

Open-Set Underwater Acoustic Recognition: A Multi-Protocol Benchmark with Mixture-of-Experts and OOD-Aware Evaluation

Luis Paulo A. Guedes and Rafael Lamêgo P. L. Peron

Abstract—This paper presents a unified Wolfset benchmark for joint motor–state recognition, motor classification under operating-condition shift, state classification with unknown-state detection, and exploratory multi-label motor-presence inference. Raw waveforms, time–frequency representations, classical classifiers, convolutional neural networks, and a specialized mixture-of-experts (MoE) are evaluated under shared file partitions. The MoE combines Low-Frequency Analysis and Recording (LOFAR)-like and Detection of Envelope Modulation on Noise (DEMON)-like views, matching the highest in-distribution (ID) Macro-F1 and out-of-distribution (OOD) rejection rate in the joint benchmark, at 0.879 and 0.625, respectively, while achieving an OOD area under the receiver operating characteristic curve (AUROC) of 0.868. A multi-criteria framework relates model selection to operational priorities. No model dominates all protocols: Mel/MFCC perform consistently in the joint task, LFCC+CNN2D achieves the highest motor Macro-F1 under operating-condition shift, and classical cepstral models best separate known and unknown states.

Keywords—Underwater acoustic recognition, open-set recognition, out-of-distribution detection, mixture-of-experts, time–frequency representations, multi-label classification.

I. INTRODUCTION

Underwater acoustic target recognition is challenged by environmental variability, noise, class imbalance, and limited generalization [1], [2], [3]. Traditional methods combine LOFAR, DEMON, and mel-frequency cepstral coefficients (MFCCs) with statistical classifiers, whereas deep models mainly use time–frequency representations, as in *DeepShip* [4].

Most studies assume closed-set conditions, although real recordings may contain unseen motor–state combinations, unknown states, or simultaneous sources. Open-set recognition [8] and out-of-distribution (OOD) detection are therefore relevant. Common approaches include maximum softmax probability [9], temperature scaling [10], and energy-based scores [11], but remain underexplored in underwater acoustics.

Mixture-of-experts (MoE) architectures [6], [7] employ specialized subnetworks and have also been applied to underwater target recognition [5].

This work contributes: (i) a unified file-level benchmark for joint motor–state, motor-only, and state-only protocols; (ii) a comparison of raw waveforms, handcrafted representations,

and classical and neural classifiers; (iii) a band-wise MoE combining LOFAR-like and DEMON-like views; and (iv) an exploratory multi-label reinterpretation of the joint backbone. The results show how protocol design affects generalization and model selection.

II. RELATED WORK

Underwater acoustic recognition has progressed from handcrafted descriptors to deep spectro-temporal models [1], [2], [3]. *DeepShip* highlighted specialized convolutional models [4], while MoE structures address intra-class diversity and inter-class similarity [5], extending local-expert [6] and sparse-gated [7] formulations. Open-set recognition handles unknown test classes [8], whereas maximum softmax [9], calibration [10], and energy scores [11] provide common OOD baselines.

III. METHODOLOGY

A. Overview

Wolfset recordings sampled at 44.1 kHz were used [14]. The selected subset contained 99 files and 23 joint motor–state labels: 11 ID classes and 12 classes reserved for joint open-set evaluation. The split comprised 47 training, 18 validation, 18 ID-test, and 16 external-test files. Splits were shared across protocols, and label encoders used only ID data.

Signals were converted to mono, resampled to 16 kHz, peak-normalized, and divided into 5.0 s windows with a 2.5 s hop. Benchmarks used up to 40 windows per file, whereas the MoE and motor-presence protocols used up to 16.

For a file x with N_x windows, the file posterior is

$$\bar{\mathbf{p}}(x) = \frac{1}{N_x} \sum_{i=1}^{N_x} \mathbf{p}_i. \quad (1)$$

Single-label prediction uses $\hat{y}(x) = \arg \max_c \bar{p}_c(x)$, whereas the multi-label task marginalizes joint-class posteriors into thresholded motor-presence probabilities.

