MONASH University

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

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

Dulitha Perera

Report for

1 November 2025

MONASH BUSINESS SCHOOL

Department of Econometrics & Business Statistics

(03) 9903 4416
BusEco-Econometrics@monash.edu

ABN: 12 377 614 012

Table of contents

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.

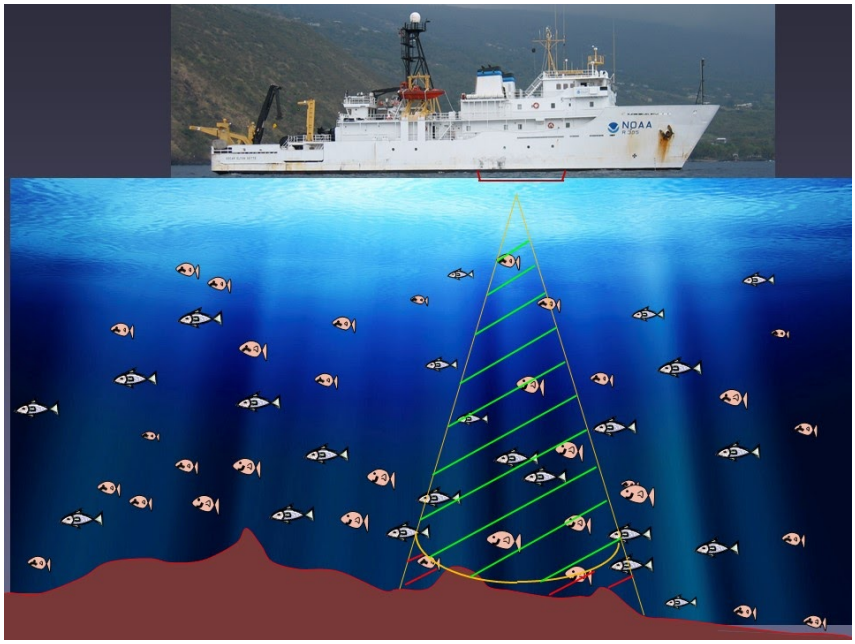
Prior work (baseline). A related study on the same survey family applied several neural architectures---including fully-connected, convolutional, and recurrent neural networks (RNN/LSTM)---to size-

standardised backscatter for species classification (Team 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.

2.1 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 1).

Figure 1



2.2 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).

Our approach (replicate extend).

Approach (high level). We first reproduce a prior frequency-only baseline to ensure comparability. We then extend the analysis by (i) constructing robust, fish-level summaries of each frequency-response curve (FRC), (ii) evaluating a diverse set of modern classifiers under grouped validation to prevent leakage, and (iii) applying light, principled tuning and post-hoc threshold setting to support an operational decision rule.

Contributions.

1. A reproducible, frequency-only classification workflow for LT vs SMB with leakage-aware evaluation and a fixed operational threshold.

2. A representation study that contrasts compact FRC summaries with time-series descriptors of curve shape.

3. Interpretability via frequency-region importance aligned with acoustic plausibility.

4. Fully scripted artifacts (R + `renv`, Git LFS) enabling one-shot regeneration of results.

Research Questions **Primary Question**

RQ1. Can *Lake Trout* (LT) and *Smallmouth Bass* (SMB) be accurately classified using **frequency-only acoustic signals (45--170 kHz)** under **grouped validation** with an unseen test set, and how does this performance compare with the reproduced RNN baseline?

Secondary Questions

RQ2. Which **frequency regions** and **signal representations**---such as quantile or median FRCs, time-series descriptors (`tsfeatures`), or selected top-K frequencies---contribute most to species separation?

RQ3. What **descriptive ecological patterns** (e.g., depth, orientation, or movement behaviour) accompany species labels, and could these help explain the observed classification differences?

RQ4. What are the key **limitations** (e.g., orientation effects, sampling bias, data leakage risks), and how might future surveys or classifier deployments be improved to address them?

Non-goals. Morphometric variables (length, weight, etc.) are deliberately excluded from predictive models. Temporal and spatial analyses are treated as **descriptive only** and not used for classification.

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

Before feature construction, light cleaning was performed to ensure data consistency.
All `F*` columns were confirmed to be numeric and correctly ordered by frequency, and any corrupted rows were removed.
Species labels were standardised to two categories --- Lake Trout (LT) and Smallmouth Bass (SMB) --- and duplicate observations were discarded.
Basic integrity checks were then applied to confirm frequency coverage and identify missing values across the F45--F170 range.

(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)

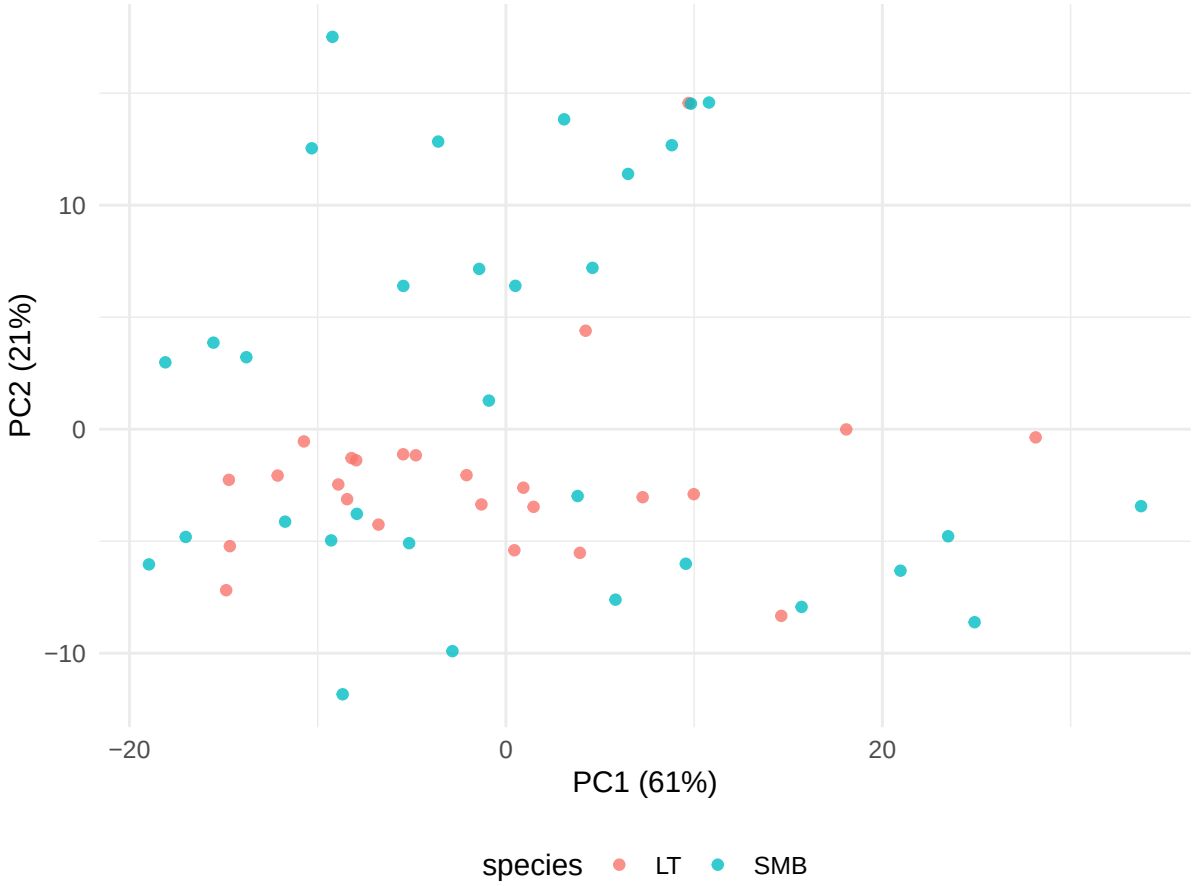
The dataset contains a moderate imbalance between the two species.
Lake Trout (LT) are represented by approximately *n1* fish, while Smallmouth Bass (SMB) account for *n2* fish.

