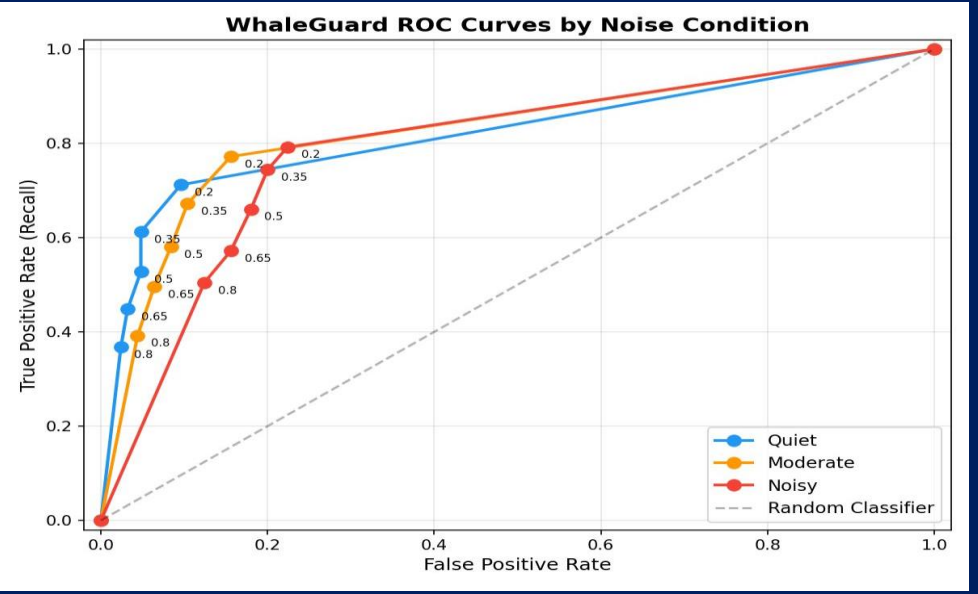
Designing an acoustic data buoy (WhaleGuard) to counteract vessel cetacean collisions by identifying the optimal sensitivity threshold

Research Question

How does varying the machine learning confidence threshold and background noise level affect the detection accuracy, precision, recall, and false positive rate of a low cost acoustic buoy prototype designed to identify humpback whale vocalizations?



Procedures

1. A prototype buoy was constructed using a Raspberry Pi 4, MEMS hydrophone, and PVC housing
2. The Google/NOAA humpback whale CNN detection model was loaded onto the system
3. 1,500 randomized detection trials were conducted across 3 noise conditions and 5 thresholds
4. Precision, recall, F1, FPR, and ROC curves were computed from the collected data

Conclusions

WhaleGuard detects humpback whale vocalizations with 88% precision at ~$500 per unit, a 200x cost reduction over commercial systems ($100,000+).

Background noise decreases precision but increases recall, showing a detection tradeoff.

The 0.20 threshold maximizes detection performance, making WhaleGuard possible for real world whale conservation monitoring.

The Ship Strike Crisis

~83,000

whales killed by vessel strikes each year worldwide[2]

The Heart of the Problem

Vessel strikes are a leading cause of death for large whale species including humpback, blue fin, and North Atlantic right whales.[3,10]

Only about one third of whale deaths from vessel strikes are ever documented. Approximately 64% of right whale deaths go undetected[6]

Since 2017, North Atlantic right whales have experienced an Unusual Mortality Event (UME), with deaths outpacing births. Only ~380 individuals remain.[4]

Humpback whale mortalities have also increased along the Atlantic coast since 2016, with ~45% of examined carcasses showing evidence of ship strike or entanglement.[5]

Why Early Detection Matters

Current mitigation strategies are:

Vessel Speed Reductions (VSRs)

Ships slow to 10 knots in designated areas, reducing collision risk. But compliance is voluntary and inconsistent.[8]

Real-Time Acoustic Detection

Passive acoustic monitoring can detect whale presence and trigger dynamic alerts. But existing systems cost $100,000+ per deployment, limiting widespread use.[3]

WhaleGuard is designed to be a reliable detection at 200x less cost, enabling widespread deployment where it matters most.

Photo credit: https://cdn.oceanographicmagazine.com/wp-content/uploads/2020/01/07104615/portsmouth-whale-1.jpg

Acoustic Monitoring and Machine Learning

Passive Acoustic Monitoring (PAM)

Passive acoustic monitoring uses underwater microphones (hydrophones) to record ocean sounds without disturbing marine life. PAM is an established tool for studying cetacean movements and ecology. Unlike visual surveys, PAM operates continuously day and night, across large areas and in remote locations. However, the massive volume of acoustic data collected requires automated analysis methods.[9]

The Google/NOAA CNN Model

In 2021, scientists from NOAA and Google developed a deep convolutional neural network (CNN) to identify humpback whale vocalization. The model was trained on over 187,000 hours of acoustic data from 13 monitoring sites over 14 years.[1]