B. Comparative Benchmarks

The protocols used *ID train*, *ID validation*, *ID test*, and *external test* partitions. The joint external set contained unseen motor–state combinations and was treated as OOD. The motor-only external set contained known motors under unseen operating conditions and therefore measured condition-shift generalization. In the state-only protocol, external recordings were divided into known and unknown states.

The joint and unknown-state protocols report ID Macro-F1, rejection rate, and area under the receiver operating characteristic curve (AUROC). Baseline file confidence was

$$s(x) = \max_k \bar{p}_k(x), \quad (2)$$

with threshold

$$\tau = Q_{0.05}(s(x) : x \in \text{ID validation}). \quad (3)$$

Files with $s(x) < \tau$ were rejected, using the same rule across baselines.

Inputs comprised raw waveforms and MFCC, Mel, short-time Fourier transform (STFT), linear-frequency cepstral coefficient (LFCC), and Bark representations from 20 Hz to 8 kHz.

The evaluated configurations were h256_c13_b64, h512_c26_b128, and h2048_c40_b256, with

$$n_{\text{fft}} = \max(512, 2 \cdot \text{hop}). \quad (4)$$

Spectral maps were resized to 128 frames and z -score normalized; classical models used temporal means and standard deviations.

Models included CNN1D, CNN2D, RBF-SVM with $C = 10$ and $\gamma = \text{scale}$, random forest with 400 trees, and an MLP with 256 and 128 units. Neural models used Adam, learning rate 10^{-3} , weight decay 10^{-4} , batch size 32, weighted cross-entropy, and early stopping with patience 5 over at most 20 epochs.

C. Specialized MoE Protocol

A specialized MoE was used for the joint task (Figure 1). Its two-view *Spec2* input ($128 \times 256 \times 2$) combined LOFAR-like Mel ($n_{\text{fft}} = 2048$, $\text{hop} = 512$, 128 bands, 20–6000 Hz) and DEMON-like envelope features ($n_{\text{fft}} = 1024$, $\text{hop} = 256$), both log-compressed and z -score normalized.

The spectrum was split into N bands, each processed by a convolutional expert with a bidirectional GRU, temporal attention, and a 128-dimensional embedding. A DEMON-energy gate produced softmax fusion weights. Training used AdamW, regularization, mixup, OneCycle scheduling, and EMA. Open-set scores were temperature-scaled, with thresholds calibrated using pseudo-OOD validation samples; baselines used the ID-quantile rule.

D. Expert-Count and Computational Audit

The number of experts was swept over $N \in \{2, 3, 4, 5\}$ with all other settings fixed. Table I reports the selection metrics and available computational records. These values were used only for architecture selection; final held-out results appear in Table VI.

TABLE I
EXPERT-COUNT ABLATION.

Experts	Val F1	ID F1	OOD Reject	AUROC
2	0.8099	1.0000	1.0000	0.8681
3	0.7190	0.8333	0.2500	0.7326
4	0.6887	0.7576	0.1875	0.7257
5	0.6887	0.7697	0.3125	0.7639

The two-expert model achieved the best validation F1, ID file-level F1, OOD rejection, and AUROC, with 2.79M parameters, 10.95 MB, and 16.84 s runtime. Larger models increased parameters, size, and runtime without improving performance.

E. Derived Multi-Label Evaluation from the Fixed Backbone

An exploratory multi-label motor-presence evaluation used the fixed MoEJointN2 backbone without fine-tuning. The Spec2 frontend, architecture, and checkpoint were unchanged. Let $p(c | x)$ denote the posterior of joint class c , and $b_k(c) \in \{0, 1\}$ indicates the presence of motor k . Its marginal probability is

$$p(\text{bit}_k | x) = \sum_c p(c | x) b_k(c). \quad (5)$$

Files were divided into non-composite training and validation subsets and composite development and test subsets. File-level mean, top- k mean, and maximum aggregation were combined with motor-specific thresholds. This analysis assesses whether a joint single-label backbone preserves information about simultaneous motors, rather than representing a fully trained multi-label system.