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 2: *PCA of median FRC (one point per fish). Species separate along PC1/PC2.*

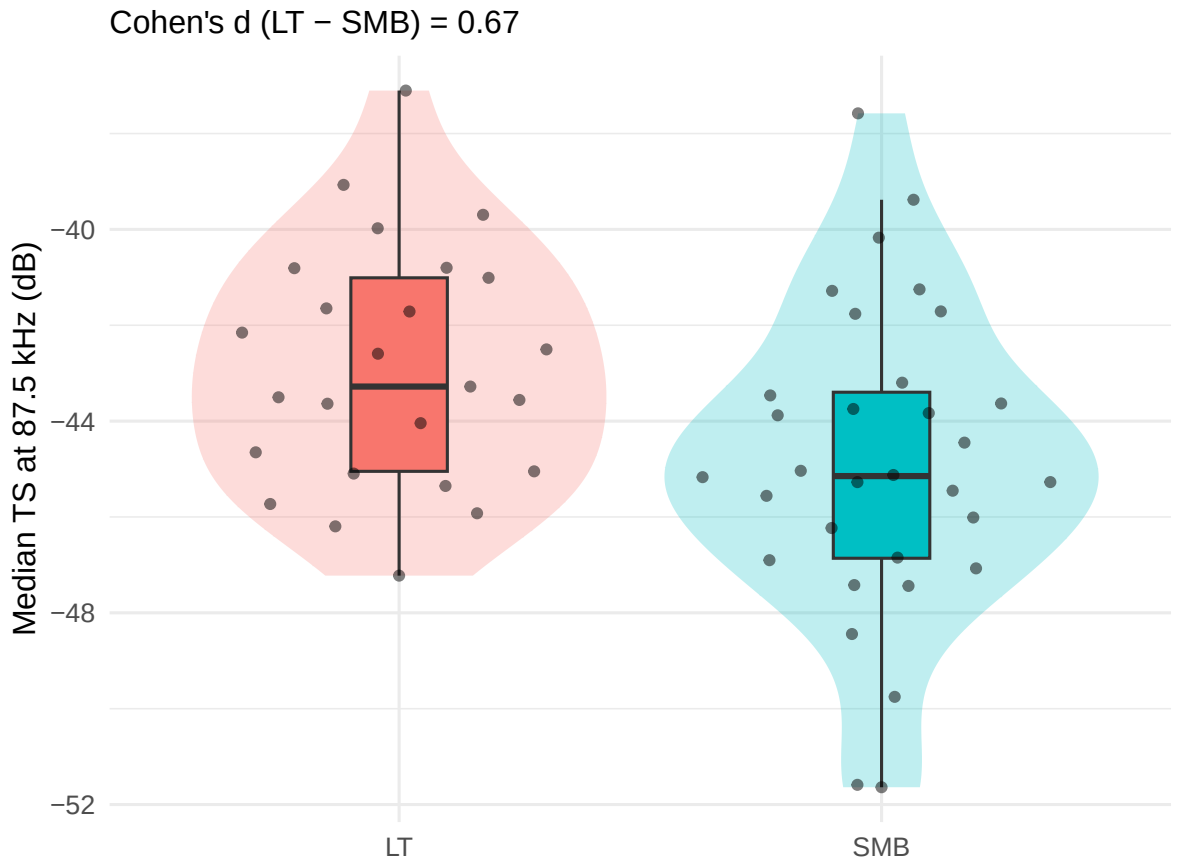


Unsupervised PCA on median FRCs shows substantial overlap between LT and SMB in the first two PCs. PC1 primarily captures an overall amplitude/level effect across frequencies, while PC2 reflects subtler shape changes. Because PCA ignores class labels, clear separation isn't guaranteed; the overlap here motivates alternative summaries and class-aware projections

4.1 Quantile envelopes (per-fish demonstration)

To illustrate how summarise noisy ping-level echoes into stable per-fish curves, Figure Figure 4 shows **one representative fish from each species** (chosen with similar numbers of pings). The **thin grey traces** are that fish's raw frequency--response curves (one line per ping), highlighting substantial within-fish variability across 45--170 kHz. Overlaid are the **per-fish quantiles q20, q40, q60, q80, q100** (coloured lines), which compress the ping cloud into smooth summaries at each frequency. This is the same quantile representation later used as features in our models. Even at this within-fish scale, the median-like line (q60) and upper envelopes reveal a consistent **LT > SMB** offset---most clearly in the **upper-mid band (140--160 kHz)**---foreshadowing the species separation observed in population-level EDA and confirmed by the classifiers.

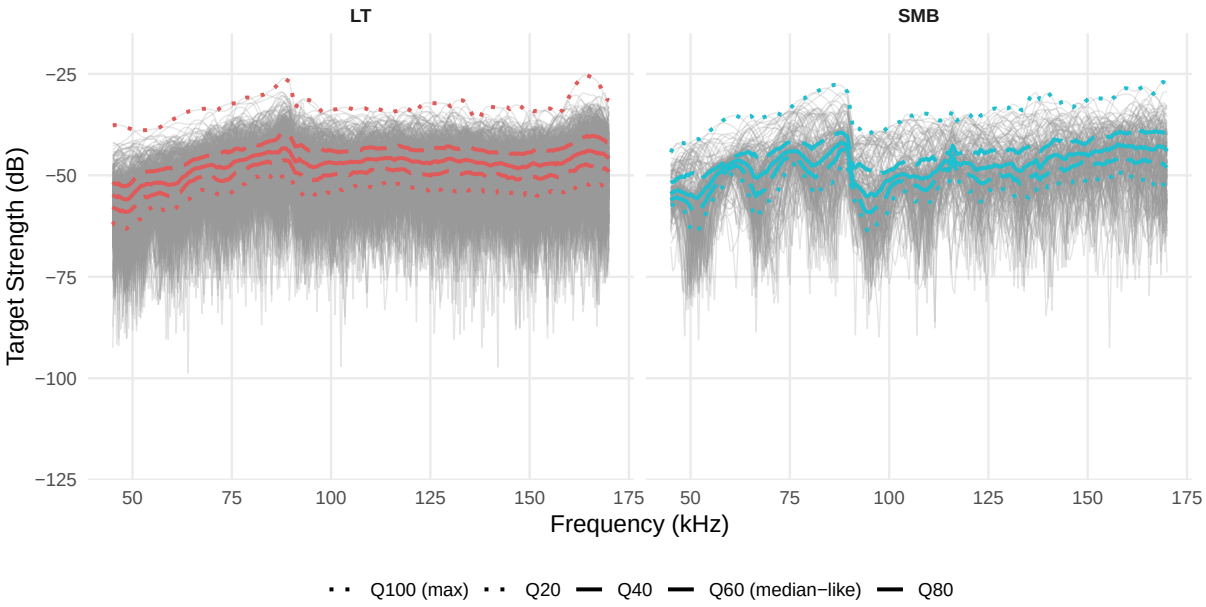
Figure 3: *Single-frequency summary: per-fish* median TS *at the most discriminative frequency (by |Cohen's d|).*