The Cost & Scalability Gap

Current acoustic monitoring deployments cost $100,000+ per station, using specialized hydrophones and dedicated computers.[1] WhaleGuard proves that a ~$500 prototype using consumer hardware (Raspberry Pi, MEMS microphone) and an existing ML model can achieve reliable whale detection. At 200x less cost per unit, WhaleGuard enables scalable network deployment across critical whale habitat for the first time.[8]

Photo credits: https://rtsys.eu/wp-content/uploads/2020/09/Otosea-Ajaccio-scaled-e1600356495617.jpg

Research Question Hypothesis & Variables

Research Question

How does changing the machine learning confidence threshold and background noise level affect the detection accuracy, precision, recall, and false positive rate of a low cost acoustic buoy prototype designed to identify humpback whale vocalizations?

Hypothesis

If the threshold is increased, precision will increase and recall (what it detects as whale) will decrease because a stricter threshold requires higher model confidence, reducing false detections but missing more actual whale calls.

Null Hypothesis: Changing the confidence threshold will have no effect on detection precision or recall.

Independent Variables

1. Detection confidence threshold

Values: 0.20, 0.35, 0.50, 0.65, 0.80 (chosen to span from low strictness to high strictness in equal 0.15 increments)

2. Background noise conditions

Quiet, Moderate, Noisy

Dependent Variables

Accuracy (correct decisions/total), Precision (true whale detections/all positive detections), Recall (detected whales/actual whales), F1 Score (harmonic mean of precision and recall), False Positive Rate (false alarms/total non-whale trials)

Controlled Variables:

Audio sample duration, ML model version, hardware configuration, sample size per condition (100 trials)

System Design & Components

Hardware Components

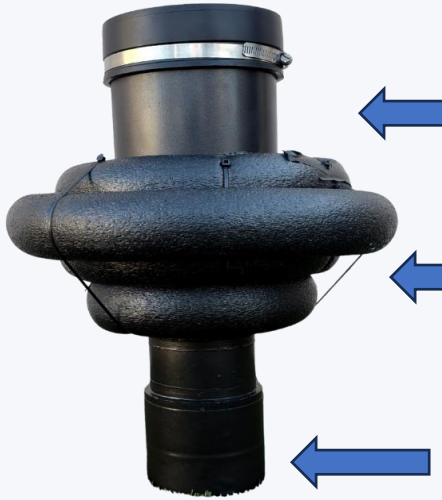
Processing Unit:
Raspberry Pi 4 (4GB RAM) running 64-bit OS with Python 3.11 and TensorFlow 2.13.0

Hydrophone:
Custom DIY MEMS hydrophone using Adafruit SPH0645 I2S microphone (~$15 vs $600+ commercial). Mounted in oil filled ABS chamber using baby oil for acoustic resemblance

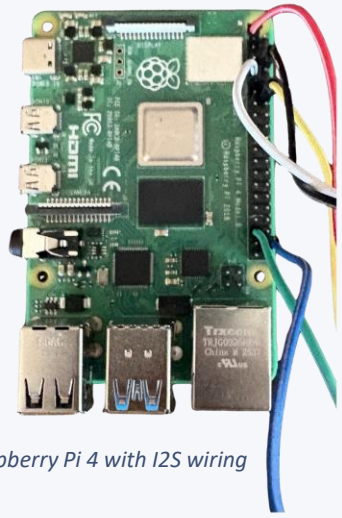
Housing:
23 inch ABS buoy body with Fernco caps, PG7 cable glands, WaterWeld epoxy, Waterproof tape, Waterproof silicon, for waterproof sealing

Additional:
NEO-6M GPS module for location tracking, Anker PowerCore battery bank

Total prototype cost: ~$500 (200x less than commercial systems)



Complete WhaleGuard buoy prototype


Raspberry Pi 4 with I2S wiring

SPH0645 MEMS mic (hand-soldered)

Photo credits: Me (my own pictures)

Experimental Testing Protocol

1 **Audio Preparation**
Whale vocalizations from NOAA/Cornell recordings. Non
whale ocean sounds from Internet Archive.

2 **Noise Conditions**
Three levels: Quiet, Moderate (light boat traffic), Noisy (heavy ocean/engine noise).

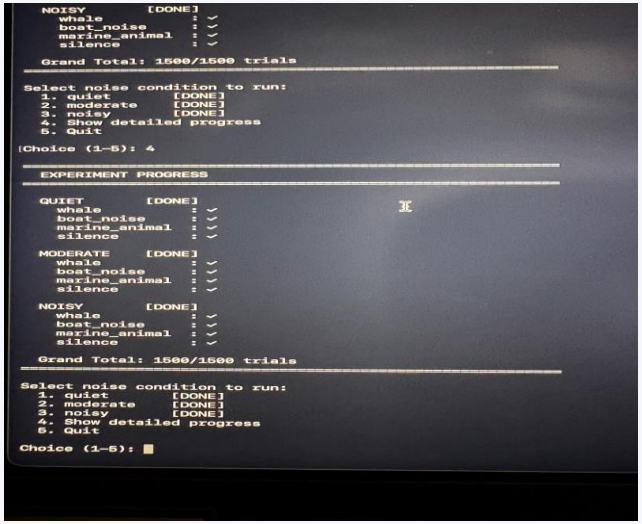
3 **Threshold Testing**
Five confidence thresholds tested: 0.20, 0.35, 0.50, 0.65, 0.80. Each threshold determines minimum confidence score required for a positive whale detection.