IV. RESULTS AND DISCUSSION

Tables II, III, and IV summarize the feature-based experiments for the *joint*, *motor-only*, and *state-only* scenarios, respectively. In each case, different acoustic representations and time–frequency resolutions are compared using SVM, RF, MLP, and CNN2D (C2D), together with closed-set and, when applicable, open-set metrics. For reference, the raw-waveform baseline is summarized in Table V, allowing a direct comparison between waveform-based and feature-based approaches.

In the *joint* scenario, the highest ID Macro-F1 was 0.879, achieved by RF+Mel, CNN2D+MFCC, MLP+Mel, and MLP/CNN2D+Bark configurations (Table II). SVM+MFCC with hop length 256 samples achieved the highest OOD rejection rate (0.625), while CNN2D+Mel with hop length 2048 samples obtained the highest AUROC (0.896). Thus, closed-set and open-set performance were not directly aligned. Compared with the raw-waveform baseline in Table V, spectral features improved ID classification and ID–OOD separation.

In the *motor-only* scenario, several configurations reached an ID Macro-F1 of 1.000. CNN2D+LFCC with hop length 2048 samples achieved the highest external Macro-F1 (0.931), compared with 0.471 for the raw-waveform baseline, highlighting the value of low-frequency harmonic and narrowband components.

In the *state-only* scenario, the highest ID Macro-F1 was 0.962. RF+MFCC and SVM/MLP+LFCC with hop length 2048 samples achieved the highest unknown-state rejection rate (0.875), while RF+MFCC obtained the highest AUROC (0.993). Spectral features therefore mainly improved known–unknown separation.

Table VI compares the proposed MoE with the strongest joint baselines.