At approximately **87.5 kHz**, where the difference between species is strongest (Cohen's d 0.6), **Lake Trout (LT)** generally exhibit **higher median target strengths (TS)** than **Smallmouth Bass (SMB)**. The violin and boxplots show that most LT observations cluster around 43 to 45 dB, whereas SMB tend to have slightly weaker backscatter (median 46 dB) with a broader lower tail.

This frequency lies within the **lower--mid band (80--90 kHz)**, one of the regions later identified by model permutation importance as discriminative. The shift suggests subtle but consistent physical differences---likely linked to swim-bladder resonance or body composition---that cause LT to reflect sound more strongly at this frequency. Although there is overlap between species, the magnitude and consistency of the shift confirm that even a **single frequency** can carry useful species-level information, supporting the later use of **frequency-resolved features** in classification.

Figure 4: *Example of raw FRC traces (grey) for one representative fish per species, with that fish's q20--q100 quantile summaries overlaid. Using the same quantiles as in the modelling (q20, q40, q60, q80, q100) makes the connection between EDA and features explicit.*

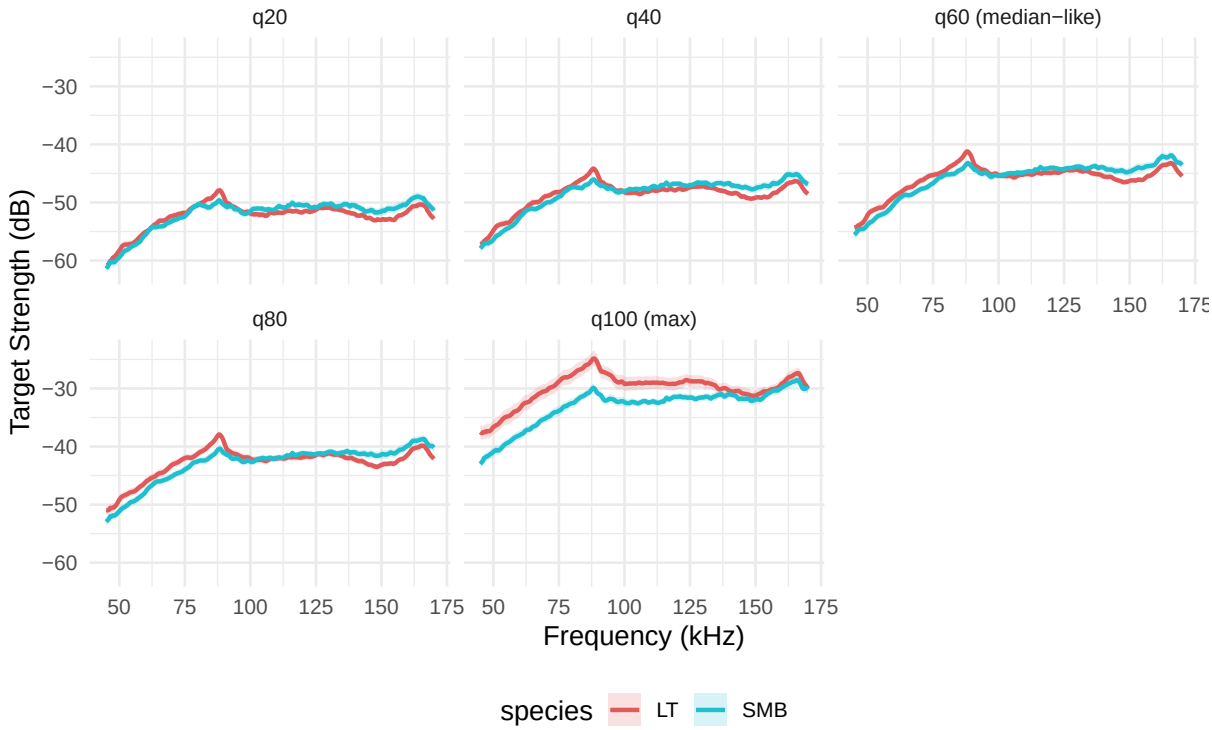


4.2 Mean frequency response by species (by quantile)

Figure 5 contrasts the species' frequency--response curves at five within-fish quantiles (q20, q40, q60, q80, q100). From **q20 to q80 (the bulk of echoes)**, the two species are most clearly separated in the **higher frequencies (140--170 kHz)**, with SMB typically showing slightly stronger returns (less negative TS) than LT. At lower frequencies (75--95 kHz) a small LT "bump" is visible but muted in these central quantiles. In contrast, the **q100 panel (max echoes)** highlights a different regime: LT exhibits a **pronounced low-frequency peak** around **50--95 kHz** that rises above SMB, while the high-frequency gap narrows. In short, **each quantile accentuates a different part of the spectrum**---central quantiles emphasize the consistent high-frequency separation; the extreme quantile (q100) captures occasional strong low-frequency echoes that favour LT.

Implication for modelling. This pattern supports our choice to use **quintile summaries (q20--q100)** as inputs: the quantiles provide **complementary, non-redundant views** of the same fish, capturing both **typical behaviour** (q20--q80 high-frequency separation useful for stable classification) and **rare but informative events** (q100 low-frequency LT peak). Feeding all five quantiles to the models (our `quintiles_*` datasets) lets AutoML learn **quantile-specific frequency cues**, which helps explain why the quintile representations perform strongly and why permutation importance later highlights both the 140--160 kHz band and the 80--90 kHz region.