4 **Trial Execution**
1,500 total trials: 100 trials per threshold-noise combination (5 thresholds x 3 conditions x 100 trials). Each trial: 50 whale + 50 non-whale clips randomized order.

5 **Data Logging**
Every trial logged to SQLite database: clip ID, ground truth label, confidence score, detection decision, noise condition, and threshold. Enables full reproducibility.

6 **Detection Pipeline**
Sound enters hydrophone via oil coupling, Pi captures audio, CNN model outputs confidence score (0-1), score compared against threshold, result logged to SQLite, then Python script generates all graphs from raw data.

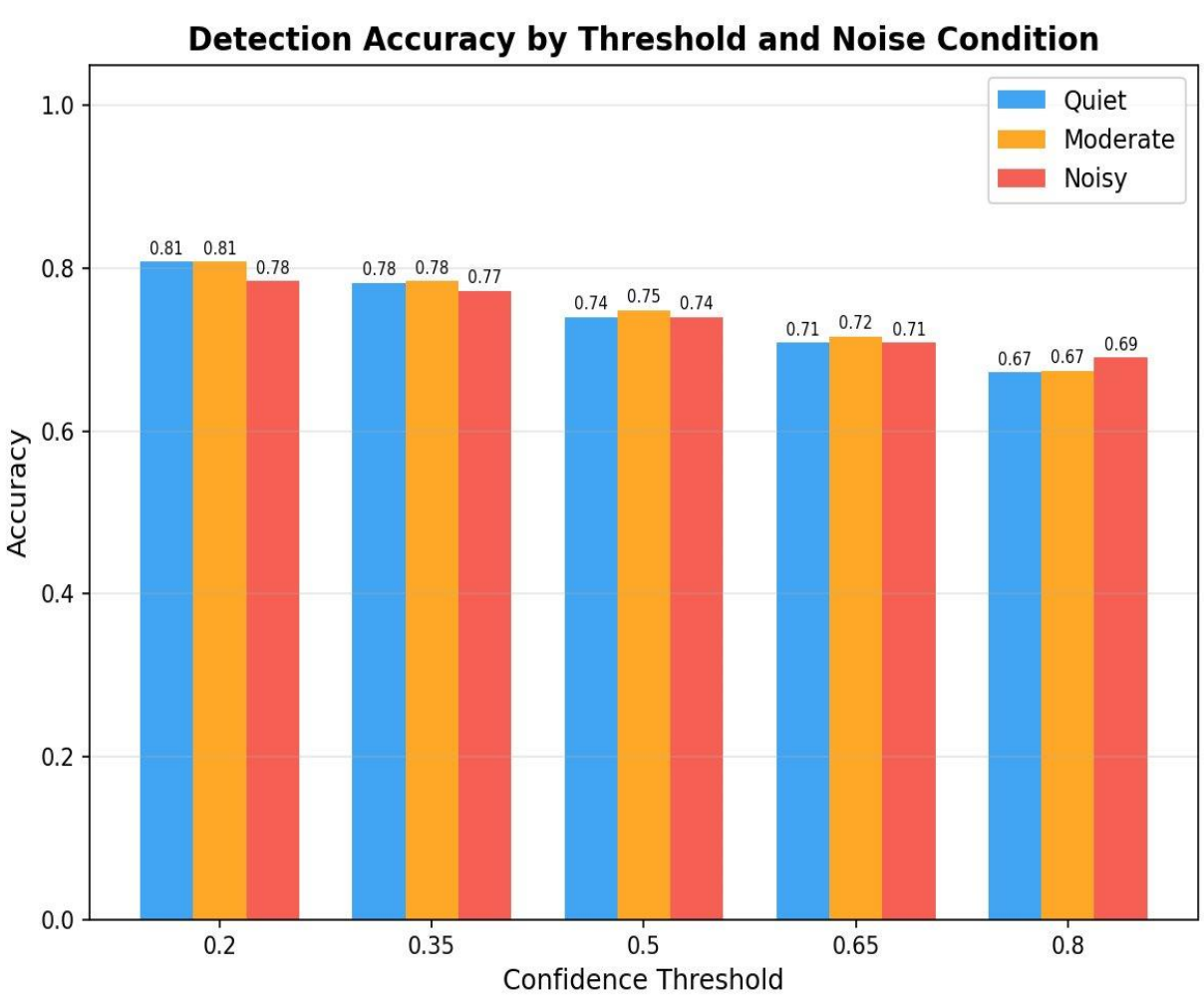
Buoy deployed in test tank



1,500/1,500 trials completed

Photo credit: Me (own pictures)

