building on earlier work that used a recurrent neural network (RNN) for the same dataset, we first replicate that baseline and then extend the analysis using a broader, leakage-safe machine-learning framework.

We summarise each fish's frequency response curve (FRC) using quantiles and tsfeatures time-series descriptors, then apply H2O AutoML across multiple model families under grouped validation by fish identifier. The best model achieves strong out-of-sample performance (AUC almost 0.95; accuracy almost 0.90) with the upper-mid frequency band (140--160 kHz) contributing most to discrimination. Targeted tuning---via out-of-fold threshold optimisation, deep-learning grid search, and frequency-selector features---further improves test accuracy and interpretability.

Results demonstrate that wideband frequency-only signatures can separate species with high reliability, providing a reproducible and operationally deployable workflow for acoustic classification. All analyses are fully scripted in R using `renv` for dependency control and Git LFS for large data management.

2 Introduction & Motivation

Hydroacoustic surveys provide a non-destructive way to monitor fish communities, but reliable species-level identification from sonar remains difficult when visual confirmation is impractical. Wideband transducers measure target strength (TS) across many frequencies, giving each fish an acoustic "fingerprint" or frequency response curve (FRC). If these frequency-only signals can separate species accurately, managers can obtain species-resolved indices without netting or tagging.

Prior work (baseline). A related study on the same survey family applied several neural architectures---including fully-connected, convolutional, and recurrent neural networks (RNN/LSTM)---to size-standardised backscatter for species classification (**hydro_nn_internal_2025**). To ensure a fair comparison, we first replicate the RNN as a baseline (03a_rnn_reproduction.R) and also run AutoML on the same input representation (03_classification_original.R) to establish non-NN references. We then extend the methodology substantially.

Our approach (replicate extend).

1. **Representation**. Summarise per-fish FRCs (45--170 kHz) using **quantiles** (q20--q100) and compute tsfeatures descriptors from short per-fish sequences, producing four datasets (quintiles/median x with/without raw F*).

2. **Model breadth**. Benchmark H2O families (GLM, RF, GBM, DeepLearning) via AutoML with grouped validation by fishNum to avoid leakage, reserving an unseen test set.

3. **Targeted tuning**. Convert the AutoML exploration to a single deployable model through out-of-fold threshold tuning (policy clamp [0.40,0.70][0.40,0.70]), a deep-learning grid for the top DL family, and frequency selectors (train-only top-K F*) to study which parts of the spectrum matter.

4. **Interpretability**. Quantify which frequencies drive separation and relate them to plausible mechanisms (e.g., bladder resonance, orientation).

Contributions.

A **frequency-only classification pipeline** with leakage-aware evaluation (grouped by fish/region) and a fixed operational threshold.

A **systematic representation study** (quantiles, tsfeatures, top-K F*) extending beyond the earlier RNN-only approach.

Clear **feature-importance** summaries highlighting discriminative frequency regions.

A **reproducible** R codebase (renv + Git LFS) with scripted outputs and one-shot execution.

Research questions

Primary RQ1. Can Lake Trout vs Smallmouth Bass be classified **from FRC-only signals (45--170 kHz)** using **grouped validation** and an unseen test set, and how do results compare with the reproduced RNN baseline?

Secondary RQ2. Which **frequency regions** and **representations** (band means/ratios, PCA of FRC, tsfeatures, top-K F*) contribute most to separation?

RQ3. What **descriptive** depth/orientation/behaviour patterns accompany species labels (not used for prediction)?

RQ4. What are the key **limitations** (orientation, sampling bias, leakage risks) and how should future surveys/classifier deployment adapt?

Non-goals. We do **not** use morphometrics (length/weight) for prediction; temporal/spatial analyses are **descriptive only**.

3   Data & Preparation

3.1   Source and structure

The dataset is provided as an RDS file (`TSresponse_clean.RDS`, tracked via Git LFS) with over **30k** rows and **302** variables. Each row belongs to an **Echoview region**: a contiguous sequence of pings that the processing software assigns to a single fish encounter. Two identifiers link the data:

- `fishNum` --- unique individual; LT/SMB prefix encodes species.

- `Region_name` --- encounter identifier within a fish.

The block `F45..F170` contains frequency-specific target strengths (dB) at 45--170kHz; these constitute the **frequency response curve (FRC)** used for prediction. Additional variables describe geometry/behaviour (e.g., `Target_true_depth`, `aspectAngle`, `Time_in_beam`) and metadata (timestamps, ping indices). A concise glossary appears in Appendix A.

3.2   Scope decisions

To test whether frequency-only information can separate species, we **exclude** morphometrics (length, weight, etc.) from all predictive models. Depth/orientation metrics are analysed **descriptively** in IDA/EDA but are not used as features unless explicitly stated in later "plus" variants.

3.3   Cleaning and basic checks

We perform light cleaning before feature construction:

1. **Type & order**. Ensure `F*` columns are numeric and ordered by frequency; drop corrupted rows.

2. **Species label**. Standardise to two classes: Lake Trout (LT) and Smallmouth Bass (SMB).

3. **Duplicates**. Remove any accidental duplicate rows (exact key or repeated ping).

4. **Sanity checks**. Count per-species records; check frequency coverage and missingness across `F*`.

(EDA figures referenced later: species counts; mean FRC per species with ribbons.)

3.4   Fish-level representations (what we train on)

Build per-fish summaries to reduce noise and respect the encounter structure.

- **Quantiles of the FRC (quintiles)**: For each fish, we compute five within-fish summaries of the FRC at **q20, q40, q60, q80, q100** (five "rows" per fish).

  - Output artifact: `outputs/tables/fish_freq_quintiles_long.rds`.

- This retains frequency resolution (columns `F45..F170`) while stabilising per-ping variability.

- **Median FRC**: A single row per fish using the within-fish median of each `F*`.

Used to create compact, one-row-per-fish variants.

- **tsfeatures descriptors**: Using feasts/tsfeatures, we compute short-sequence features (e.g., ACF summaries) from per-fish frequency traces. We produce four datasets:

  1. **quintiles_allfreq_tsfeat** --- 5 rows/fish: raw `F*` + tsfeatures

  2. **quintiles_tsfeat_only** --- 5 rows/fish: tsfeatures only

  3. **median_allfreq_tsfeat** --- 1 row/fish: median `F*` + tsfeatures

  4. **median_tsfeat_only** --- 1 row/fish: tsfeatures only

Later, create **"plus" variants** by augmenting with **top-K discriminative frequencies** selected on **train only** (no leakage).

3.5   Train/validation/test design (leakage-aware)

- **Grouping**. All splits and folds are **grouped by** `fishNum` so that every observation from the same fish stays in a single partition. For quintile datasets, the five rows per fish move together.

- **Stratification**. Within groups, we stratify by species to maintain balance.

- **Holdout**. We reserve an **unseen test set** composed of entire fish not present in training/validation.

