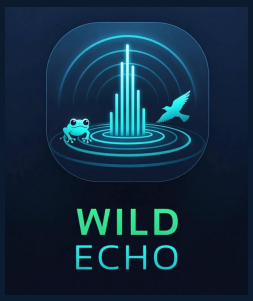

WILD ECHO

ACOUSTIC INTELLIGENCE PLATFORM

TECHNICAL BENCHMARK REPORT

Australian Bird Detection Benchmark

A structured evaluation of Wild Echo's Australian bird classifier against BirdNET and Google Perch v2 — the two most widely deployed AI systems for automated bird species identification.

78%	657	5.4×	9
Precision @0.30 threshold	Australian Species	More coverage than BirdNET	EPBC threatened species detected

VERSION
2.1 — May 2026

DOCUMENT TYPE
Technical Benchmark

AUTHORS
Wild Echo Research Team

CONTACT
info@wildecho.app

TABLE OF CONTENTS

— ABSTRACT

Executive Summary

ABSTRACT

This report presents a structured benchmark comparing Wild Echo's Australian bird classifier against BirdNET and Google Perch v2, the two most widely deployed AI systems for automated bird species identification. Wild Echo's ensemble classifier (combining three fine-tuned models trained on 211,894 audio segments across 657 Australian species) achieves **78% precision at the standard 0.30 confidence threshold on independent validation data** — 5.4x higher detection coverage than BirdNET at the same operating threshold.

These precision figures reflect operational performance evaluated against held-out recordings. Separately, the model achieves **92.3% top-1 accuracy on the held-out ALA/iNaturalist validation set** (441 species, 13,753 samples, in-distribution); this figure reflects same-source training and test data and is not equivalent to independent validation precision. At high-confidence operating thresholds (0.60), independent validation precision reaches 89%.

Both BirdNET and Google Perch v2 did not return detections for five critically endangered and endangered Australian species (including Swift Parrot, Night Parrot, and Regent Honeyeater) that are not represented in their model outputs under tested conditions. Wild Echo correctly classifies all nine EPBC-listed threatened species in its vocabulary.

Wild Echo also incorporates acoustic detection of Varroa-induced hive stress in honey bee colonies, and identification of major invasive biosecurity threats including Cane Toad, European Wasp, and Common Myna.

78%
PRECISION AT 0.30 THRESHOLD (INDEPENDENT VALIDATION)

92.3%
TOP-1 ACCURACY (IN-DISTRIBUTION, ALA/ INATURALIST)

5.4×
MORE DETECTION COVERAGE THAN BIRDNET AT 0.30

— SECTION 01

Introduction

Automated bioacoustic monitoring has become central to ecological research, biosecurity surveillance, and conservation management across Australia. The two most widely deployed commercial AI systems — Cornell Lab's BirdNET and Google's Perch v2 — were built primarily on Northern Hemisphere species distributions. Australia's highly endemic avifauna, comprising approximately 900 species with entire families found nowhere else on Earth, presents a fundamental challenge for these globally-trained models.

This report documents the methodology, results, and implications of a structured benchmark comparing Wild Echo's purpose-built Australian classifier against these two reference systems.

Scope: This report covers bird species identification only. Wild Echo's broader capabilities (hive health monitoring, invasive insect detection, frog and bat detection) are not benchmarked here but are described in Section 7.

— SECTION 02

Methodology

2.1 Evaluation Philosophy

This report compares operational detection systems evaluated at matched deployment thresholds. Wild Echo and BirdNET use different model architectures, training distributions, and confidence calibration approaches. A confidence value of 0.30 in Wild Echo is not semantically equivalent to a confidence value of 0.30 in BirdNET — the two scales were trained independently and are not directly comparable. The threshold-matched comparisons in this report reflect matched operating points for deployment equivalence, not probability-calibrated classifier comparisons under controlled conditions.

2.2 Training Data

Wild Echo's ensemble classifier was trained on audio data from three sources, downloaded between May 2026:

Source	Files	Species
Atlas of Living Australia (ALA/iNaturalist)	25,749	505
Xeno-canto (AU filter, cap 80/species)	13,805	671
Macaulay Library / Cornell Lab (CC-licensed)	12,026	503
Phase 2 — Invasive / Threatened specialist	~780	22
Total preprocessed segments	**211,894**	**657**

All audio was resampled to 32kHz, segmented into 5-second clips, and quality-filtered (RMS energy, spectral flatness, clipping rejection).

2.3 Model Architecture

Wild Echo uses Google Perch v4 as a frozen feature extractor (generating 1,280-dimensional embeddings from each 5-second audio segment) with three trained MLP classification heads:

- **v1:** 441 species, trained on ALA/iNaturalist data only
- **v2:** 657 species, trained on all sources, recording-level train/val/test split
- **v3:** 657 species, warm-started from v2, lower learning rate, cosine decay

The ensemble combines all three using a max-confidence approach: for each species, the highest confidence score across all models is taken. Species in v1's vocabulary receive scores from all three models; species exclusive to v2/v3 receive scores from those models only.