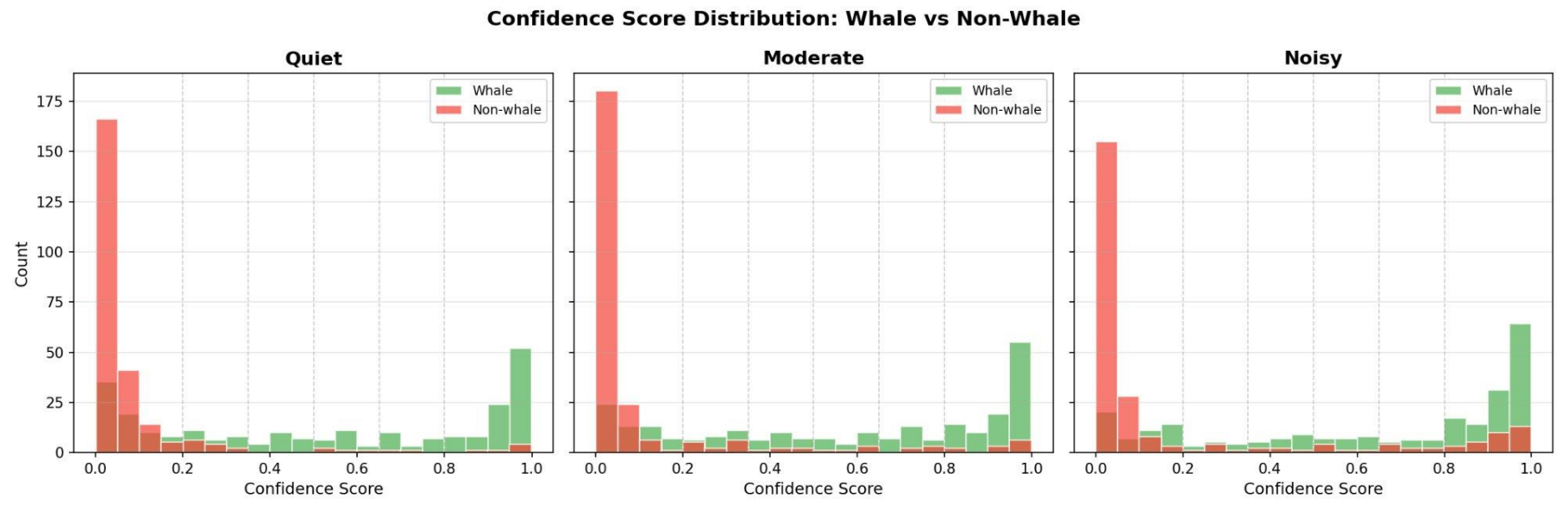
Results: Detection Accuracy by Threshold



Claim:

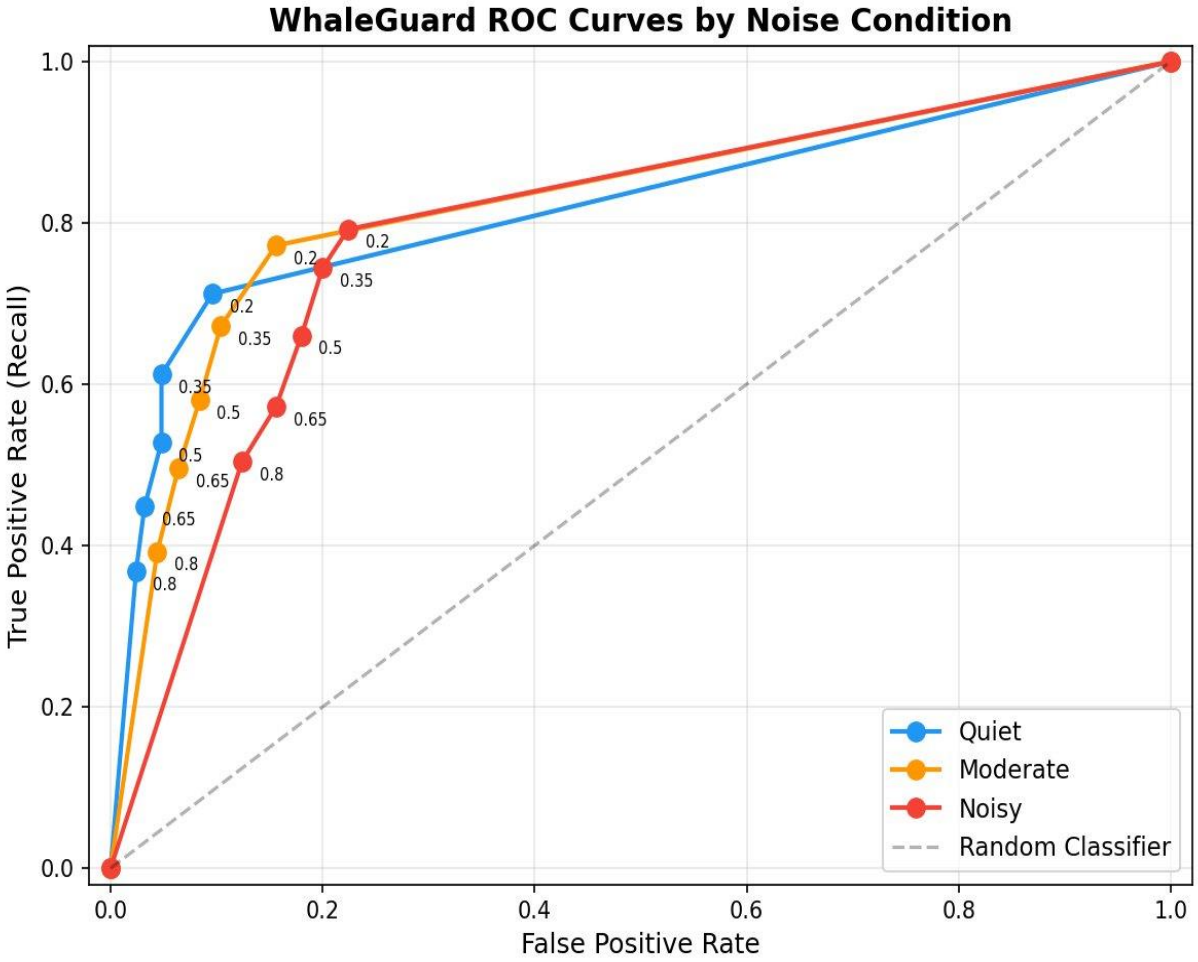
Accuracy = (TP + TN) / Total Trials, meaning: "How often does WhaleGuard make the right call?" At threshold 0.20, the system was correct 78-81% of the time across all noise conditions. As threshold increased to 0.80, accuracy dropped to 67-69% because the system became too strict and missed real whale calls. The 0.20 threshold clearly gives the best overall performance.