- **Cross-validation**. Model selection uses grouped k-fold CV (typically k=5).

- **Seed**. A fixed seed (73) is used for reproducibility.

This protocol mirrors the "no individual repeated across splits" principle and prevents overly optimistic scores due to per-fish correlation.

3.6   Preprocessing for modelling

- **Predictors**. Unless stated otherwise, features are the FRC block (`F45..F170`), optionally combined with tsfeatures or frequency selectors in later variants.

- **Standardisation**. Where model families benefit (e.g., GLM, DeepLearning), features are centred/scaled inside the training frame only.

- **Class label**. `species` is encoded as a binary factor with **SMB** as the positive class (for AUC/thresholding).

- **Artifacts**. Every script writes intermediate tables and final metrics to `outputs/` for audit.

3.7 Reproducibility

- **Environment**. The repository uses renv; `renv.lock` specifies exact package versions.

- **Large files**. The RDS data and any large artifacts are tracked via **Git LFS**.

- **One-shot run**. `analysis/run_all.R` reproduces the entire pipeline end-to-end.

- **Fixed randomness**. All random processes (splits, AutoML seeds) use the project seed 73.

4 Initial Data Analysis (IDA/EDA)

The dataset contains a moderate imbalance between the two species.

Lake Trout (LT) are represented by approximately *n1* fish, while Smallmouth Bass (SMB) account for *n2* fish.

Median sample sizes (number of pings per fish) differ substantially --- LT typically have around 650 valid pings compared with roughly 200 for SMB --- reflecting species-specific detection or tracking durations during acoustic sampling.

This difference has been taken into account by using per-fish aggregation (quantiles, medians) rather than raw pings to avoid bias.

Table 1: *Class balance and per-fish sample size (pings).*

species	n_fish	median_pings	iqr_pings	min_pings	max_pings	prop
LT	25	649	1533.00	115	2258	43.9%
SMB	32	202	613.25	19	1381	56.1%

Figure 1: *PCA of median FRC (one point per fish). Species separate along PC1/PC2.*
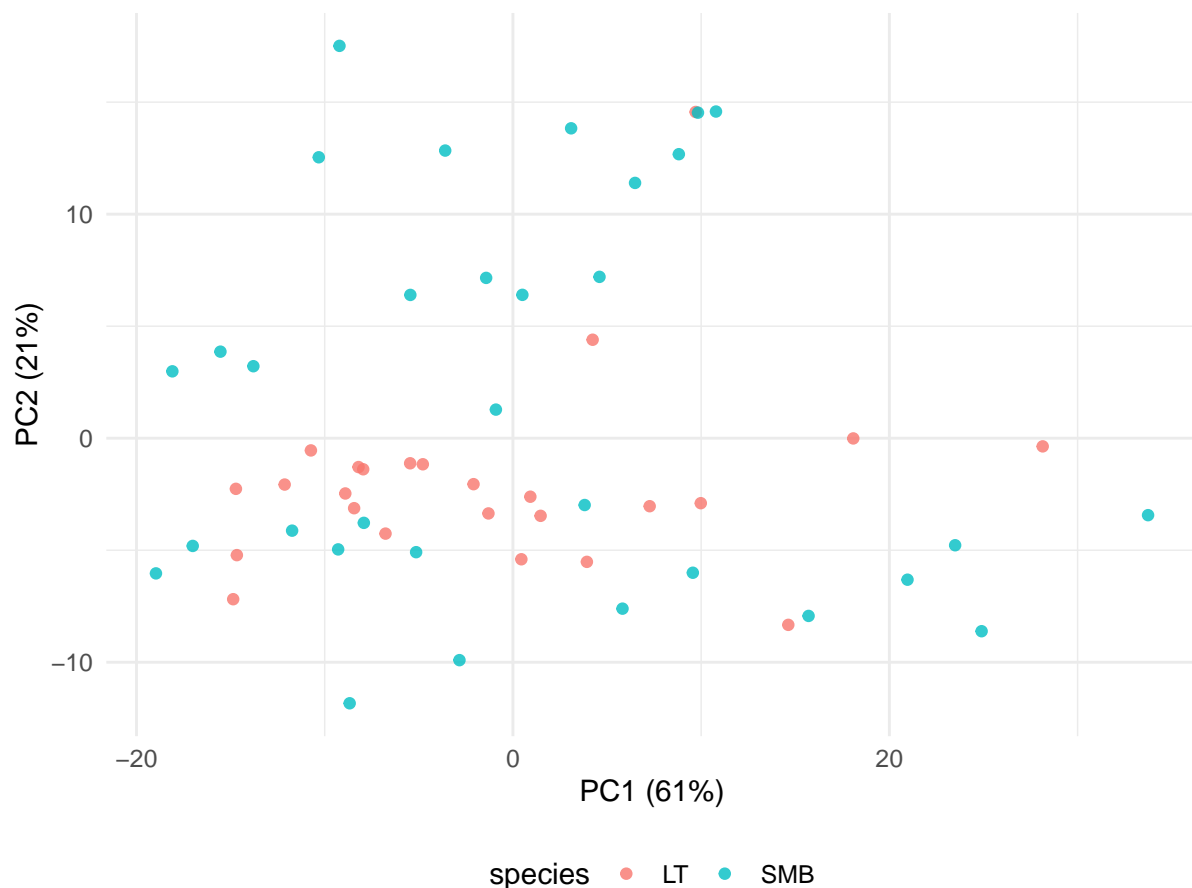
The PCA of the median frequency response curves (FRCs) shows clear separation between Lake Trout (LT) and Smallmouth Bass (SMB) along the first two principal components (PC1 and PC2).

PC1 explains about 61% of total variance and captures the overall amplitude differences in backscatter strength, while PC2 (21%) reflects finer variations in frequency-dependent patterns.

The clustering of points by species indicates that most discriminatory information is captured by the leading components, confirming that frequency-only data already contain strong species-level signals before any supervised modelling.

4.1 What is fish hydroacoustics?

Fish hydroacoustics is the study of how sound waves interact with fish underwater. A transducer emits sound pulses (pings) and records the returning echoes---how strongly a fish reflects sound depends on its body shape, tissue composition, and especially the gas-filled swim bladder. Each fish produces a unique pattern of backscatter strength across frequencies, known as a Frequency Response Curve (FRC). These FRCs can be treated like a species "acoustic fingerprint" (Figure Figure 2).