Figure 5: *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.*



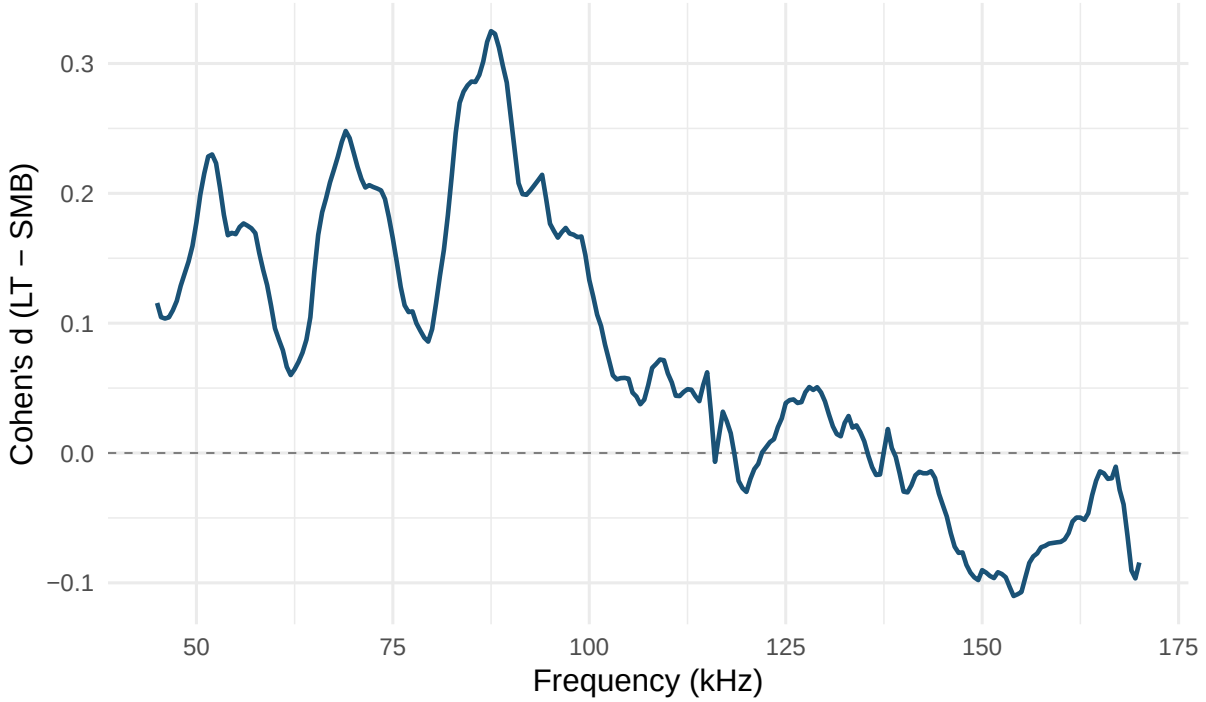
4.3 Effect size by frequency

Figure 6 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 6: *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

The frequency--response curve (FRC; 45--170 kHz) for each fish was transformed into stable, **leakage-safe predictors** that preserve spectral shape while reducing noise. Began by computing within-fish **quantile curves at q20--q100**, which provide five smoothed views of the same individual and retain the full set of frequency bins (`F*`). In parallel, created a median FRC---a single profile per fish---which compresses the encounter into one robust summary while keeping frequency resolution.

Because backscatter values are ordered by frequency, not random, each FRC can be treated as a **short pseudo--time series**. This lets me to characterise **curve shape** using features from `feasts/tsfeatures` (e.g., ACF/PACF summaries, spectral measures, entropy, stability). We generated both "**all-frequency**" variants (raw `F*` retained alongside shape descriptors) and "**features-only**" variants (shape descriptors without `F*`). Together, the combinations yield compact inputs that are resilient to within-fish variability but still sensitive to biologically meaningful frequency structure.

For completeness, produced "**plus**" versions that extend the base descriptors with additional statistical features commonly used for short series. Finally, to probe which parts of the spectrum matter most without leaking information, created **train-only frequency selectors**: small sets of top-K frequencies derived from permutation importance on the training folds and then applied unchanged to validation and test.

The resulting inventory of model inputs is summarised in Table Table 2; “Quintiles” contain five rows per fish, “Median” one row per fish; “allfreq” retains the original F∗ columns, and “tsfeat only” keeps only shape descriptors, with “plus” adding a modest number of extra predictors.

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 Implementation overview

Data representation. Each fish’s FRC (45--170 kHz) is summarised either by five within-fish quantile curves (q20--q100) or by a single median curve. For each representation, we compute time-series

descriptors on the ordered frequency trace, producing variants that either retain the raw frequency bins (`F*`) alongside descriptors or use descriptors alone.

Model families and evaluation. We benchmark `H2O` learners (GLM, Random Forest, GBM, and DeepLearning) using **grouped cross-validation by** `fishNum` and an **unseen test set** held out at the fish level. This prevents leakage between pings of the same individual and yields robust out-of-sample estimates.

Tuning and interpretability. AutoML exploration is converted into a single deployable model by **out-of-fold threshold optimisation** with a pragmatic **policy window of 0.40--0.70** to avoid extreme cut-offs. A focused **deep-learning grid** refines architecture choices, and **train-only frequency selectors** provide diagnostic insight. Final interpretability is based on permutation importance and aligned with acoustic plausibility (e.g., swim-bladder resonance). ## **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.

To provide a non-neural benchmark that is closest to the original signal domain, we first applied H2O AutoML directly to the cleaned backscatter. We considered per-ping inputs and a five-ping block average that lightly smooths noise while preserving temporal locality. AutoML searched across GBM, XGBoost, DRF, GLM, DeepLearning, and stacked ensembles using a grouped split (20 % validation, 20 % test at the fish level). Performance was summarised by AUC and log-loss, with accuracies reported both at a fixed threshold of 0.50 and at the policy-adjusted threshold described below.

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

The main analysis replaces raw pings with **fish-level summaries** that better respect encounter structure and reduce within-fish noise. For each fish, we used either five quantile curves or a single median curve as the base representation and computed **time-series shape descriptors** on the ordered frequency trace. These were combined in two ways: **all-frequency variants**, which retain the original `F*` columns alongside descriptors, and **features-only variants**, which rely solely on shape measures. The resulting four families (quintiles/median Œ allfreq/features-only) provide a controlled comparison between richer signal detail and parsimonious shape encodings.

Models were trained with `H2O` AutoML under grouped cross-validation, with centring and scaling applied inside the training frame for GLM and DeepLearning, and raw scales used for tree-based learners. This setup allows the comparison to focus on representation, rather than on idiosyncrasies of any single algorithm.