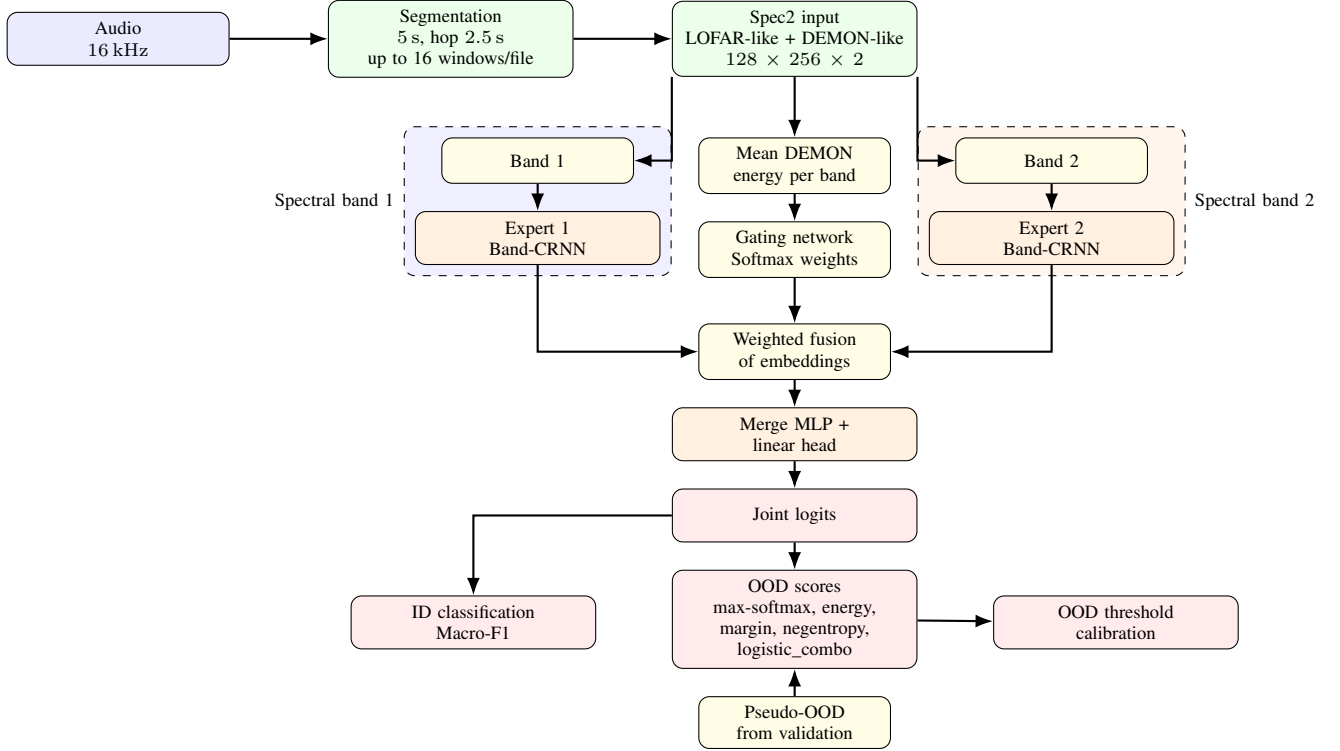


Fig. 1. Schematic of Protocol 4: specialized MoE model for the joint task, with Spec2 input, two spectral experts, and explicit OOD-score calibration.

TABLE II

FEATURE-BASED RESULTS FOR THE JOINT ID/OOD BENCHMARK. C2D DENOTES THE CNN2D MODEL. THE BEST VALUE IN EACH ROW WITHIN EACH METRIC BLOCK IS HIGHLIGHTED IN GREEN.

Feature	SR [kHz]	FFT Size	Hop Length	Extra Parameters	ID Macro-F1				OOD Reject				OOD AUROC			
					SVM	RF	MLP	C2D	SVM	RF	MLP	C2D	SVM	RF	MLP	C2D
MFCC	16	512	256	$n_c = 13$	0.833	0.833	0.833	0.879	0.625	0.375	0.562	0.250	0.875	0.767	0.806	0.774
	16	1024	512	$n_c = 26$	0.833	0.719	0.794	0.879	0.562	0.438	0.562	0.375	0.819	0.792	0.788	0.778
	16	4096	2048	$n_c = 40$	0.712	0.712	0.833	0.833	0.500	0.375	0.250	0.188	0.865	0.837	0.792	0.833
Mel Spectrogram	16	512	256	$n_b = 64$	0.758	0.758	0.818	0.770	0.438	0.312	0.250	0.250	0.809	0.806	0.733	0.809
	16	1024	512	$n_b = 128$	0.773	0.879	0.879	0.770	0.438	0.375	0.375	0.188	0.799	0.816	0.757	0.726
	16	4096	2048	$n_b = 256$	0.773	0.879	0.794	0.833	0.250	0.312	0.125	0.062	0.736	0.806	0.653	0.896
STFT	16	512	256	$n_b = 64$	0.788	0.758	0.758	0.770	0.188	0.312	0.188	0.188	0.792	0.809	0.705	0.712
	16	1024	512	$n_b = 128$	0.773	0.780	0.773	0.770	0.375	0.125	0.125	0.062	0.753	0.812	0.691	0.830
	16	4096	2048	$n_b = 256$	0.758	0.758	0.773	0.600	0.312	0.188	0.125	0.562	0.802	0.802	0.729	0.736
LFCC	16	512	256	$n_c = 13$	0.773	0.758	0.758	0.788	0.312	0.250	0.250	0.250	0.684	0.750	0.632	0.767
	16	1024	512	$n_c = 26$	0.773	0.758	0.652	0.758	0.188	0.375	0.062	0.250	0.778	0.788	0.681	0.688
	16	4096	2048	$n_c = 40$	0.758	0.648	0.758	0.682	0.188	0.188	0.375	0.062	0.868	0.806	0.719	0.708
Bark Spectrogram	16	512	256	$n_b = 64$	0.697	0.697	0.818	0.770	0.312	0.375	0.188	0.188	0.809	0.809	0.760	0.733
	16	1024	512	$n_b = 128$	0.697	0.758	0.879	0.879	0.375	0.438	0.188	0.062	0.781	0.837	0.740	0.639
	16	4096	2048	$n_b = 256$	0.733	0.712	0.794	0.788	0.375	0.312	0.188	0.062	0.743	0.826	0.656	0.882

No baseline dominated all protocols, as performance depended on the evaluation objective and data distribution. Mel and MFCC best balanced ID classification and ID–OOD separation in the joint task, whereas LFCC+CNN2D performed best under operating-condition shift. MFCC/LFCC with classical models were more effective for unknown-state detection, indicating that feature and model selection should be protocol-specific.