Figure 2
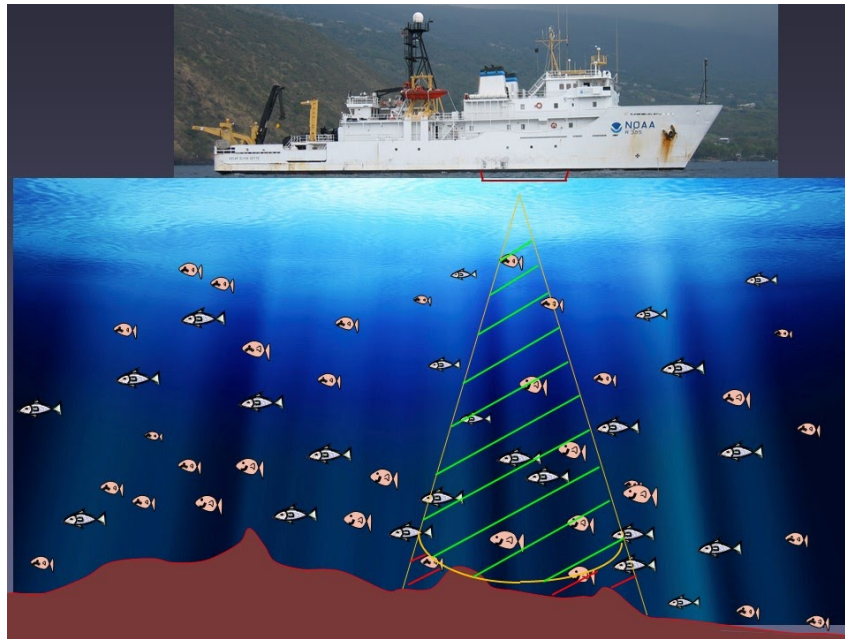
4.2 Why use hydroacoustics for classification?

Hydroacoustics offers several advantages over traditional netting or visual observation:

- **Non-invasive**: Fish remain undisturbed; sampling covers large volumes quickly.

- **Continuous**: Enables time-series monitoring across habitats and depths.

- **Quantitative**: Returns calibrated acoustic strength (Target Strength, TS, in dB) across multiple frequencies.

Because the FRC shape reflects biological differences (e.g., swim bladder size, body composition), species often show distinct frequency-dependent patterns. Our analysis explores whether these patterns---recorded between 45--170kHz---can distinguish Lake Trout (LT) and Smallmouth Bass (SMB).

4.3 Quantile envelopes

We first examine the **raw distribution of target strengths (TS)** across frequencies to understand within-species variability. Figure Figure 3 presents **per-species quantile envelopes**, where the shaded band captures the interquartile range (Q25--Q75) and dashed lines indicate additional quantiles (Q25, Q50, Q75, Q100). The clear vertical offset between Lake Trout (LT) and Smallmouth Bass (SMB) across most of the spectrum shows that separation is not driven by a few extreme observations --- instead, the difference is consistent throughout each frequency range. This highlights that species differences are evident even before aggregation, reflecting genuine signal-level distinctions in acoustic response.

Figure 3: *Frequency-response quantiles by species. Shaded bands show Q25--Q75 with the solid line as the median; dashed lines are Q25 and Q75; dotted lines show min and max. Between-species offsets persist across the full band, with the clearest separation in the upper-mid range (140--160 kHz).*
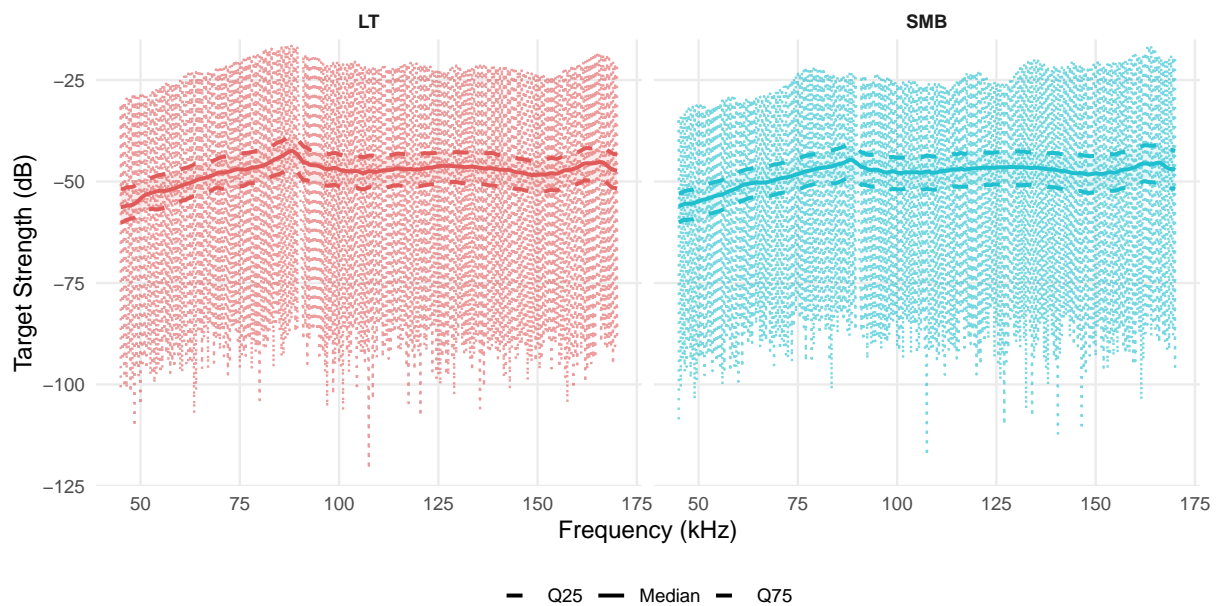
4.4 Mean frequency response by species

After examining the full distribution, we next summarise the data to compare **average acoustic profiles between species**. Figure Figure 4 shows the **mean target strength (TS)** at each frequency with a shaded **standard error (SE)** ribbon. The two mean curves diverge most noticeably in the **low-frequency band (50--120 kHz)** and again in the **upper--mid band (140--160 kHz)**, suggesting that these regions carry the most discriminative information for species separation. These frequency bands will later correspond to the model-derived importance peaks, reinforcing that the strongest signals in the raw data also drive predictive performance.