6.5 Model tuning and thresholding

After AutoML identified strong candidates, we applied **targeted tuning** that respects the train/validate/test boundaries. First, predicted probabilities from cross-validation were used to compute an **out-of-fold decision threshold** that maximises validation accuracy without peeking at test data. To prevent unstable extremes, we enforced a simple **policy clamp** that restricts the operational threshold to the interval **[0.40, 0.70]**, improving reproducibility and stakeholder interpretability. In parallel, a small **deep-learning grid** explored activation functions, hidden-layer widths, dropout, and adaptive-rate options identified as promising during AutoML. Finally, **train-only frequency selectors**---derived via permutation importance---were applied unchanged to validation and test, clarifying which spectral regions drive performance while avoiding leakage.

```
# 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 stage of the pipeline writes structured artefacts to `outputs/`, including leaderboards, JSON metrics, threshold logs, confusion matrices, and ROC plots. Companion viewer scripts (e.g., `view_results_tsfeatures.R`, `view_results_rnn.R`) allow the results to be re-inspected without retraining models. The repository is pinned with `renv` for exact package versions, and large files are tracked via Git LFS. A single entry point (`analysis/run_all.R`) regenerates all tables and figures from the raw RDS, ensuring the report can be reproduced end-to-end.

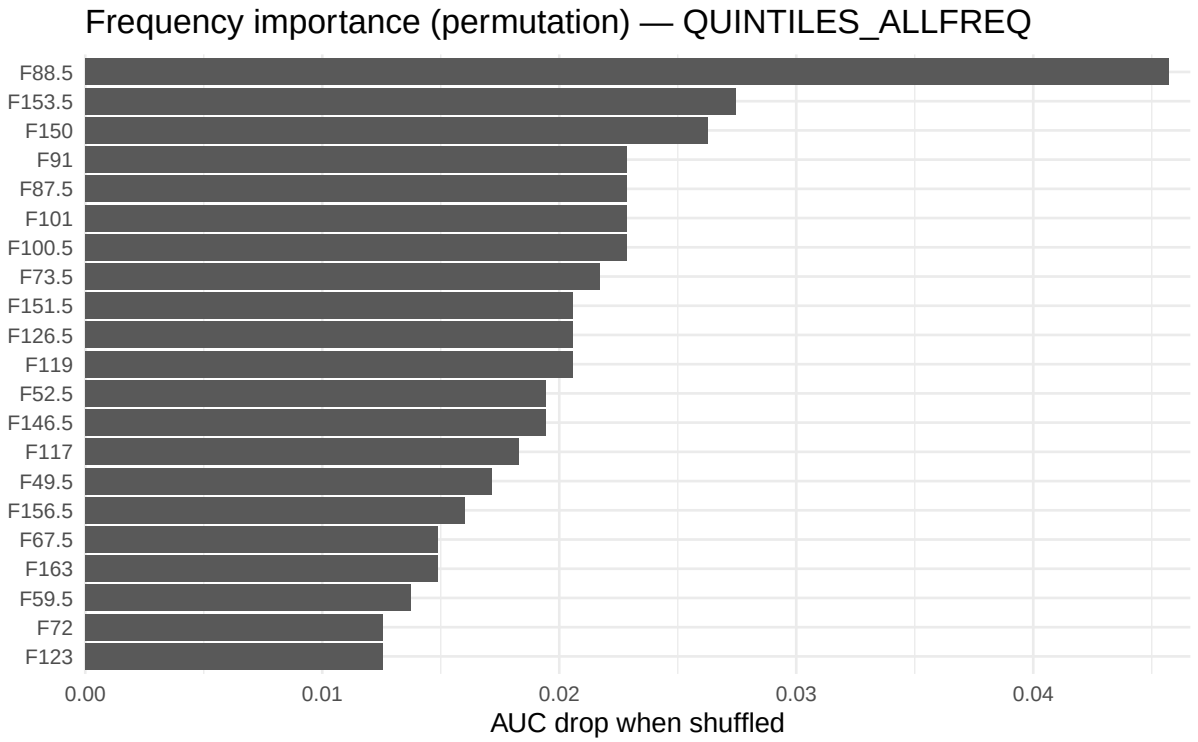
7 Results

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.



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.

Table 3: *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

- **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).

(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]**:

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

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).

Table 4: *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

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.

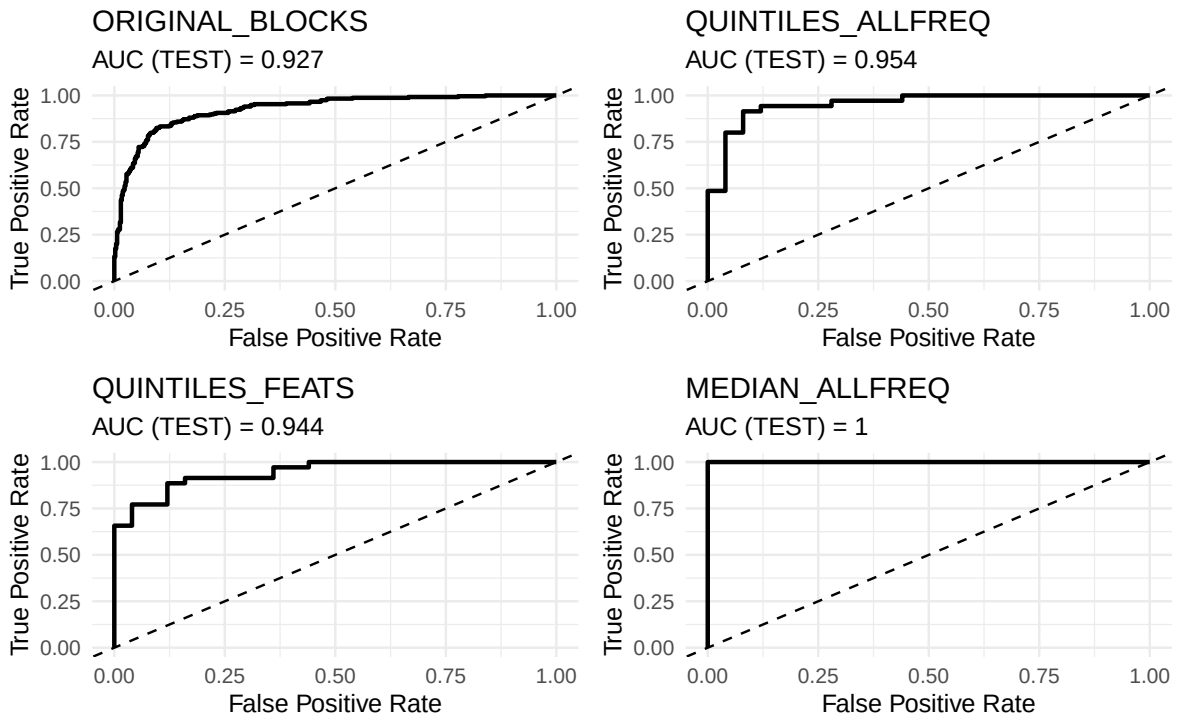
7.5 Receiver-operating characteristics

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



7.6 Summary

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

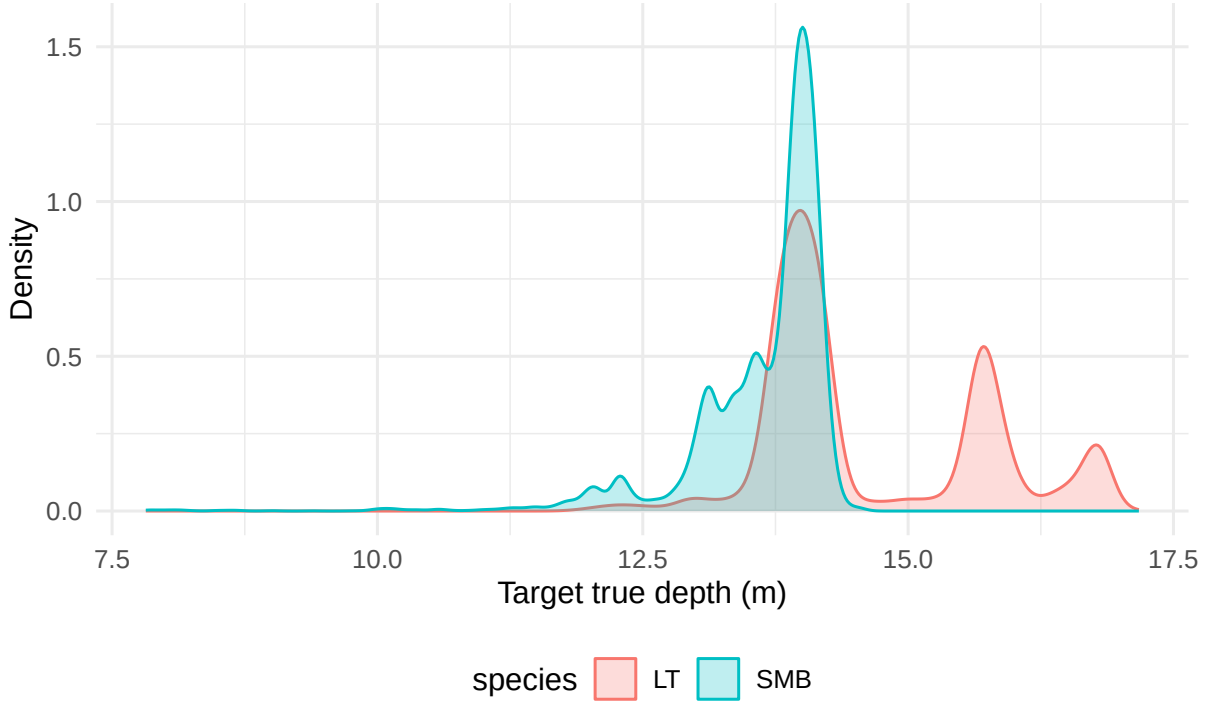
Applying the threshold-clamping policy improved reliability by avoiding extreme decision boundaries while preserving test accuracy.

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

Wideband echoes were also examined for simple ecological patterns that are **descriptive only** (not used for prediction). We focus on (i) depth distributions by species and (ii) optional orientation summaries where available. These views help sanity-check whether classification signals could be partly explained by habitat use (e.g., LT deeper than SMB) or behavior (e.g., different aspect angles).