A. Multi-Criteria Model Selection

To formalize these trade-offs, model selection used a multi-criteria decision-making formulation based on normalized criteria and operational preference weights [12], [13]. For criterion k , scenario c , and model m , min–max normalization was

$$\widetilde{M}_k^{(c)}(m) = \frac{M_k^{(c)}(m) - \min_{j \in \mathcal{M}_c} M_k^{(c)}(j)}{\max_{j \in \mathcal{M}_c} M_k^{(c)}(j) - \min_{j \in \mathcal{M}_c} M_k^{(c)}(j) + \varepsilon}, \quad (6)$$

TABLE III

FEATURE-BASED RESULTS FOR MOTOR CLASSIFICATION UNDER OPERATING-CONDITION SHIFT. C2D DENOTES THE CNN2D MODEL. THE BEST VALUE IN EACH ROW WITHIN EACH METRIC BLOCK IS HIGHLIGHTED IN GREEN.

Feature	SR [kHz]	FFT Size	Hop Length	Extra Parameters	Val Macro-F1				ID Macro-F1				External Macro-F1			
					SVM	RF	MLP	C2D	SVM	RF	MLP	C2D	SVM	RF	MLP	C2D
MFCC	16	512	256	$n_c = 13$	0.922	0.778	0.922	0.922	0.893	0.871	0.893	0.758	0.694	0.702	0.694	0.876
	16	1024	512	$n_c = 26$	0.922	0.738	0.922	0.880	0.893	0.871	0.893	0.914	0.524	0.627	0.681	0.819
	16	4096	2048	$n_c = 40$	0.871	0.728	0.778	0.922	0.842	0.871	0.893	1.000	0.527	0.488	0.574	0.738
Mel Spectrogram	16	512	256	$n_b = 64$	0.922	0.922	0.922	0.880	0.871	0.922	1.000	0.826	0.702	0.702	0.702	0.693
	16	1024	512	$n_b = 128$	0.778	0.922	0.922	0.947	0.922	0.922	1.000	0.851	0.702	0.702	0.694	0.560
	16	4096	2048	$n_b = 256$	0.778	0.778	0.778	0.947	0.842	0.871	1.000	0.947	0.702	0.702	0.694	0.406
STFT	16	512	256	$n_b = 64$	0.922	0.922	0.922	0.922	0.920	0.948	0.920	0.958	0.751	0.702	0.702	0.776
	16	1024	512	$n_b = 128$	0.922	0.728	0.922	0.958	0.920	0.948	0.920	0.958	0.702	0.702	0.694	0.824
	16	4096	2048	$n_b = 256$	0.922	0.728	0.922	0.922	0.871	0.920	0.920	1.000	0.665	0.702	0.767	0.674
LFCC	16	512	256	$n_c = 13$	0.920	0.920	1.000	1.000	0.948	1.000	0.948	0.825	0.747	0.751	0.680	0.863
	16	1024	512	$n_c = 26$	0.778	0.778	0.778	1.000	0.920	0.948	0.920	0.743	0.747	0.640	0.694	0.801
	16	4096	2048	$n_c = 40$	0.728	0.728	0.778	0.958	0.948	0.948	1.000	0.743	0.705	0.604	0.655	0.931
Bark Spectrogram	16	512	256	$n_b = 64$	0.922	0.922	0.922	0.947	0.922	0.922	1.000	0.787	0.627	0.627	0.694	0.603
	16	1024	512	$n_b = 128$	0.922	0.922	0.922	0.958	0.871	0.871	1.000	0.958	0.627	0.627	0.694	0.653
	16	4096	2048	$n_b = 256$	0.778	0.922	0.922	0.880	0.871	0.871	1.000	0.738	0.702	0.702	0.694	0.500