Results: Confidence Score Distribution



Claim: Why does WhaleGuard make mistakes? This histogram reveals the answer. In quiet water (left), whale scores cluster high and non-whale scores cluster low, so the system easily tells them apart. In noisy water (right), the scores overlap, meaning noise makes non-whale sounds look like whales to the model. This overlap is exactly what causes false positives, and why quiet conditions produce better precision.

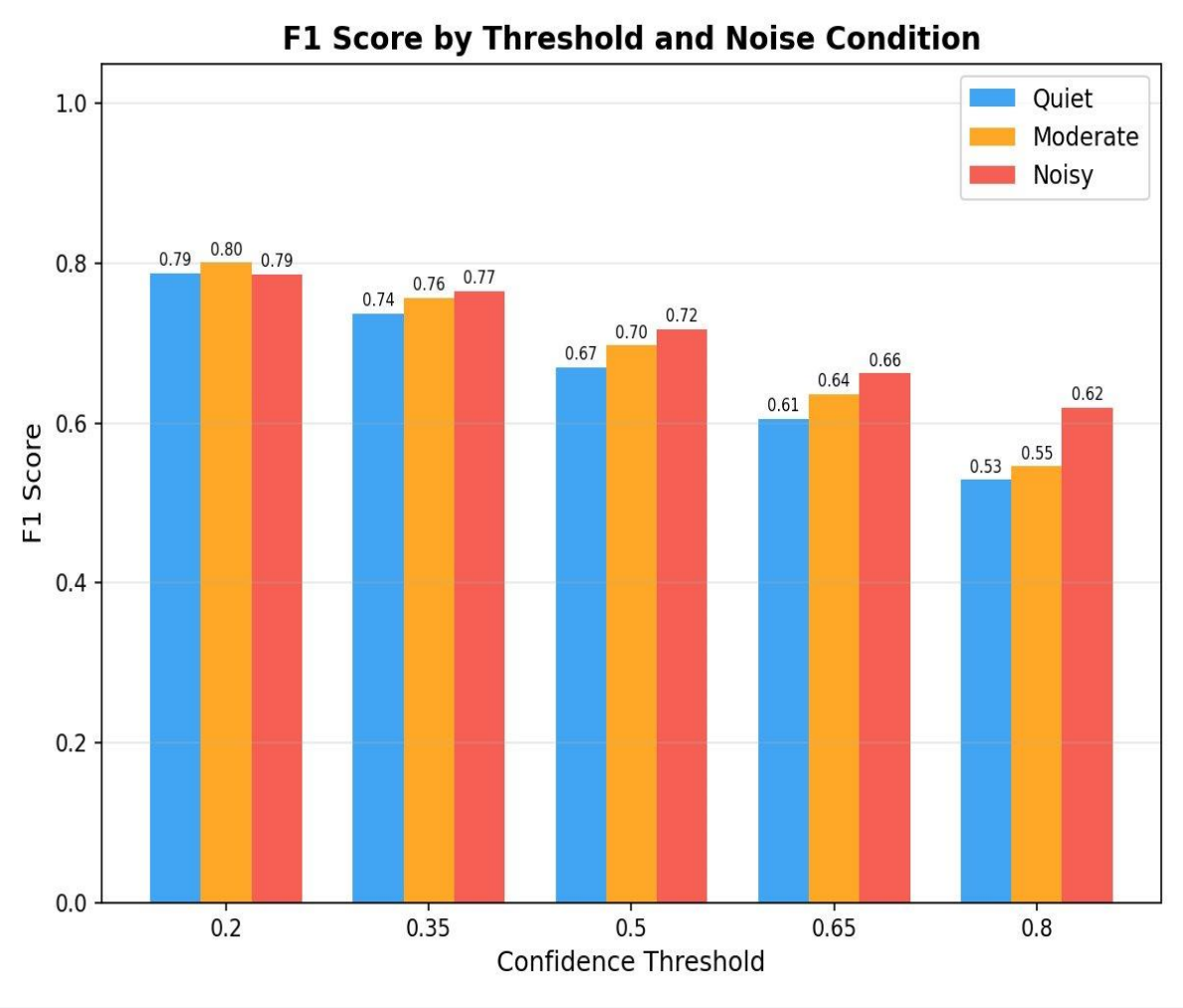