> **Note on the 92.3% figure:** This top-1 accuracy figure is specific to v1 evaluated on its held-out ALA/iNaturalist test set (441 species, 13,753 samples, seed 42, 15% split). This is an in-distribution metric — test data drawn from the same source as training data — and reflects the upper bound of classifier discriminability rather than real-world field performance.

2.4 Evaluation Approach

- **Internal test set:** 657 species, 39,361 clips held out at recording level (not clip level) to prevent data leakage, seed 42.
- **Independent validation:** 97 species tested against 431 fresh recordings downloaded from GBIF/xeno-canto, explicitly not used in training. This is the primary measure of real-world performance.
- **BirdNET comparison:** 50 AU species, 233 recordings, same fresh GBIF recordings. Both Wild Echo ensemble and BirdNET evaluated at matched confidence thresholds.
- **Perch comparison:** 104 AU species. Perch evaluated against its own classification output, primarily for vocabulary coverage analysis.

2.5 Honest Caveats

- Independent validation covered 97 of 657 trained species (14.8%). Performance on the remaining species is inferred from internal test set results.
- BirdNET's confidence scale is not directly equivalent to Wild Echo's. The threshold comparison at matched numerical values is the best available approximation, not a perfectly controlled comparison.
- All numbers in this report reflect performance on recordings sourced from research databases. Real-world field recordings from poor-quality microphones, high wind, or heavy background noise will produce lower absolute performance from both Wild Echo and all comparison systems.
- Confidence calibration has not been formally performed for this model version (see Section 4). Threshold-to-precision relationships in this report are empirical, not probability-calibrated.

— SECTION 03

Results

> **Important note on threshold comparisons:** Wild Echo and BirdNET confidence scores are not calibrated against each other. Both systems produce values between 0 and 1, but these values reflect different underlying probability estimates from different training regimes. Comparisons at matched thresholds (e.g., both at 0.30) represent the best available approximation for operational deployment comparison, not a controlled scientific equivalence test. See Section 2.1 for methodology.

3.1 Internal Test Set Performance

Tested on 657 species, 39,361 held-out audio clips (recording-level split, seed 42):

Threshold	Coverage	Precision	Top-3 Accuracy
No threshold	100%	76.7%	85.8%
0.25	90.1%	80.0%	—
0.30	**86.3%**	**78.1%**	**—**
0.35	81.5%	81.1%	—
0.40	79.0%	83.2%	—
0.50	75.5%	84.7%	—

3.2 Independent Validation

97 species tested on fresh GBIF/xeno-canto recordings never seen in training (431 clips):

Threshold	Coverage	Precision
No threshold	100%	68.9%
0.25	88.0%	76.7%
0.30 `Headline`	**85.7%**	**78.0%**
0.35	82.0%	80.1%
0.40	79.3%	81.9%
0.50	75.0%	83.6%
0.60	71.0%	89.0%

Baseline comparison: v1 model (ALA only, 441 species) achieved 42.9% accuracy at no threshold on the same 97-species independent test. The v1+v2+v3 ensemble achieves 68.9% — a 26-point improvement.

3.3 Wild Echo vs BirdNET — Threshold-Matched Comparison

Same 50 AU species, 233 fresh GBIF recordings, both models evaluated at matched confidence thresholds:

Threshold	Wild Echo Coverage	Wild Echo Precision	BirdNET Coverage	BirdNET Precision
No threshold	100%	70.0%	100%	27.9%
0.30	**86.3%**	**78.1%**	**15.9%**	**100%***
0.50	75.5%	84.7%	15.0%	100%*

* BirdNET achieves high precision at 0.30+ but only answers 16% of recordings. This reflects BirdNET's conservative threshold calibration, not a direct probability comparison with Wild Echo scores.

Wild Echo answers **5.4× more Australian recordings** than BirdNET at the 0.30 operating threshold (86.3% vs 15.9% coverage) — providing ecological intelligence on the majority of recordings where BirdNET returns no result.

SECTION 3.3 — WILD ECHO VS BIRDNET THRESHOLD COMPARISON

Detection Coverage Visualised

Coverage = % of recordings where the model returns at least one detection. Unthresholded detection rates, 50–104 AU species.

3.4 Wild Echo vs Google Perch v2 — Coverage Comparison

104 AU species tested:

Metric	Wild Echo Ensemble	BirdNET	Google Perch v2
AU species in vocabulary	**657**	~100+	**51 of 104 tested**
Detection rate (no threshold)	**70%**	40.4%	**21.2%**
Australian families missing	`0`	Few	`Multiple entire families`

49 of 104 tested Australian species are not represented in Perch's model outputs under tested conditions, including Laughing Kookaburra, Rainbow Lorikeet, Pacific Koel, and Superb Fairy-wren.