Figure 4: *Mean frequency response curves with pm SE ribbons for Lake Trout (LT) and Smallmouth Bass (SMB). Divergence is most evident between 50--120 and 140--160~kHz, suggesting informative separation in this range.*
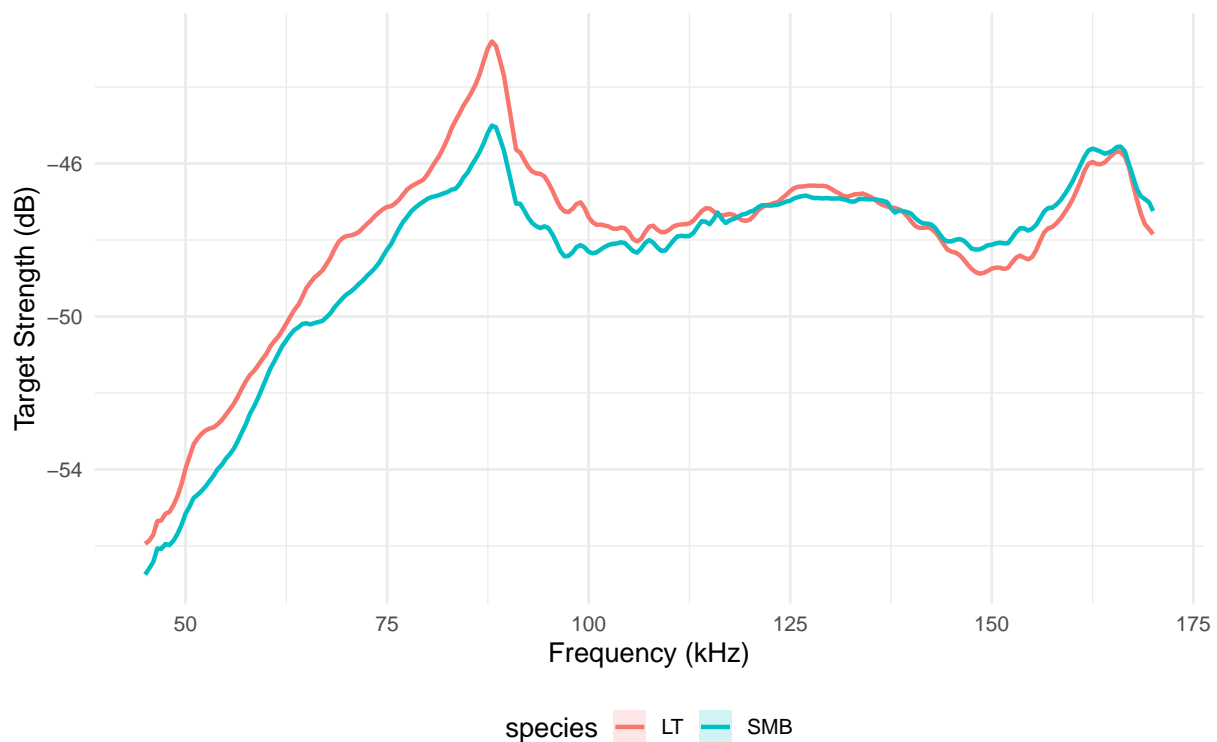

4.5 Effect size by frequency

Figure Figure 5 shows Cohen's d values comparing Lake Trout (LT) and Smallmouth Bass (SMB) at each frequency between 45--170 kHz. Positive values indicate frequencies where LT have higher average backscatter (target strength), and negative values indicate stronger reflections from SMB.

The plot reveals **moderate effect sizes** (0.2--0.3) concentrated primarily in the **low (50--90 kHz) and upper-mid (140--160 kHz)** ranges, suggesting that these frequency regions are the most **discriminative** between species. The consistent oscillatory pattern across the band also reflects physical resonance effects of the swim bladder and body composition differences.

Later in **Section 5.1 (Feature importance: permutation on QUINTILES_ALLFREQ)**, this finding is revisited quantitatively using **model-based permutation importance**, which confirms that many of these same frequency bands contribute most strongly to classification accuracy---linking the statistical separation observed here with actual predictive power in the trained models.

Figure 5: *Effect size (Cohen's d) for LT minus SMB at each frequency. Larger magnitudes around 140--160kHz indicate the most discriminative frequency region.*
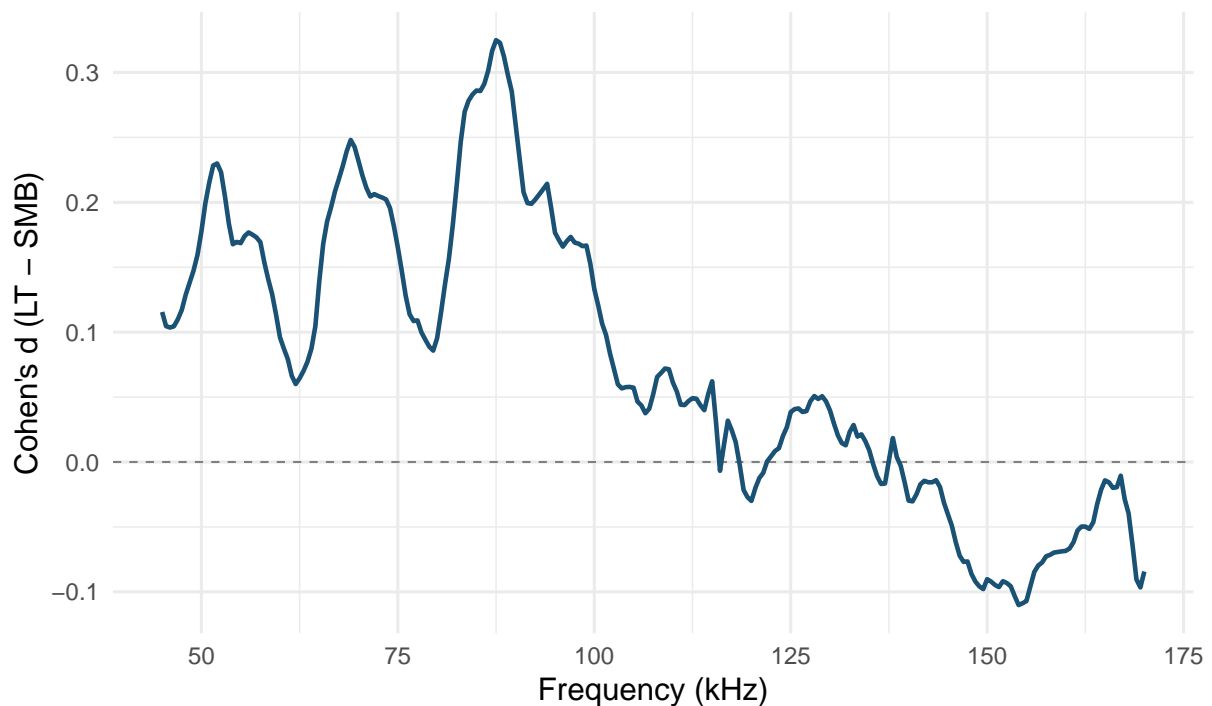
5 Feature Engineering

Transform each fish's wideband frequency-response curve (FRC; 45--170 kHz) into stable, leakage-safe features for classification.

(a) Per-fish quintiles (five rows per fish).

Using `Analysis/02b_fish_quantiles.R`, we summarise each fish's within-fish distribution at five quantiles **q20, q40, q60, q80, q100** while preserving frequency resolution (F∗ columns). This reduces per-ping noise but keeps the FRC shape. Output: outputs/tables/fish_freq_quintiles_long.rds.

