

Avaliação de Desempenho e Generalização de Redes Neurais para Classificação Acústica Subaquática no *Dataset* IARA

Leandro Kfuri de Oliveira¹, Rigel Procópio Fernandes², José M. de Seixas³ e Allyson Andrade da Silva¹

¹Centro de Instrução Almirante Alexandrino (CIAA), Penha/Rio de Janeiro - Brasil

²Diretoria de Sistemas de Armas da Marinha (DSAM), Centro/Rio de Janeiro - Brasil

³Universidade Federal do Rio de Janeiro (UFRJ), Laboratório de Processamento de Sinais, COPPE/Poli, Rio de Janeiro - Brasil

Resumo—Este artigo apresenta uma avaliação sistemática de quatro arquiteturas de redes neurais profundas, a saber, CNN Baseline, CRNN, Transformer-DWC e Conformer, para classificação de contatos acústicos subaquáticos no *dataset* IARA, coletado na Baía de Santos. Os modelos são avaliados em três cenários: condições ideais, i.e., sem adição de ruído externo; desempenho sob ruído gaussiano sintético e sob ruído de fundo real; e generalização geográfica entre o IARA (Brasil) e o ShipsEar (Espanha). A CRNN obteve a maior acurácia em condições ideais ($66,14\% \pm 1,17$) e o Conformer manteve o desempenho mais estável sob ruído gaussiano (-5 dB: $45,79\% \pm 1,86$). Sob ruído oceânico real extremo, contudo, as quatro arquiteturas apresentam acurácia estatisticamente indistinguível, e a diferença entre elas manifesta-se na estrutura do erro: entre os modelos baseados em atenção, a taxa de Falso Negativo, i.e., embarcação classificada como ruído de fundo, varia de $15,61\% \pm 14,66$ no Transformer-DWC a $40,52\% \pm 6,53$ no Conformer. Na transferência IARA→ShipsEar, o Conformer alcançou $41,29\%$, superando em $10,06$ pontos percentuais o *baseline* CNN reportado no trabalho original do IARA; a avaliação na direção inversa foi conduzida apenas para dois dos quatro modelos, sendo portanto parcial. Os resultados oferecem subsídios para a seleção de arquiteturas em sistemas de monitoramento de águas jurisdicionais brasileiras.

Palavras-Chave—Sonar Passivo, Redes Neurais, SisGAAz.

I. INTRODUÇÃO

O Sistema de Gerenciamento da Amazônia Azul (SisGAAz) é a principal iniciativa da Marinha do Brasil para o monitoramento contínuo de cerca de 5,6 milhões de km² de águas jurisdicionais, por meio de hidrofones fixos que captam sinais acústicos sem comprometer a furtividade do sistema.

O espectro de contatos de interesse não se restringe a embarcações mercantes: veículos de superfície não tripulados (VSNT), que operam em missões de patrulha, inspeção e escolta, também o compõem [1]. A consciência situacional plena depende de duas etapas complementares — classificação do contato, foco deste trabalho, e estimação de sua direção de chegada (*Direction of Arrival*, DOA). Rocha et al. [2] abordam essa segunda etapa com um arranjo tetraédrico de quatro hidrofones, reduzindo o erro angular RMS de $21,4^\circ$ para $2,2^\circ$ com janela de integração de 5,12 s sob razão sinal-ruído (*Signal-to-Noise Ratio*, SNR) de -14 dB.

Leandro Kfuri de Oliveira, kfuri.naval@hotmail.com;
Rigel P. Fernandes, rigelfernandes@gmail.com; José M. de Seixas, seixas@lps.ufrj.br; Allyson Andrade da Silva, allyson.andrade@marinha.mil.br.

A classificação automática de sonar passivo enfrenta dois desafios: (i) a baixa SNR imposta pelo ambiente oceânico [3]; e (ii) a variabilidade das assinaturas acústicas com parâmetros operacionais como rotação do eixo, carga e velocidade [4]. Técnicas de aprendizado profundo têm demonstrado capacidade de reconhecimento de padrões nesses cenários [5], [6].

Silva et al. [7] publicaram o IARA (*an Acoustical Recordings Archive*), o maior *dataset* público de acústica subaquática brasileira, com 1.825 gravações da Baía de Santos. Este trabalho utiliza um subconjunto de 1.257 gravações. Os *baselines* (Random Forest, MLP, CNN) usam o *Pipeline A* — vetores espectrais 1D —, com o MLP alcançando 67,48% de acurácia balanceada e F1 de 66,72% via 5×2 *cross-validation* com *grid search*. Este trabalho adota o *Pipeline B*, com espectrogramas Mel 2D que preservam a estrutura espectral-temporal, sem *grid search*.

Três aspectos inéditos são explorados: (1) arquiteturas modernas com o *Pipeline B*; (2) desempenho sob ruído gaussiano sintético e ruído real da Baía de Santos (DC E); e (3) generalização geográfica entre IARA e ShipsEar [8].