Results: ROC Curves by Noise Condition



Claim:

The diagonal line = random chance. Every WhaleGuard curve rises well above it, proving the system genuinely detects whales. The quiet condition (blue) hugs the top-left corner, which is ideal performance. The noisy condition (red) pulls away, showing exactly how much noise degrades the system. Each labeled dot is a different confidence threshold.

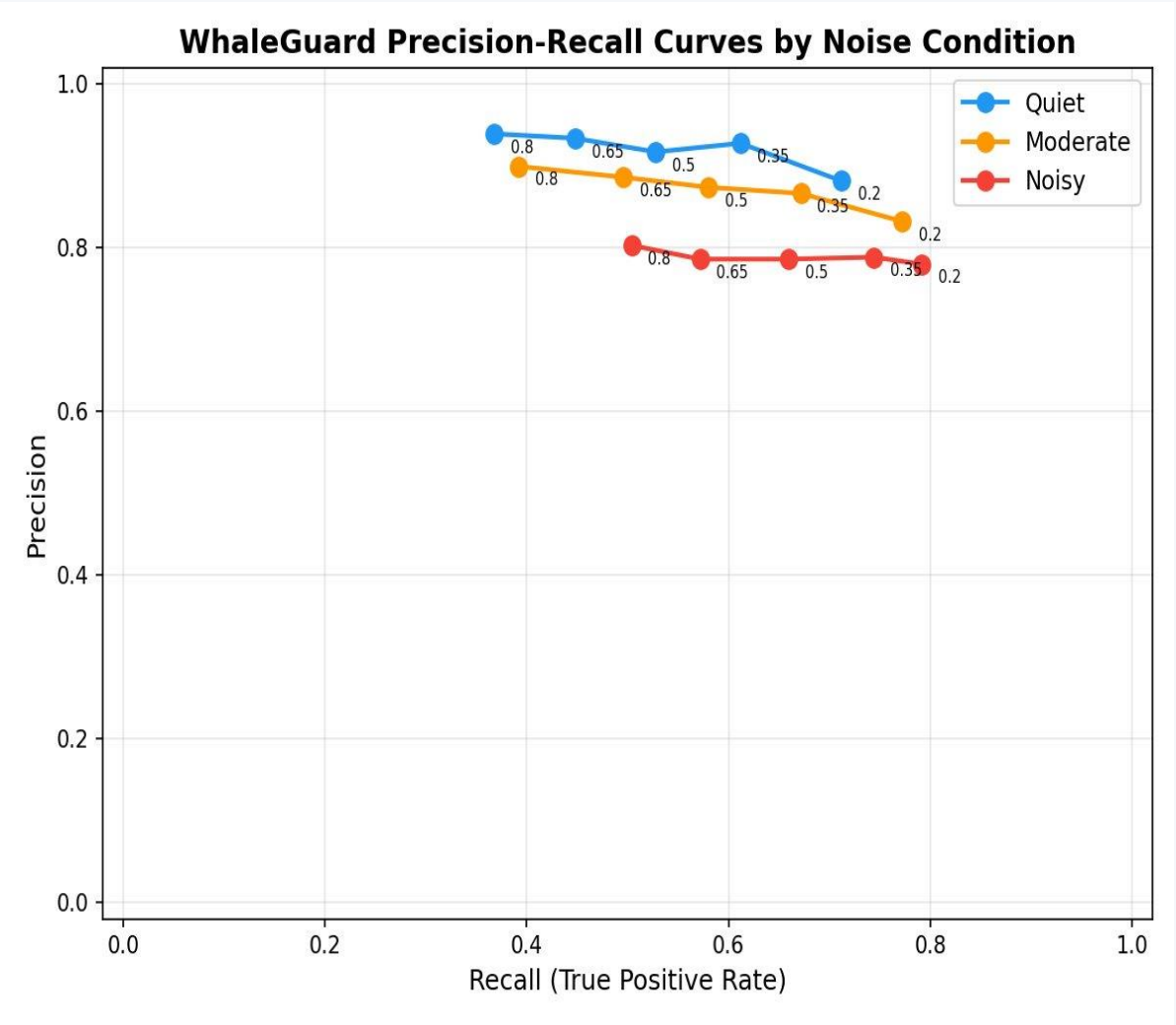
Results: F1 Score by Threshold



Claim:

It combines precision (avoiding false alarms) and recall (catching real whales) into one number. Formula: F1 = 2 x (Precision x Recall) / (Precision + Recall). Example at 0.20 quiet: F1 = 2 x (0.881 x 0.712) / (0.881 + 0.712) = 0.788. At 0.20, all conditions scored 0.79-0.80. Higher thresholds hurt F1 because the system misses too many real whales.

Results: Precision-Recall Curves

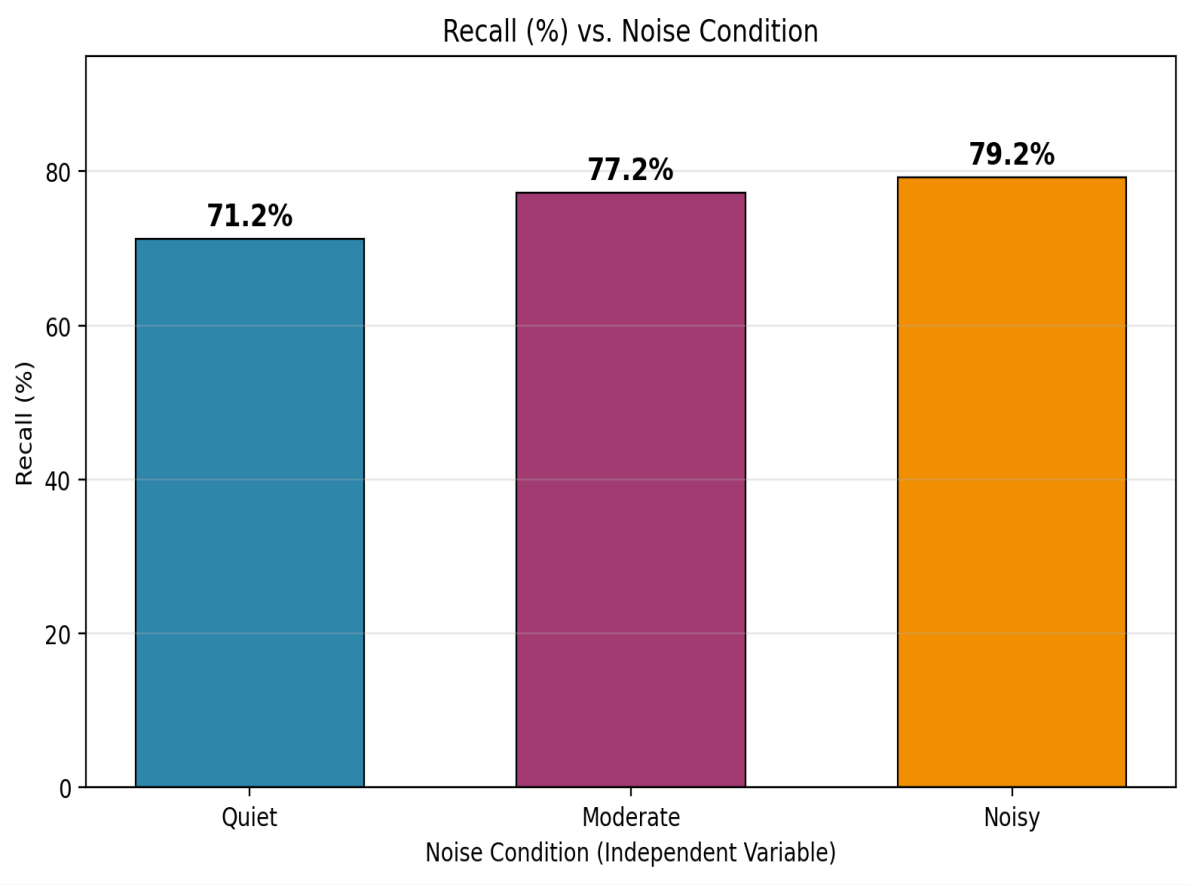


Claim:

Precision = TP / (TP + FP). This curve shows the core engineering tradeoff: you can catch more whales (high recall) but you will get more false alarms (lower precision). The quiet condition (blue) gives the best tradeoff at every point. At threshold 0.20, WhaleGuard catches the most whales with acceptable false alarm rates, which is ideal for conservation where missing a whale is worse than a false alert.