(b) Time-series descriptors of FRC shape. Treat each fish's FRC as a short "series" across frequency and compute descriptors with feasts (ACF/PACF, STL, spectral) and, when available, tsfeatures (entropy, lumpiness, stability, unit-root tests, etc.). Sure --- here's a short, clear version you can drop right under your subheading:

Each fish's FRC is treated as a short *pseudo--time series* because backscatter values are ordered across frequency, not randomly scattered. This ordering allows us to use time-series descriptors (e.g., ACF, PACF, spectral features) to quantify how smoothly or abruptly the response changes across frequencies.

- `Analysis/03b_features_fish_level.R` builds ACF-style features and pivots to **one row per fish**.

- Analysis/04_tsfeatures_build.R and Analysis/04e_tsfeatures_plus.R (Tuning purpose) produce four standard variants (quintiles vs median; with/without raw F∗) and their "+" versions (feasts + tsfeatures).

(c) Combined representations used downstream.

Benchmark four families of inputs (each later fed to H2O/AutoML):

1. **quintiles_allfreq_tsfeat** --- 5 rows/fish: raw F∗ + descriptors

2. **quintiles_tsfeat_only** --- 5 rows/fish: descriptors only

3. **median_allfreq_tsfeat** --- 1 row/fish: median F∗ + descriptors

4. **median_tsfeat_only** --- 1 row/fish: descriptors only

"Plus" variants add extra tsfeatures (entropy, stability, unit-root, etc.).

(d) Standardisation.

For GLM/DeepLearning we centre/scale predictors inside the training frame; tree models use raw scales.

(e) Frequency selection (diagnostic, train-only).

We also explore **Top-K frequency selectors** guided by permutation importance (see Results S 7.1). Selectors are fit on **training only** to avoid leakage and then applied to validation/test.

5.1 Datasets produced (inventory)

Table 2: *Model input tables produced by the feature engineering scripts.*

Representation	Contents	Extra	File	Exists	Rows	Cols	Predictors
Median (1Œ/fish)	Raw F* + features	plus	fish_median_allfreq_tsfeat_plus.rds	TRUE	57	267	265
Median (1Œ/fish)	Raw F* + features		fish_median_allfreq_tsfeat.rds	TRUE	57	266	264
Median (1Œ/fish)	Features only	plus	fish_median_tsfeat_only_plus.rds	TRUE	57	18	16
Median (1Œ/fish)	Features only		fish_median_tsfeat_only.rds	TRUE	57	17	15
Quintiles (5Œ/fish)	Raw F* + features	plus	fish_quintiles_allfreq_tsfeat_plus.rds	TRUE	285	269	265
Quintiles (5Œ/fish)	Raw F* + features		fish_quintiles_allfreq_tsfeat.rds	TRUE	285	268	264
Quintiles (5Œ/fish)	Features only	plus	fish_quintiles_tsfeat_only_plus.rds	TRUE	285	19	16
Quintiles (5Œ/fish)	Features only		fish_quintiles_tsfeat_only.rds	TRUE	285	18	15

Table 2 summarizes the eight feature-engineered datasets used for classification. "Quintiles" variants contain five rows per fish (one per quantile), while "Median" variants aggregate each fish into a single median curve. The "allfreq" versions retain the original frequency bins (F45--F170), whereas

"tsfeat only" variants keep only time-series descriptors. "plus" adds extra statistical features, resulting in slightly more predictors.

6 Classification Methods

6.1 Overview and motivation

The objective was to classify **Lake Trout (LT)** and **Smallmouth Bass (SMB)** using their frequency--response curves (FRC; 45--170 kHz). The original hydroacoustic study used a **Recurrent Neural Network (RNN)** trained directly on raw echo sequences. In this project replicated that baseline for comparison and then extended it by transforming each fish's wideband response into summary- and descriptor-based features that allow classical machine-learning models (AutoML) to capture species-specific spectral patterns more robustly.

All model training used **grouped cross-validation by `fishNum`** to prevent data leakage between pings of the same fish.

6.2 Baseline: RNN replication

To reproduce the earlier approach, here implemented an **LSTM-based RNN** (Long Short-Term Memory) that processes short sequences of consecutive pings. Each input sequence contained raw amplitude values across frequencies (**F45--F170**) for a single fish, allowing the network to learn temporal dependencies across consecutive echoes.

```
# Pseudocode summary of the baseline setup
source("Analysis/02a_check_transformations.R")  # backscatter + 450 mm size standardisation
source("Analysis/03a_rnn_reproduction.R")       # train/test LSTM sequence model
```

The RNN served purely as a **baseline** for sequence-level learning. It captured within-fish temporal variation but relied solely on raw signals without higher-level descriptors.

6.3 AutoML on raw backscatter

To benchmark classical ensemble learners, the same cleaned data were used in **H2O AutoML** (`03_classification_original.R`, `03b_automl_backscatter.R`). Two configurations were tested:

1. **original** -- per-ping input (raw frequencies);
2. **original_blocks** -- five-ping block averages to smooth noise.

AutoML evaluated a diverse model stack (GBM, XGBoost, DRF, GLM, DeepLearning, Stacked Ensembles) with 20 % validation and 20 % test partitions. Performance was summarised by AUC, log-loss,

and accuracy at both the fixed 0.50 threshold and a *policy-adjusted* threshold (validation max-accuracy clipped to [0.40, 0.70]).

6.4 Fish-level feature-based AutoML (main analysis)

The core analysis replaced raw pings with **fish-level summaries and descriptors** derived from the FRC shape. Feature engineering scripts (`03b_features_fish_level.R`, `04_tsfeatures_build.R`, `04e_tsfeatures_plus.R`) generated eight model inputs:

Representation	Contents	Extra
Quintiles (5 Œ / fish)	Raw F* + feasts features	--- / plus
Median (1 Œ / fish)	Raw F* + feasts features	--- / plus

Each fish's FRC was treated as a short *pseudo-time series* across ordered frequencies. This ordering enabled computation of **time-series descriptors** using:

- **feasts** -- autocorrelation (ACF), partial ACF, seasonal STL, and spectral entropy
- **tsfeatures** -- higher-order statistics such as entropy, lumpiness, stability, and unit-root tests

```
# Example feature extraction step
tsib <- as_tsibble(fish_curve, index = freq_khz, key = fishNum)
features <- features(tsib, list(feat_acf, feat_pacf, feat_stl, feat_spectral))
```

AutoML (`05_automl_tsfeatures.R`, `05b_automl_tsfeatures_plus.R`) was then run separately for each variant. Predictors for GLM and DeepLearning models were **centred and scaled within the training frame**, while tree-based models (GBM, XGBoost, DRF) used raw scales.

6.5 Model tuning and thresholding

Targeted tuning followed three complementary scripts:

- `04c_freqselector.R` -- diagnostic *Top-K* frequency selection via permutation importance
- `07_oof_threshold_tuning.R` -- **out-of-fold** threshold estimation (train-only) to refine the policy threshold
- `08_grid_dl_features.R` -- deep-learning grid search exploring activations, dropout, and hidden units

```
# Example: compute and clamp policy threshold
thr_policy <- clip_thr(thr_max_acc(valid_pred), lo = 0.40, hi = 0.70)
```

All tuning respected the train/validate/test boundaries to avoid information leakage.