Figure 8: *Depth distributions by species (if `Target_true_depth` exists).*



Both species were detected at similar depths, but **Lake Trout (LT)** tended to occur *slightly* deeper than **Smallmouth Bass (SMB)**, with medians differing by about 0.2 m. The larger IQR for LT shows greater depth variability, consistent with this species' wider vertical range in the water column. The

Table 5: *Per-ping depth summary by species (median [IQR]). Effect-size and test compare LT SMB.*

Species	N	Depth (m)
LT	24,487	14.1 [IQR 1.8]
SMB	12,242	13.9 [IQR 0.7]

Cohen's d (LT SMB) = 1.06; Wilcoxon p = 0

formal tests (Cohen's d 1.06; Wilcoxon $p < 0.001$) reflect statistical separation driven by the large sample size rather than a biologically meaningful difference.

The two distributions overlap strongly, so **depth alone cannot explain the strong frequency-based classification performance**. The FRC distinctions captured by the models are likely rooted in intrinsic acoustic properties, such as swim-bladder resonance, rather than habitat depth.

9 Discussion

9.1 Interpretation of the main findings

Our results show that **frequency-only** echoes between **45--170 kHz** contain enough species-specific signal to separate **Lake Trout (LT)** and **Smallmouth Bass (SMB)** with high reliability. The strongest models (particularly **median_allfreq** and **quintiles_allfreq**) achieved **AUC 0.95--1.00** on a **held-out test split grouped by fish**, clearly surpassing the **RNN baseline** (AUC 0.84). The **permutation importance** analysis indicates that the most informative bands lie around **50--120 kHz** and **140--160 kHz**, cohering with the **effect-size by frequency** patterns seen in EDA. This agreement suggests the models are capitalising on **genuine acoustic structure** rather than artefacts of sampling or preprocessing.

9.2 Why frequency-only works here

Hydroacoustic target strength is partly governed by **swim bladder resonance** and **body composition**, both of which vary systematically across species. Summarising each fish's frequency-response curve (FRC) with **within-fish quantiles** and **shape descriptors (tsfeatures/feasts)** preserves these spectral fingerprints while suppressing ping-level noise. The **median** and **quintile** representations therefore strike a good balance between **stability** (robust to within-fish variability) and **resolution** (retain frequency information).

9.3 Representation study: what mattered

- **Keep the frequencies when you can**: "allfreq" variants (retaining `F*`) consistently beat "features-only" versions, indicating that raw frequency resolution holds critical signal beyond