TABLE IV

FEATURE-BASED RESULTS FOR THE STATE-ONLY ID/OOD BENCHMARK. C2D DENOTES CNN2D, AND THE COMMON 16 KHz SAMPLING RATE IS OMITTED. OPEN-SET PERFORMANCE IS MEASURED BY UNKNOWN-STATE REJECTION AND ID-VERSUS-UNKNOWN AUROC. THE BEST VALUE IN EACH ROW AND METRIC BLOCK IS HIGHLIGHTED IN GREEN.

Feature	FFT Size	Hop Length	Extra Parameters	ID Macro-F1				Unknown Reject				AUROC			
				SVM	RF	MLP	C2D	SVM	RF	MLP	C2D	SVM	RF	MLP	C2D
MFCC	512	256	$n_c = 13$	0.887	0.679	0.887	0.873	0.500	0.375	0.250	0.250	0.924	0.889	0.924	0.611
	1024	512	$n_c = 26$	0.887	0.679	0.679	0.887	0.875	0.500	0.625	0.375	0.958	0.972	0.986	0.819
	4096	2048	$n_c = 40$	0.887	0.679	0.679	0.887	0.750	0.875	0.500	0.000	0.944	0.993	0.931	0.889
Mel Spectrogram	512	256	$n_b = 64$	0.887	0.753	0.887	0.887	0.500	0.625	0.250	0.500	0.958	0.903	0.896	0.903
	1024	512	$n_b = 128$	0.962	0.679	0.962	0.829	0.500	0.500	0.250	0.500	0.951	0.896	0.896	0.896
	4096	2048	$n_b = 256$	0.887	0.679	0.962	0.810	0.750	0.625	0.250	0.125	0.965	0.910	0.882	0.875
STFT	512	256	$n_b = 64$	0.962	0.753	0.962	0.962	0.500	0.625	0.500	0.750	0.924	0.875	0.903	0.958
	1024	512	$n_b = 128$	0.962	0.753	0.962	0.905	0.500	0.500	0.250	0.625	0.924	0.882	0.882	0.736
	4096	2048	$n_b = 256$	0.962	0.962	0.962	0.762	0.375	0.750	0.000	0.375	0.896	0.861	0.868	0.799
LFCC	512	256	$n_c = 13$	0.962	0.707	0.753	0.962	0.625	0.625	0.500	0.000	0.833	0.833	0.812	0.785
	1024	512	$n_c = 26$	0.962	0.707	0.887	0.915	0.750	0.625	0.500	0.375	0.903	0.840	0.833	0.701
	4096	2048	$n_c = 40$	0.962	0.633	0.841	0.726	0.875	0.375	0.875	0.000	0.882	0.806	0.854	0.292
Bark Spectrogram	512	256	$n_b = 64$	0.887	0.679	0.887	0.829	0.375	0.750	0.125	0.250	0.938	0.910	0.889	0.785
	1024	512	$n_b = 128$	0.887	0.679	0.887	0.924	0.750	0.750	0.375	0.000	0.944	0.938	0.924	0.743
	4096	2048	$n_b = 256$	0.887	0.679	0.887	0.924	0.875	0.625	0.250	0.500	0.944	0.924	0.847	0.889

TABLE V

RAW-WAVEFORM BASELINE ACROSS ALL BENCHMARKS USING CNN1D.

METRIC DEFINITIONS ARE PROVIDED IN SUBSECTION III-B.

Benchmark	Val	ID	External F1	Reject	AUROC
Joint	–	0.682	–	0.438	0.726
Motor-only	0.903	0.914	0.471	–	–
State-only	–	0.962	–	0.625	0.889

and the normalized criteria were combined as

$$\mathcal{S}^{(c)}(m) = \prod_{k=1}^{K_c} \left(\widetilde{M}_k^{(c)}(m) + \varepsilon \right)^{w_k^{(c)}}, \quad \sum_k w_k^{(c)} = 1. \quad (7)$$

The selected configuration was $m_c^* = \arg \max_{m \in \mathcal{M}_c} \mathcal{S}^{(c)}(m)$.