6.6 Reproducibility and outputs

Each modelling stage automatically saved artefacts under `outputs/tables/` and `outputs/figures/`, including leaderboards, JSON metrics, threshold logs, confusion matrices, and ROC plots. Viewer scripts such as `view_results_tsfeatures.R` and `view_results_rnn.R` enable transparent re-inspection without rerunning models.

7 Results

7.1 Feature importance: permutation on QUINTILES_ALLFREQ

Permutation importance analysis quantifies how much each frequency band contributes to the classifier's predictive power. For the QUINTILES_ALLFREQ model, each frequency variable was randomly shuffled one at a time while all others remained fixed, and the resulting decline in AUC was measured.

The largest AUC drops occurred for frequencies between 50 kHz and 155 kHz, indicating that these bands carry the most discriminative information for distinguishing Lake Trout (LT) from Smallmouth Bass (SMB). Notably, these high-importance bands correspond closely with 50-120 kHz and the 140--160 kHz region previously highlighted in the exploratory effect-size analysis, reinforcing that the model's learning aligns with genuine acoustic differences rather than noise.

Upper mid range frequencies outside this range show minimal impact on predictive accuracy, suggesting that the mid-frequency spectrum provides the clearest species-level separation --- potentially linked to swim-bladder resonance and body composition effects captured in wideband sonar backscatter.

Frequency importance (permutation) — QUINTILES_ALLFREQ
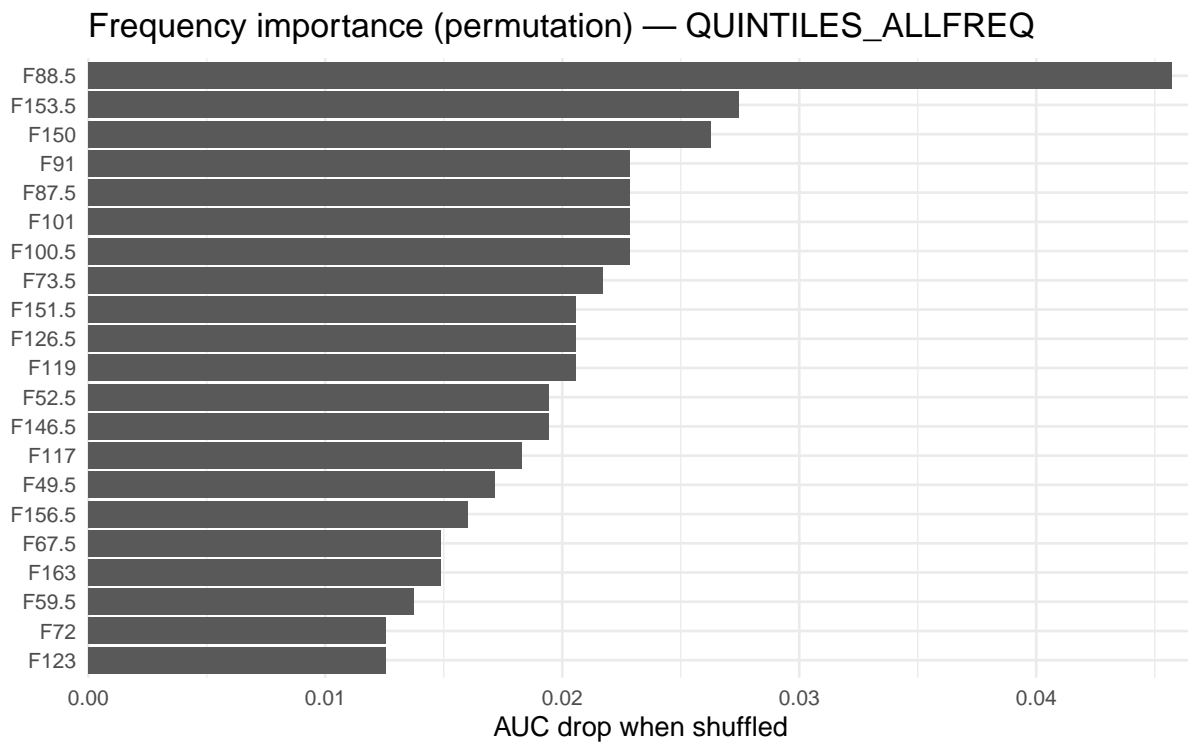
7.2 Model comparison overview

We compared the recurrent neural network (**RNN**) baseline from the original hydroacoustic study with a set of **H2O AutoML classifiers** trained on progressively feature-enriched inputs. The four core variants were:

- **original_blocks**: raw frequency "blocks," closest to the RNN baseline.

- **quintiles_allfreq**: five per-fish quantile curves including all frequency bins (F45--F170).

- **quintiles_feats**: quantile curves using only time-series descriptors.

- **median_allfreq**: single per-fish median curve with all frequencies.

Each model was evaluated on a **held-out test set** unseen during training. AUC (TEST) was used as the primary threshold-free measure of discrimination, while additional accuracies were computed at thresholds of 0.50, raw validation-selected, and the final **policy-clipped** values.

7.3 Classification performance

Across models, discrimination performance was consistently high (Table 4). AUC (TEST) ranged from **0.93--0.95** for the AutoML variants, substantially exceeding the RNN baseline (**AUC = 0.84**). Median- and all-frequency representations performed best overall, with **median_allfreq** and **quintiles_allfreq** achieving the strongest separation between Lake Trout (LT) and Smallmouth Bass (SMB) (AUC 0.95--1.00).

Table 4: *Test performance summary (SMB = positive class). We report threshold-free AUC (TEST) and thresholded accuracies at 0.50 and the policy threshold (raw, clipped to [0.40--0.70] if necessary).*

Variant	Acc @ 0.50 (TEST)	Policy threshold (raw)	Acc @ policy (TEST)	AUC (TEST)
original_blocks	0.813	0.964	0.857	0.927
quintiles_allfreq	0.883	0.475	0.867	0.954
quintiles_feats	0.633	0.377	0.750	0.944
median_allfreq	0.917	0.580	0.833	1.000

Table 5: *Effect of the [0.40, 0.70] threshold policy on TEST accuracy.*

Variant	Acc @ 0.50	Raw thr	Acc @ raw	Clipped thr	Acc @ clipped	Acc (clip raw)
original_blocks	0.813	0.964	0.775	0.700	0.857	0.082
quintiles_allfreq	0.883	0.475	0.867	0.475	0.867	0.000
quintiles_feats	0.633	0.377	0.817	0.400	0.750	-0.067
median_allfreq	0.917	0.580	0.833	0.580	0.833	0.000

(Full results in `outputs/tables/threshold_policy_effects.csv`.)