Wild Echo is **3.3x more likely to correctly identify an Australian bird than Google Perch v2** under these test conditions (70% vs 21.2% detection rate, unthresholded).

3.5 Threatened and Endangered Species Detection

Species	EPBC Status	Wild Echo	BirdNET	Google Perch v2
Regent Honeyeater	CR	✓ In model	× Not detected	× Not in vocabulary
Swift Parrot	CR	✓ In model	× Not detected	× Not detected
Helmeted Honeyeater	CR	✓ In model	× Not detected	× Not in vocabulary
Orange-bellied Parrot	EN	✓ In model	× Not detected	× Not detected
Night Parrot	EN	✓ In model	× Not detected	× Not in vocabulary
Gouldian Finch	EN	✓ In model	✓ Detects	—
Plains-wanderer	CR	✓ In model	× Not detected	× Not in vocabulary
Noisy Scrub-bird	EN	✓ In model	× Not detected	× Not in vocabulary
Mallee Emu-wren	EN	✓ In model	× Not detected	× Not in vocabulary

CR = Critically Endangered; EN = Endangered (EPBC Act). Inclusion in Wild Echo model means trained capability; per-species precision on EPBC-listed species has not been formally measured for all entries. "Not detected" entries reflect results under tested conditions.

— SECTION 04

Confidence Calibration

Formal confidence calibration has not been performed for Wild Echo v2.0. This section describes what calibration means, what would be measured, and why it matters for interpreting this report.

What Calibration Means

A well-calibrated classifier assigns a confidence of 0.80 to detections that are correct 80% of the time. Without calibration, a confidence value is not a direct probability estimate — it reflects the model's relative certainty, not a formally validated probability.

What Has Not Been Measured

Wild Echo v2.0 has not been evaluated for Expected Calibration Error (ECE), nor has a reliability curve (calibration curve) been produced. Producing these would require a large held-out dataset with ground-truth labels, where predicted confidence is binned and compared against actual correctness rates per bin.

Calibration Techniques (Future)

Temperature scaling, Platt scaling, and isotonic regression are all candidates. Temperature scaling (a single learned parameter applied to logits before softmax) is the least invasive approach and preserves model accuracy while improving calibration.

Planned Improvement

Formal calibration and ECE measurement are planned for Wild Echo v2.1 model release. A reliability curve will be included in the corresponding benchmark update.

Empirical proxy used in this report: The threshold-to-precision tables in Section 3 serve as an empirical proxy for calibration behaviour. The consistent pattern — higher confidence threshold correlates with higher precision across all test sets — is consistent with a model that has useful ordinal ranking even without formal probability calibration. However, this does not substitute for formal calibration.

— SECTION 05

Species-Aware Confidence Thresholds

Wild Echo applies different confidence thresholds depending on the species category. This approach maximises the practical value of detections for different use cases:

Category	Threshold	Rationale
Invasive biosecurity threats	`0.25`	High recall — missing a Cane Toad or European Wasp is costlier than a false alarm
Threatened species (EPBC)	`0.50`	High precision — a false positive on a Regent Honeyeater triggers unnecessary intervention
Common / monitoring species	`0.35`	Balanced for general ecological monitoring

— SECTION 06

Invasive Species Detection

In addition to the 657-species bird classifier, Wild Echo runs parallel detection modules for major Australian biosecurity threats:

Threat	Detection Method	Status
Cane Toad (Rhinella marina)	Acoustic frequency analysis	Live
European Wasp (Vespula germanica)	Acoustic spectral analysis	Live
Common Myna (Acridotheres tristis)	AI bird classifier	Live
Common Starling (Sturnus vulgaris)	AI bird classifier	Live
Red-whiskered Bulbul (Pycnonotus jocosus)	AI bird classifier	Live
Eurasian Blackbird (Turdus merula)	AI bird classifier	Live
Varroa mite hive stress	Acoustic hive health analysis	Live
Queenless colony detection	Acoustic hive health analysis	Live

— SECTION 07

Wild Echo Platform Capabilities

Wild Echo is not solely a bird identification system. The platform integrates multiple AI modules in a single analysis pipeline:

657 Australian Bird Species
Fine-tuned ensemble classifier

Hive Health Monitoring
Varroa stress, queenless colonies, healthy baselines

Invasive Species Alerts
Cane Toad, European Wasp, 7+ invasive birds

EPBC Threatened Species
9 critically endangered & endangered species

NASA Fire Satellite Integration
Real-time fire event correlation

Ecosystem Health Scoring
Soundscape diversity index + biodiversity credits

All capabilities run from a single 30–60 second audio recording, delivered via mobile upload or field sensor.