The weights reflect operational priorities. Joint-task weights were 0.40 for ID Macro-F1 and 0.30 each for OOD rejection and AUROC. Motor-only weights were 0.15, 0.35, and 0.50 for validation, ID-test, and external-condition Macro-F1, respectively. State-only weights were 0.30 each for ID Macro-F1 and unknown-state rejection and 0.40 for AUROC.

Among the evaluated models, MoEJointN2 ranked first with $\mathcal{S} \approx 0.967$, followed by SVM+MFCC with hop length 256 and $\mathcal{S} \approx 0.908$. CNN2D+STFT was selected for motor-only classification with hop length 512 ($\mathcal{S} \approx 0.819$) and state-only classification with hop length 256 ($\mathcal{S} \approx 0.935$). Rankings depend on the weights.

TABLE VI
COMPARISON OF THE SELECTED MoE AND REPRESENTATIVE JOINT ID/OOD BASELINES.

Category	Feature / Input	Model	ID Macro-F1	OOD Reject	OOD AUROC	Comment
Waveform baseline	Raw waveform	CNN1D	0.682	0.438	0.726	Lowest closed-set accuracy, with only moderate open-set behavior.
Classical baseline	MFCC ($n_c = 13$), FFT = 512, Hop = 256	SVM	0.833	0.625	0.875	Highest baseline rejection rate and high AUROC, but lower ID accuracy.
Classical baseline	Mel ($n_b = 128$), FFT = 1024, Hop = 512	RF	0.879	0.375	0.816	Highest ID accuracy among classical baselines, but weaker rejection.
Deep baseline	MFCC ($n_c = 13$), FFT = 512, Hop = 256	CNN2D	0.879	0.250	0.774	High ID accuracy but weaker open-set behavior.
Deep baseline	Mel ($n_b = 256$), FFT = 4096, Hop = 2048	CNN2D	0.833	0.062	0.896	Highest AUROC, but low threshold-based rejection and lower ID accuracy.
Proposed	LOFAR-like + DEMON-like Spec2	MoEJointN2	0.879	0.625	0.868	Ties the highest ID Macro-F1 and rejection rate while maintaining competitive AUROC, providing the strongest weighted trade-off under the stated criteria.

B. Multiple-Motor Examples

Table VII shows representative composite recordings. The fixed joint backbone detects both motors in E1 and E5, but only the dominant motor in E2–E4 and E6, likely because the single-label softmax suppresses weaker sources. Marginalization cannot recover motors with low joint-class posteriors, indicating partial transfer and the need for dedicated multi-label training.

TABLE VII
REPRESENTATIVE COMPOSITE-SET EXAMPLES (WAV-LEVEL) USING MOTOR-BITS WITH MAX AGGREGATION AND OPTIMIZED PER-MOTOR THRESHOLDS [0.80, 0.40, 0.15, 0.75, 0.35].

ID	True	Pred. ($P(M_1, \dots, M_5) \rightarrow$ Decision
E1	M3+M4	[0.10, 0.05, 0.58, 0.80, 0.05] $\rightarrow$ M3+M4
E2	M2+M4	[0.25, 0.65, 0.13, 0.14, 0.10] $\rightarrow$ M2
E3	M4+M5	[0.07, 0.05, 0.04, 0.98, 0.02] $\rightarrow$ M4
E4	M2+M5	[0.26, 0.69, 0.14, 0.11, 0.07] $\rightarrow$ M2
E5	M2+M3	[0.32, 0.82, 0.17, 0.19, 0.09] $\rightarrow$ M2+M3
E6	M1+M5	[0.96, 0.11, 0.02, 0.02, 0.01] $\rightarrow$ M1

Legend:

E1=A00220T00N00.wav; E2=A02020T00N00.wav;
E3=A00011T00N00.wav; E4=A02002T00N00.wav;
E5=A02200T00N00.wav; E6=A20002T00N00.wav.