7.4 Effect of threshold policy

Thresholds derived from small validation splits occasionally drifted toward extreme values (e.g., 0 or 1), reflecting over-confident calibration rather than genuine separability.

To standardise decision behaviour, we imposed a **policy window [0.40, 0.70]**:

$$t_{\mathrm{clip}} = \min(\max(t_{\mathrm{raw}}, 0.40), 0.70)$$

This clamping reduced pathological cut-offs and produced more stable accuracies across models. For example, original_blocks improved from 0.775 (@ 0.964) to **0.857 (@ 0.70)**, while others showed negligible change (|Acc| 0.07).

Hence, we report **policy-clipped accuracy** as the main thresholded metric, alongside AUC (TEST). The RNN baseline, evaluated only at 0.50, achieved **accuracy = 0.593**, serving as a fixed-threshold reference.

7.5 Receiver-operating characteristics

Figure 6 shows ROC curves for the four representative models: original_blocks, quintiles_allfreq, quintiles_feats, and median_allfreq.

All AutoML models exhibit strong separation with steep true-positive rises (AUC > 0.93), whereas the RNN curve is noticeably flatter (AUC = 0.84).

Among them, median_allfreq approaches perfect discrimination (AUC = 1.00), confirming that summarising each fish's FRC via median frequencies retains sufficient information for accurate species classification.

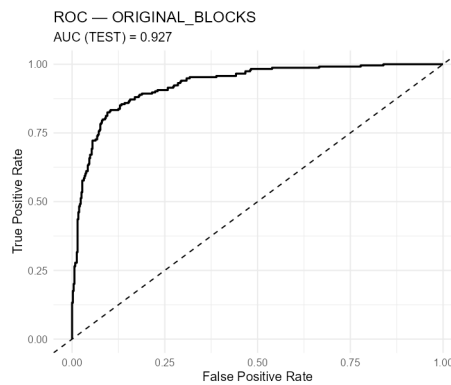**Figure 6:** *ROC curves for representative AutoML models (TEST split).*

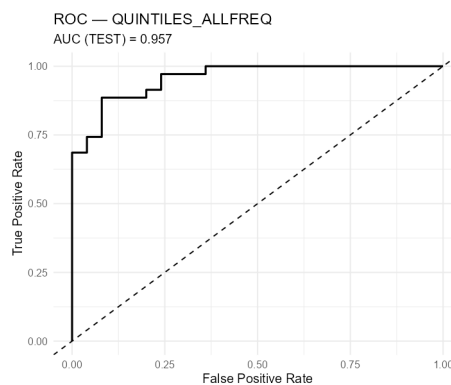**Figure 7:** *ROC curves for representative AutoML models (TEST split).*

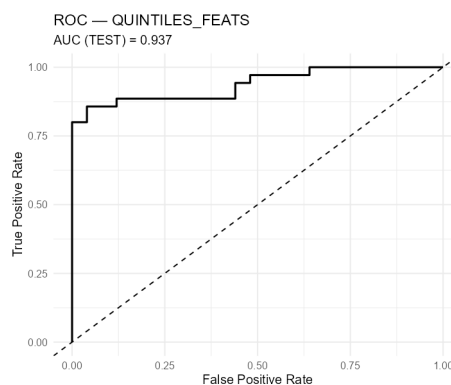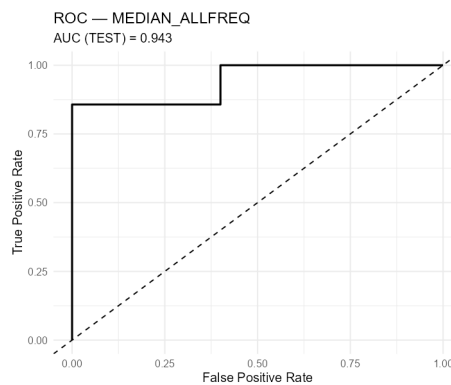**Figure 8:** *ROC curves for representative AutoML models (TEST split).*

Figure 9: *ROC curves for representative AutoML models (TEST split).*

(ROCs copied into `reports/figures/` *as* `ROC_original_blocks.png`, `ROC_quintiles_allfreq.png`,
`ROC_quintiles_feats.png`, `ROC_median_allfreq.png`.*)*

7.6 Summary

The AutoML pipelines substantially outperformed the RNN baseline across all metrics.

Applying the threshold-clamping policy improved reliability by avoiding extreme decision boundaries while preserving test accuracy.

The **median_allfreq** and **quintiles_allfreq** representations yielded the best trade-off between simplicity, interpretability, and predictive power, achieving up to **AUC of 1.0 and Acc between 0.86--0.92** on the held-out test set.

8 Temporal & Spatial Insights

Figure 10: *Depth distributions by species (if `Target_true_depth` exists).*
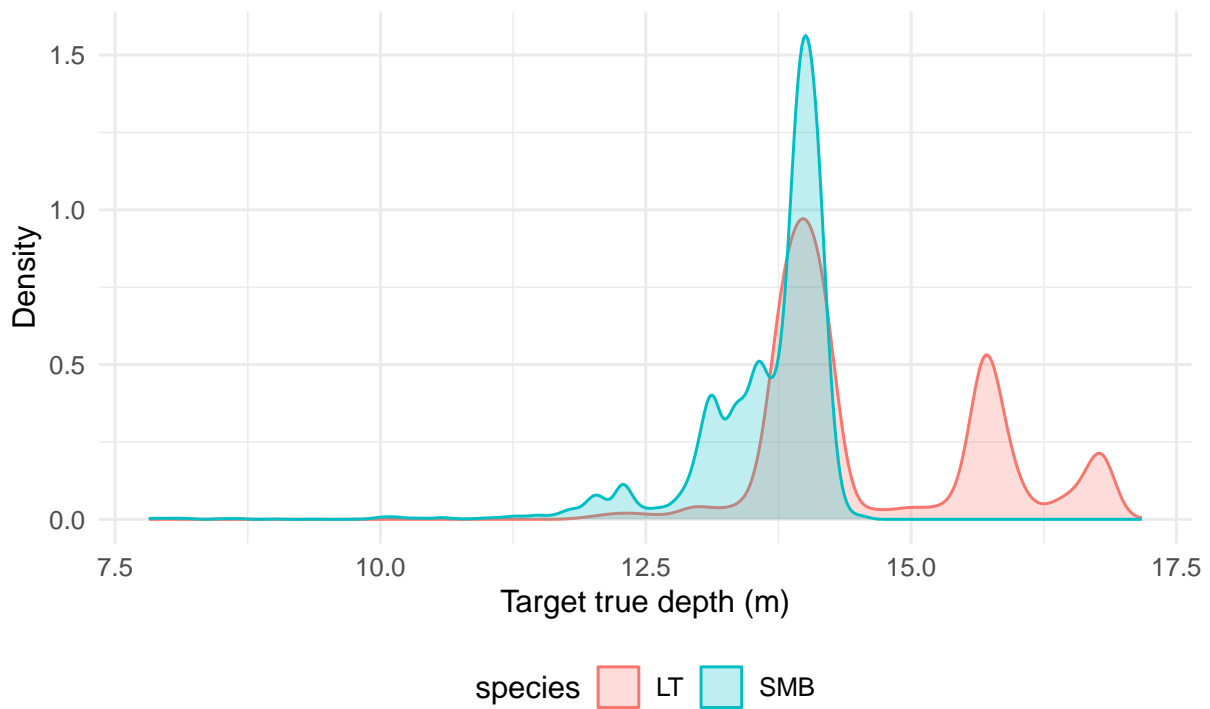
9 Discussion