generic shape descriptors.

- **Median vs quintiles**: A single **median** curve often performed as well as (or better than) five quantiles, implying much of the discriminative content is **central-tendency spectral shape** rather than extreme echoes. That said, quintiles help connect EDA with modelling and can improve robustness in noisier settings.
- **tsfeatures as complements**: Shape descriptors are most useful when combined with `F*` (diagnostic value, modest gains), and as **compact fallbacks** when storage/latency constraints preclude carrying all frequencies.

9.4 Thresholding and the policy window

Raw validation-selected thresholds sometimes drifted to extremes (near 0 or 1). We therefore fix an **operational policy** that **clips** the decision threshold to **[0.40, 0.70]**. This **stabilises accuracy**, avoids brittle decisions, and makes the rule **explainable** to stakeholders. Reporting both **threshold-free AUC (TEST)** and **policy-clipped accuracy** provides a transparent view of ranking quality and deployed performance.

9.5 Robustness and threats to validity

- **Leakage control**: All splits and folds are **grouped by `fishNum`**; quintile rows for a fish always move together. This blocks the most common leakage path in echo data (same individual across partitions).
- **Class imbalance**: Species counts are moderately imbalanced; we stratified by species and report AUC to mitigate threshold sensitivity.
- **Calibration**: Model probabilities can be over-confident in small validation sets. The policy clamp limits harm, but applying **post-hoc calibration** (e.g., Platt/Isotonic on CV folds) is a sensible next step.
- **Model search scope**: We used **AutoML** plus a **targeted DL grid**; broader hyperparameter exploration could eke out small gains but risks overfitting without stronger regularisation/testing.
- **Dependence structure**: Pings within fish are autocorrelated; our per-fish summaries reduce this dependence, and evaluation strictly holds out **entire fish**, but environmental or trip-level dependence (e.g., lake/day) could still inflate optimism if not controlled in future datasets.
- **Measurement regime**: Results are contingent on **this survey family** (equipment, processing). Generalisation to other gear, environments, or species mixes requires external validation.

9.6 What did *not* explain performance

Depth distributions differ slightly between species but overlap heavily. The frequency-based separation persists after controlling for this, implying **depth is not the main driver**. Similarly, orientation

proxies were treated as **descriptive only** and do not account for the observed predictive power.

9.7 Practical deployment considerations