V. CONCLUSION

This work evaluated joint motor–state, motor-only, state-only, and exploratory motor-presence protocols. The two-expert MoE tied the best joint ID Macro-F1 (0.879) and OOD rejection rate (0.625), with an AUROC of 0.868. Increasing N to five increased the parameter count from 2.79M to 6.93M without improving performance. CNN2D+LFCC achieved the best motor Macro-F1 under operating-condition shift (0.931), while RF+MFCC achieved the highest unknown-state AUROC (0.993). Model selection was protocol-dependent. Limitations include the use of a single dataset and small external partitions. Future work will examine cross-dataset evaluation, repeated trials, confidence intervals, dedicated multi-label training, and improved OOD calibration.

REFERENCES

- [1] Z. Lei, X. Lei, N. Wang and Q. Zhang, “Present status and challenges of underwater acoustic target recognition technology: A review,” *Frontiers in Physics*, vol. 10, p. 1044890, 2022, doi: 10.3389/fphy.2022.1044890.
- [2] X. Luo, L. Chen, H. Zhou and H. Cao, “A Survey of Underwater Acoustic Target Recognition Methods Based on Machine Learning,” *Journal of Marine Science and Engineering*, vol. 11, no. 2, p. 384, 2023, doi: 10.3390/jmse11020384.
- [3] H. I. Hummel, R. van der Mei and S. Bhulai, “A survey on machine learning in ship radiated noise,” *Ocean Engineering*, vol. 298, p. 117252, 2024, doi: 10.1016/j.oceaneng.2024.117252.
- [4] M. Irfan, J. Zheng, S. Ali, M. Iqbal, Z. Masood and U. Hamid, “DeepShip: An underwater acoustic benchmark dataset and a separable convolution based autoencoder for classification,” *Expert Systems with Applications*, vol. 183, p. 115270, 2021, doi: 10.1016/j.eswa.2021.115270.
- [5] Y. Xie, J. Ren and J. Xu, “Unraveling complex data diversity in underwater acoustic target recognition through convolution-based mixture of experts,” *Expert Systems with Applications*, vol. 249, p. 123431, 2024, doi: 10.1016/j.eswa.2024.123431.
- [6] R. A. Jacobs, M. I. Jordan, S. J. Nowlan and G. E. Hinton, “Adaptive Mixtures of Local Experts,” *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991, doi: 10.1162/neco.1991.3.1.79.
- [7] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. V. Le, G. E. Hinton and J. Dean, “Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer,” in *Proc. International Conference on Learning Representations (ICLR)*, 2017.
- [8] W. J. Scheirer, A. Rocha, A. Sapkota and T. E. Boulton, “Toward Open Set Recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1757–1772, 2013, doi: 10.1109/TPAMI.2012.256.
- [9] D. Hendrycks and K. Gimpel, “A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks,” in *Proc. International Conference on Learning Representations (ICLR)*, 2017.
- [10] C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger, “On Calibration of Modern Neural Networks,” in *Proc. 34th International Conference on Machine Learning (ICML)*, vol. 70, pp. 1321–1330, 2017.
- [11] W. Liu, X. Wang, J. D. Owens and Y. Li, “Energy-based Out-of-distribution Detection,” in *Advances in Neural Information Processing Systems*, vol. 33, pp. 21464–21475, 2020.
- [12] R. L. Keeney and H. Raiffa, *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. New York: John Wiley & Sons, 1976.
- [13] E. Triantaphyllou, *Multi-Criteria Decision Making Methods: A Comparative Study*. Dordrecht: Kluwer Academic Publishers, 2000, doi: 10.1007/978-1-4757-3157-6.
- [14] V. Lobo, N. Pessanha Santos and R. Moura, “A High-Quality Underwater Acoustic Dataset for Algorithm Development and Analysis,” *Scientific Data*, vol. 12, p. 1323, 2025, doi: 10.1038/s41597-025-05564-x.