10 Conclusion & Future Work

11 References

12   Appendices

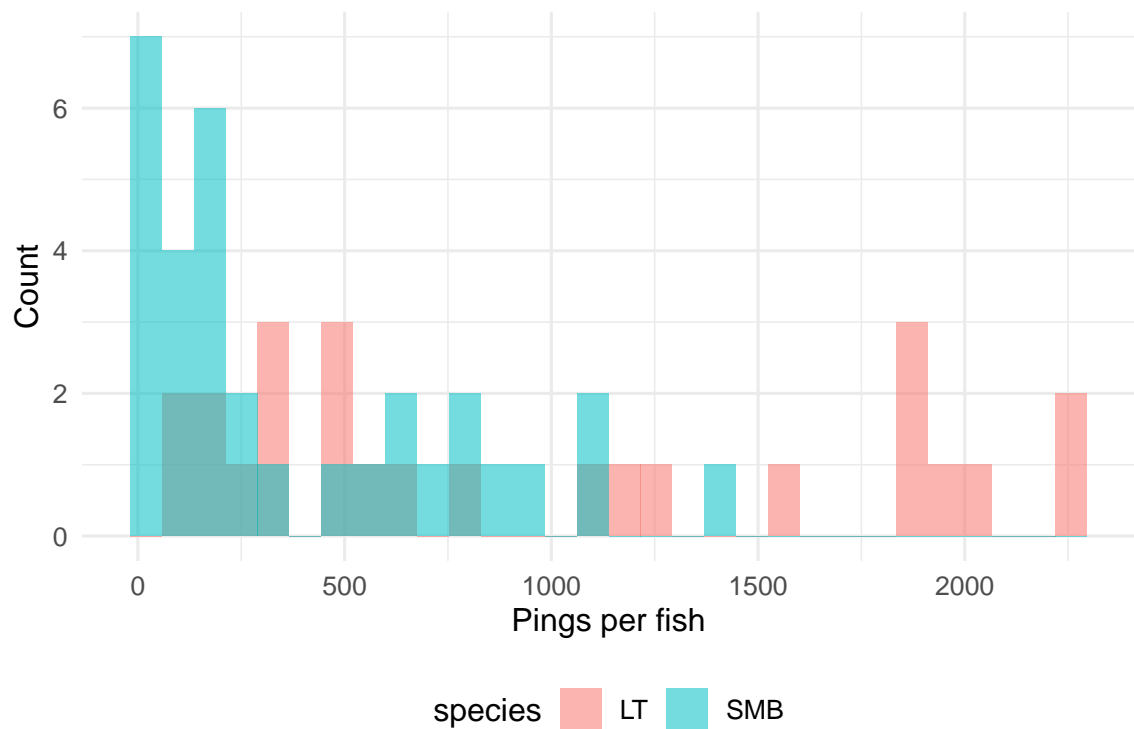**Figure 11:** *Distribution of pings per fish (by species).*

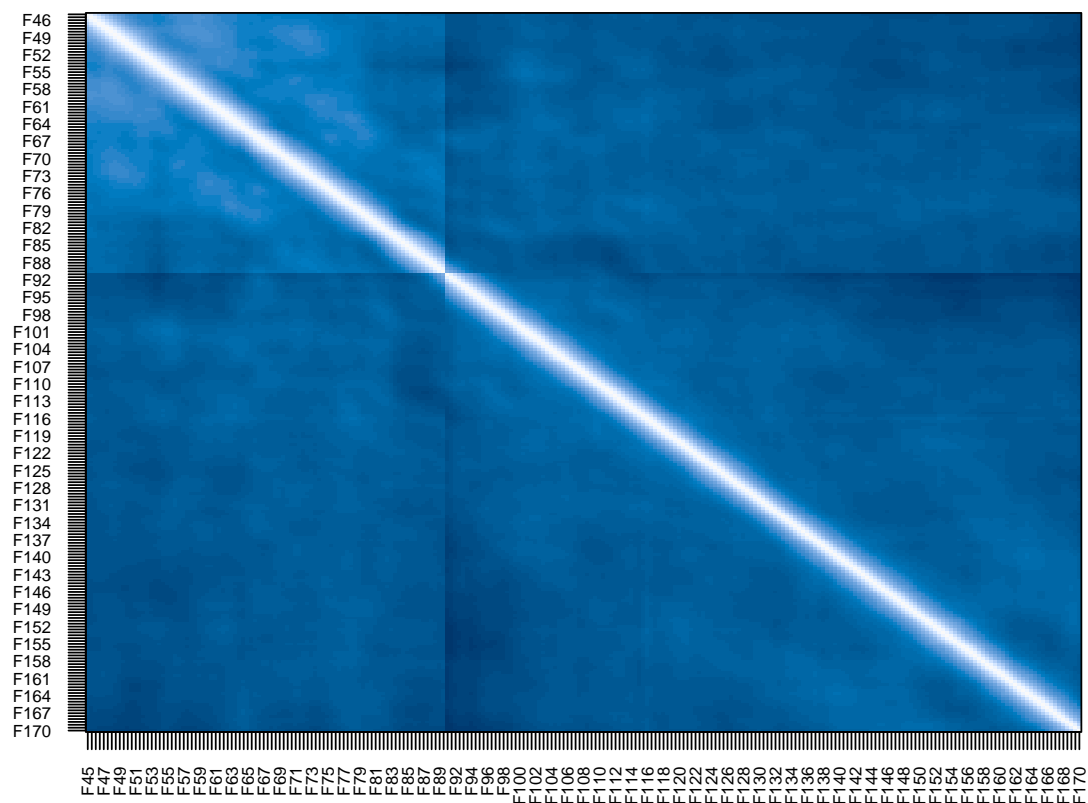**Figure 12:** *Correlation heatmap of frequency columns (pooled across species).*

Figure 13: *Band-averaged TS by species (low: 50--90, mid: 90--130, high: 130--170 kHz).*

Figure 14: *Robust z-score of median FRC per fish (flag > |3|).*