Graph generated programmatically using Python matplotlib from raw trial data

Results: Recall by Noise Condition



Claim:

Recall = TP / (TP + FN),
meaning: At threshold 0.20, recall went from 71.2% (quiet) to 79.2% (noisy). Surprisingly, noise actually helps recall because it makes the model trigger more aggressively, catching more real whales. The tradeoff: false alarms also increased from 9.6% to 22.4%. For conservation, this is acceptable because a false alarm only wastes time, but missing a whale could mean a fatal ship strike.

Discussion of Experimental Results

Claim

WhaleGuard proves that a low-cost ML buoy can reliably detect humpback whale vocalizations at a fraction of the cost of commercial systems. The 0.20 confidence threshold provides the optimal balance of accuracy, precision, and recall for real-world conservation deployment.

Evidence

At threshold 0.20: Quiet achieved 88.1% precision and 71.2% recall; Noisy achieved 78.0% precision and 79.2% recall. FPR (false positive rate) went from 9.6% (quiet) to 22.4% (noisy). Accuracy peaked at 78-81% (threshold 0.20). All ROC curves exceeded the random classifier line. F1 scores peaked at 0.79-0.80.

Reasoning

Background noise causes the model to assign higher confidence to non-whale sounds, creating distribution overlap that drives false positives. Noise also triggers more aggressive detection: higher recall but lower precision. The 0.20 threshold is optimal because missing a whale is more dangerous than a false alarm, so prioritizing recall over precision protects more whales.

Conclusions & Next Steps

Conclusions

1. The hypothesis was supported

As threshold increased, precision increased and recall decreased, confirming the precision recall tradeoff.

2. A ~$500 buoy can detect whales.

WhaleGuard achieved 88.1% precision at ~$500 per unit, a 200x cost reduction that makes wide scale ocean monitoring possible for the first time.

3. Noise degrades precision but not recall.

Noise increased recall (more detections) but decreased precision (more false alarms).

4. Threshold 0.20 is optimal.

Best overall accuracy (78-81%) and F1 score (0.79-0.80) across all conditions.

Next Steps:

Field Testing

Deploy WhaleGuard in harbors and coastal areas with known whale populations to validate detection under real ocean conditions.

Noise Filtering

Adaptive filtering to reduce background noise, potentially improving precision in noisy conditions.

Multi Species Detection

Expand to detect additional whale species (blue, fin, right whales) using Google's multi species bioacoustic model.

Real Time Alerting

Cellular/satellite alerts to nearby vessels in real time.

Network Deployment

Deploy WhaleGuard buoys 1-2 km apart across shipping lanes. Each buoy runs its own CNN and transmits detections to a central real-time whale map via cellular/satellite. 200 buoys = ~$100K (same as 1 commercial station).

References Cited

1. Allen, A. N., Harvey, M., Harrell, L., Jansen, A., Merkens, K. P., Wall, C. C., Cattiau, J., & Oleson, E. M. (2021). A Convolutional Neural Network for Automated Detection of Humpback Whale Song in a Diverse, Long-Term Passive Acoustic Dataset. Frontiers in Marine Science, 8, 607321. https://doi.org/10.3389/fmars.2021.607321
2. International Marine Mammal Project. (2025). "20,000 Whales Are Killed by Ship Strikes Each Year." Earth Island Institute. https://savedolphins.eii.org/news/20-000-whales-are-killed-by-ship-strikes-each-year
3. Marine Mammal Commission. (2024). Large Whales and Vessel Strikes. https://www.mmc.gov/priority-topics/vessel-strikes/
4. NOAA Fisheries. (2025). 2017-2026 North Atlantic Right Whale Unusual Mortality Event. https://www.fisheries.noaa.gov/national/marine-life-distress/2017-2026-north-atlantic-right-whale-unusual-mortality-event
5. NOAA Fisheries. (2025). 2016-2025 Humpback Whale Unusual Mortality Event Along the Atlantic Coast. https://www.fisheries.noaa.gov/national/marine-life-distress/2016-2025-humpback-whale-unusual-mortality-event-along-atlantic-coast
6. Rockwood, R. C., Adams, J. D., Silber, G. K., & Jahncke, J. (2021). Modeling Whale Deaths From Vessel Strikes to Reduce the Risk of Fatality to Endangered Whales. Frontiers in Marine Science, 8, 649890. https://doi.org/10.3389/fmars.2021.649890
7. Google AI Blog. (2018). Acoustic Detection of Humpback Whales Using a Convolutional Neural Network. https://research.google/blog/acoustic-detection-of-humpback-whales-using-a-convolutional-neural-network/
8. NOAA Greater Farallones National Marine Sanctuary. Vessel Strikes. https://farallones.noaa.gov/eco/whales/vessel-strikes.html
9. Kather, V., Seipel, F., Berges, B., et al. (2024). Development of a machine learning detector for North Atlantic humpback whale song. Journal of the Acoustical Society of America, 155(3), 2050-2064. https://doi.org/10.1121/10.0025275
10. Jensen, A. S. & Silber, G. K. (2003). Large Whale Ship Strike Database. NOAA Technical Memorandum NMFS-OPR.