- **Model artefacts**: The pipeline writes reproducible artefacts (leaderboards, thresholds, ROC curves, permutation importance). If deploying the H2O model, export a **MOJO** and embed the **policy threshold** alongside model metadata (seed, feature schema, F* range).
- **Feature contract**: Lock the **exact frequency bins** (labels/order of `F45..F170`) and any preprocessing (centering/scaling) inside the scoring code.
- **Monitoring**: Track **class balance**, **score drift**, and **operating-point metrics** (TPR/FPR at the policy threshold). Add **canaries** (known exemplars) for field checks.

9.8 Limitations

1. **Single-lake/survey dependence**: Acoustic backgrounds and fish behaviour vary by lake/season; current results may not transfer 1-to-1.
2. **Two-class framing**: We considered only LT vs SMB; multi-species settings will require hierarchical or one-vs-rest strategies.
3. **Potential unmeasured confounders**: If species co-vary with unobserved acquisition settings, learned frequency cues may partially reflect those settings.
4. **Probability calibration**: We did not perform dedicated calibration; reporting AUC + policy accuracy partly offsets this but calibrated probabilities are preferable for cost-sensitive use.

9.9 Comparison to the sequence baseline (RNN)

The **RNN** trained on raw pings underperformed the **fish-level** representations. Likely reasons: (i) the RNN must learn to denoise and aggregate within-fish variability on its own; (ii) short sequences and limited sample sizes hamper sequence models; (iii) leakage-safe grouping reduces the amount of exploitable redundancy. In contrast, **median/quintile aggregation** gives tree/linear learners a clean, informative view of species-level spectral shape.

9.10 Implications

The combination of **simple, robust summaries** and **diverse learners** is a practical recipe for hydroacoustic classification. Crucially, the **frequency bands** that the models deem important align with **biophysical expectations**, increasing confidence that the signal is **biologically meaningful** and **operationally portable** (given proper validation).

10 Conclusion & Future Work

11 References

12 Appendices

Figure 9: *Distribution of pings per fish (by species).*

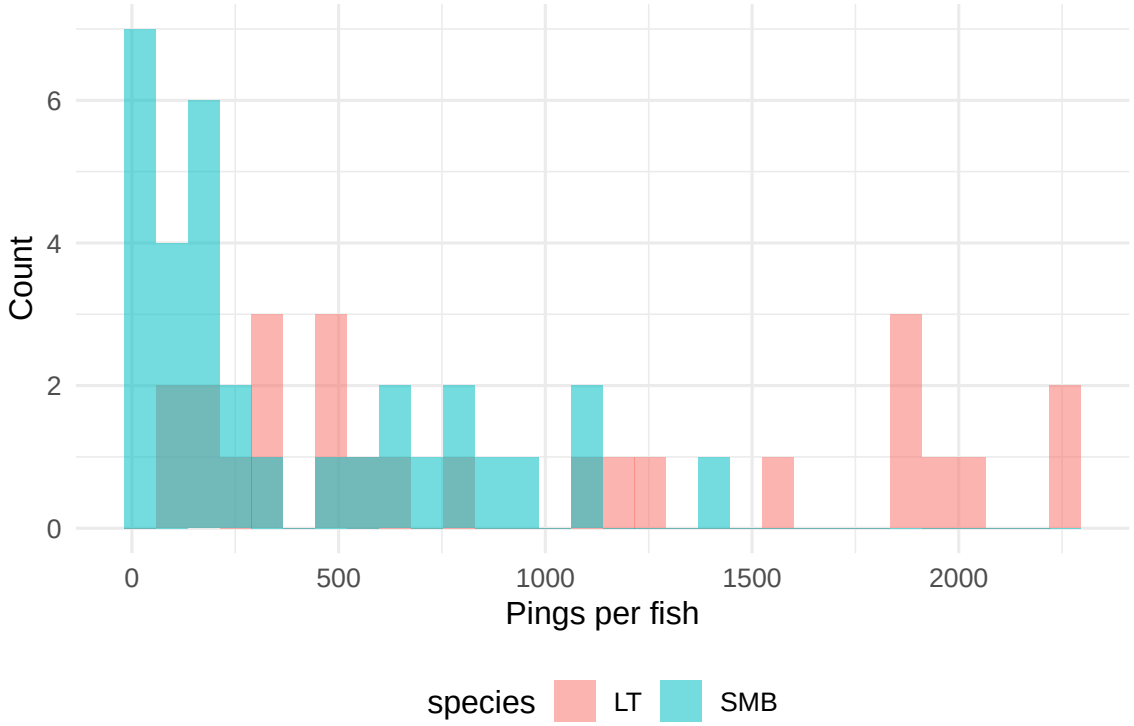


Figure 10: *Correlation heatmap of frequency columns (pooled across species).*

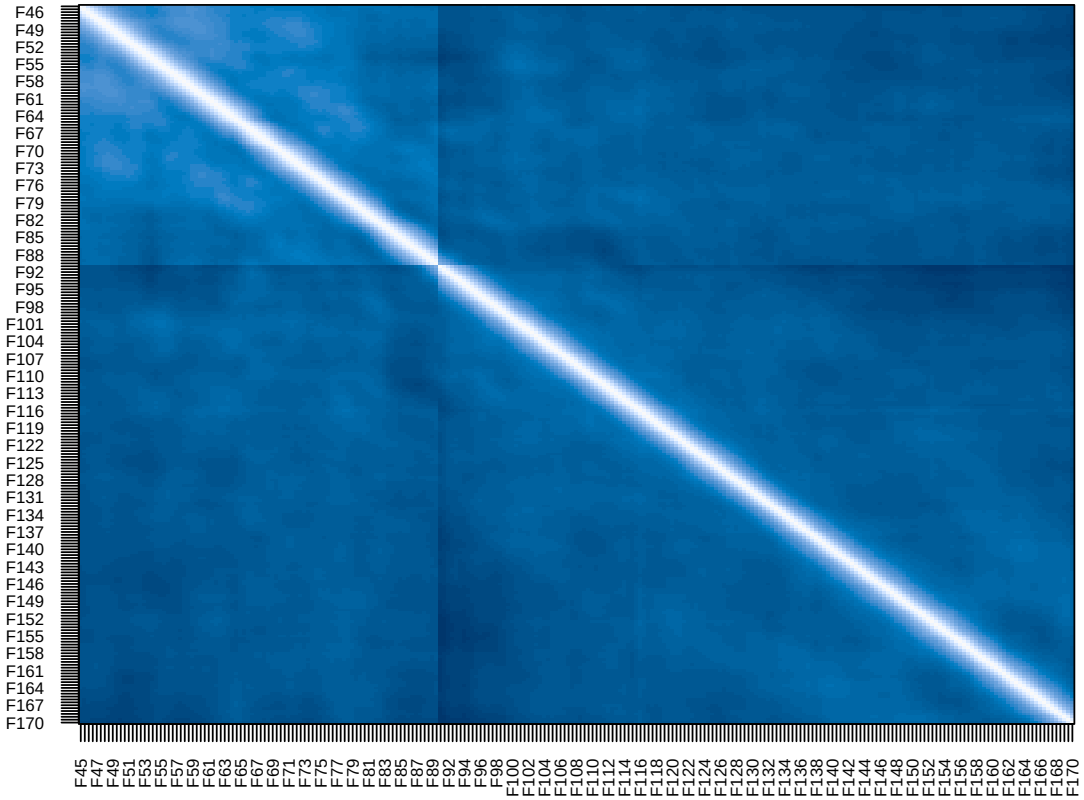


Figure 11: *Band-averaged TS by species (low: 50--90, mid: 90--130, high: 130--170 kHz).*

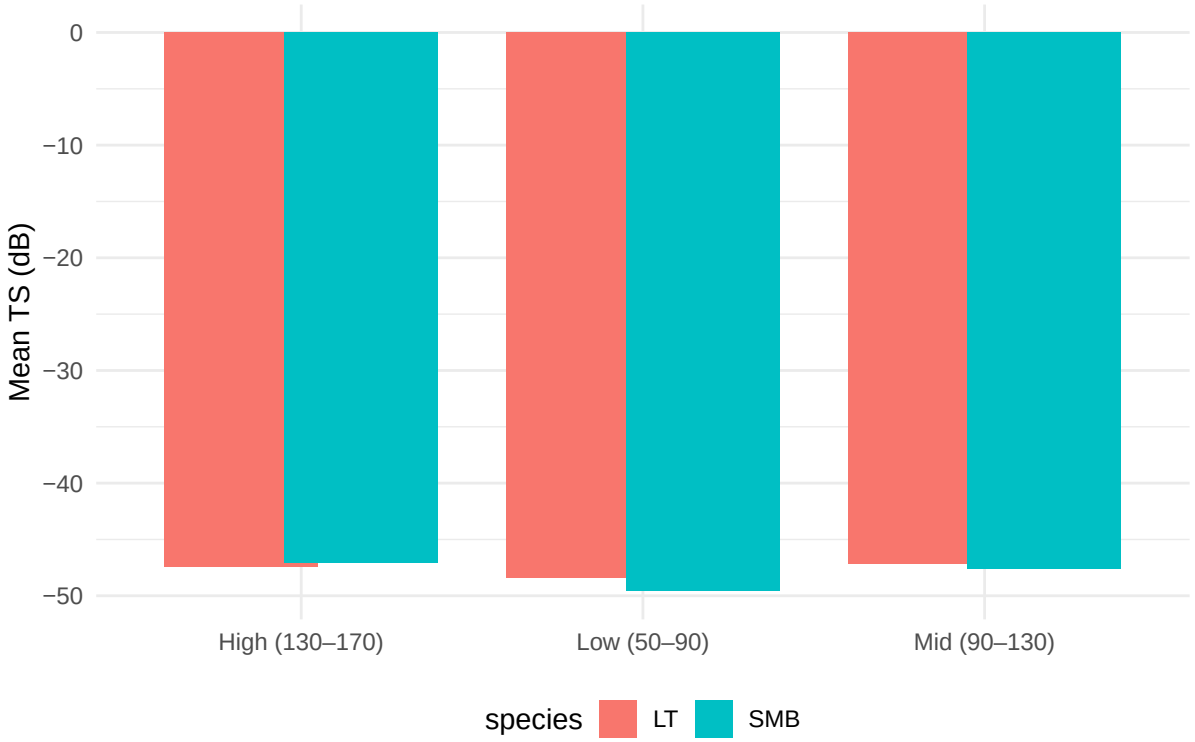
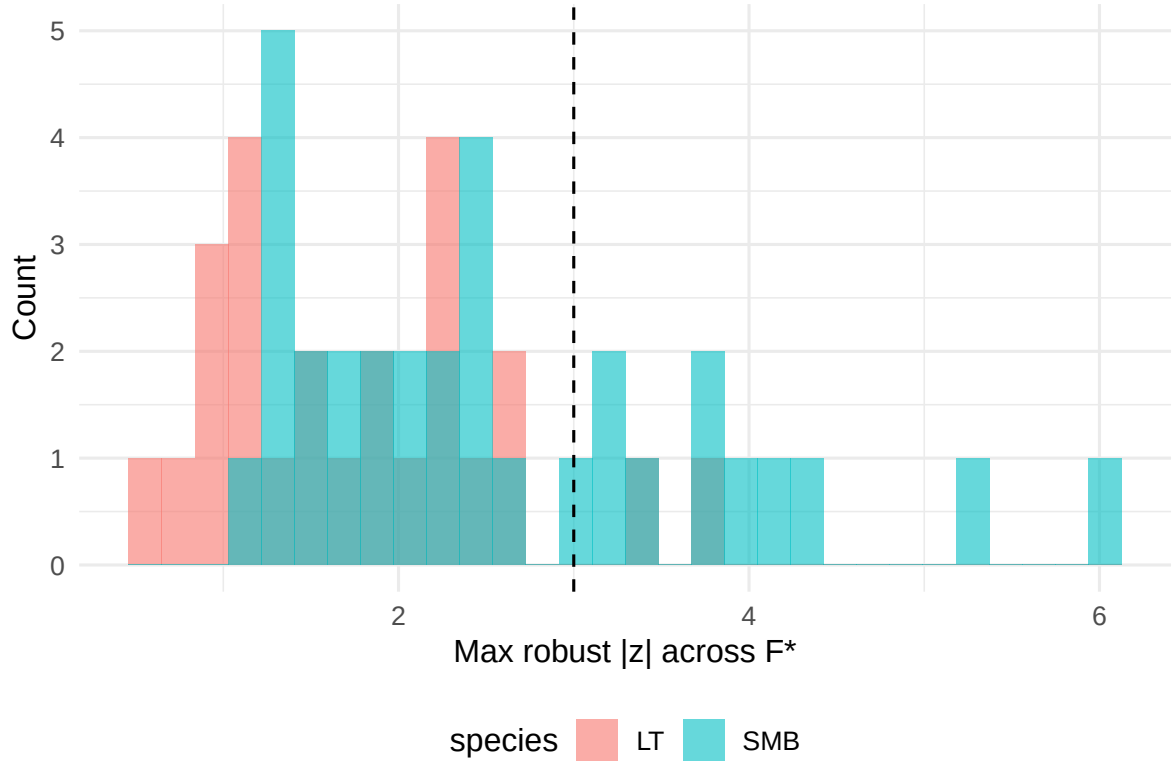


Figure 12: *Robust z-score of median FRC per fish (flag $> |3|$).*



References

Team, IH (2025). *Baseline Neural Networks for Frequency-Only Hydroacoustic Classification*. Unpublished internal report. Lake survey family baseline (RNN/LSTM) reproduced in this project.