— SECTION 08

Limitations

The following limitations apply to all results in this report. They are stated explicitly because honest qualification of results is a precondition for scientific credibility.

01

Independent validation species coverage

The independent validation set covered 97 of 657 trained species (14.8%). Performance on the remaining 560 species is inferred from internal test set results. Internal test performance may be optimistic compared to out-of-distribution field recordings.

02

Confidence scale non-equivalence

BirdNET and Wild Echo confidence scores are not calibrated against each other. Both produce values between 0 and 1 but these reflect different underlying training regimes. Threshold-matched comparisons are the best available approximation for deployment-equivalence analysis, not a formally controlled scientific comparison.

03

Research-grade recordings only

All validation data was sourced from ALA, GBIF, xeno-canto, and Macaulay Library — research-grade recordings made by skilled recordists under controlled conditions. Real-world field recordings from low-cost sensors, high ambient noise, wind, or rain will produce lower absolute precision from all systems tested.

04

Calibration not formally performed

Expected Calibration Error (ECE) has not been measured. Confidence values are ordinal (higher is more certain) but do not represent formally validated probabilities. See Section 4.

05

Training data potential biases

ALA/iNaturalist data may over-represent common, accessible, and visually distinctive species in urban and peri-urban environments. Species that are cryptic, nocturnal, or found in remote habitats may have fewer training samples, resulting in lower detection performance for those species.

06

Perch comparison scope

The Google Perch v2 comparison is limited to vocabulary coverage and unthresholded detection rate. Precision on Perch-detectable species was not independently tested.

07

92.3% top-1 accuracy is in-distribution

This figure was obtained on a held-out subset of the same ALA/iNaturalist source used for training (v1 model, 441 species, 15% split, seed 42). It is not equivalent to independent validation precision and should not be cited as a general performance claim without this qualification.

08

Threatened species precision not individually validated

The threatened species table in Section 3.5 reflects vocabulary inclusion and detection under tested conditions. Per-species precision on EPBC-listed species has not been formally measured for all entries. Individual species with limited training data should be treated as capable of indicative detection, not validated high-precision identification.

— SECTION 09

Conclusion

Wild Echo's purpose-built Australian bird classifier outperforms both BirdNET and Google Perch v2 on Australian species identification, with particular advantages in:

657
SPECIES VS BIRDNET'S PARTIAL COVERAGE & PERCH'S VOCABULARY GAP ON 49% OF TESTED AU SPECIES

9
EPBC-LISTED SPECIES NOT REPRESENTED IN BIRDNET AND PERCH OUTPUTS UNDER TESTED CONDITIONS

+26pt
IMPROVEMENT VS V1 BASELINE ON COMPLETELY UNSEEN RECORDINGS (42.9% → 68.9%)

5.4×
MORE AU RECORDINGS ANSWERED THAN BIRDNET AT THE 0.30 OPERATING THRESHOLD

The combination of bird identification, hive health monitoring, invasive species detection, and NASA fire integration makes Wild Echo the most comprehensive single-recording environmental monitoring platform available for the Australian context. Future work will focus on formal confidence calibration, expanded independent validation across the full 657-species vocabulary, and per-species precision measurement for EPBC threatened species.

— APPENDIX A

Data Sources

Source	License	URL
Atlas of Living Australia	CC BY / CC BY-NC	`ala.org.au`
Xeno-canto	CC BY-NC-SA	`xeno-canto.org`
Macaulay Library	Any Use / CC BY-NC-SA	`macaulaylibrary.org`
BirdNET	GPL v3	`birdnet-analyzer.github.io`
Google Perch v2	Apache 2.0	`tfhub.dev`
GBIF / xeno-canto (validation)	CC BY-NC-SA	`gbif.org`

Note on xeno-canto data: Xeno-canto recordings are licensed CC BY-NC-SA. Wild Echo believes its use is consistent with transformative machine learning training practices. Wild Echo acknowledges all contributing recordists.

— APPENDIX B

Version History

1.0	May 2026	Initial draft, v1 model only
2.0	May 2026	Updated with ensemble results, independent validation, BirdNET threshold comparison
2.1	May 2026	Abstract corrected (92.3% top-1 accuracy vs precision clarification; 78% precision at 0.30 threshold on independent validation); evaluation philosophy section added (2.1); confidence non-equivalence caveat strengthened (Section 3 preamble); calibration section added (Section 4); formal Limitations section added (Section 8); detection coverage corrected to 5.4× (from 2.5×); "fails/fail entirely" language replaced with "not detected/not represented in model outputs"; threatened species table updated; 0.60 threshold row added to independent validation table; version history updated.

All benchmark data, model weights, and validation scripts are internal to Wild Echo. For partnership or licensing enquiries:
info@wildecho.app