O restante deste trabalho está organizado da seguinte forma: a Seção II revisa os trabalhos relacionados; a Seção III descreve o método, incluindo o *dataset*, o pré-processamento e os protocolos de avaliação; a Seção IV apresenta os resultados nos três cenários avaliados; e a Seção V conclui com recomendações para o SisGAAz.

II. TRABALHOS RELACIONADOS

A. Bases de dados públicas

A escassez de bases públicas rotuladas é reconhecida como um dos principais entraves à reprodutibilidade em reconhecimento de alvos acústicos subaquáticos (*Underwater Acoustic Target Recognition* — UATR) [6]. Santos-Domínguez et al. [8] introduziram o ShipsEar, com 90 gravações de 11 tipos de embarcações coletadas em águas rasas próximas ao porto de Vigo, Espanha, entre 2012 e 2013, tornando-se a referência internacional mais utilizada para o desenvolvimento de classificadores. Irfan et al. [9] apresentaram o DeepShip, com 613 gravações totalizando aproximadamente 47 h, coletadas em águas profundas próximas a Vancouver, Canadá, com seleção de janelas baseada em dados do sistema de identificação automática (AIS) e raio de exclusão de 2 km. As duas bases apresentam características complementares: o ShipsEar oferece maior diversidade de tipos de embarcação

em águas rasas, enquanto o DeepShip fornece maior volume de dados em profundidades superiores, porém sem gravações de ruído ambiente.

Silva et al. [7] estabelecem o IARA como referência brasileira, com 129 h e 34 min de áudio em 1.825 gravações coletadas na Baía de Santos. O trabalho avalia Random Forest, MLP e CNN sob 5×2 *cross-validation* com *grid search* extensivo, e conduz experimentos específicos sobre o impacto da distância ao ponto de maior aproximação, da localização geográfica e da profundidade de instalação. A base inclui gravações de ruído de fundo, o que viabiliza os experimentos de robustez com ruído oceânico real conduzidos no presente trabalho.

B. Extração de características

A abordagem predominante em UATR combina um extrator de características com um classificador não linear [10], [6]. Entre as representações mais empregadas estão o espectro de potência obtido por transformada de Fourier de curta duração, o espectrograma Mel e a análise de baixa frequência com normalização por janela deslizante em dois passos (LOFAR). Yin et al. [11] demonstraram a eficácia do espectro LOFAR combinado com redes convolucionais (VGGNet adaptada), reportando ainda que a degradação sob ruído oceânico real gravado é substancialmente maior do que sob ruído gaussiano sintético de SNR nominal equivalente — padrão consistente com o observado neste trabalho (Seção IV) e que reforça a insuficiência de avaliações baseadas exclusivamente em ruído sintético. Ke et al. [12] propuseram um algoritmo de separação supervisionada de características para cenários com poucos dados rotulados. Fernandes e Apolinário [13] exploraram a classificação de alvos subaquáticos no ShipsEar com seleção otimizada de características por algoritmos genéticos, obtendo ganho de 13 pontos percentuais em relação ao uso do conjunto completo de atributos e evidenciando a importância da engenharia de características para esse domínio. Yang et al. [14] propuseram a concatenação de sub-bandas do espectrograma Mel associada a mecanismos de atenção multidomínio, indicando que a escolha da representação tempo-frequência permanece um fator determinante de desempenho.

C. Arquiteturas profundas

Tang et al. [5] incorporam modelagem temporal 2D a arquiteturas convolucionais consolidadas (SE ResNet e MSRDN), avaliando o DeepShip; representações tempo-frequência superam entradas no domínio do tempo em 1–2 pontos percentuais, ao custo de maior número de parâmetros. Liu et al. [15] propõem uma CRNN com blocos convolucionais e LSTM empilhada para capturar dependências temporais em espectrogramas Mel tridimensionais, associada a aumento de dados.

Com a consolidação do Transformer [16] e sua extensão a dados bidimensionais via *patch embedding* [17], mecanismos de autoatenção passaram a ser aplicados ao processamento de sinais acústicos subaquáticos. Wang et al. [18] propõem uma rede de fusão espaço-temporal combinando Transformer com Convolução Separável em Profundidade (DWC), aplicada à classificação de modulação em comunicação acústica — domínio distinto do reconhecimento de ruído irradiado. A

adaptação dessa estratégia a espectrogramas de ruído naval, conduzida neste trabalho, constitui transposição entre domínios ainda não caracterizada. Feng e Zhu [19] aplicam Transformer ao UATR, e Ren et al. [20] propõem um *front-end* aprendível. Xiong et al. [21] associam a posição da normalização à estabilidade do gradiente em sequências longas, questão retomada na Seção III. O Conformer [22], originalmente para reconhecimento de fala, integra convoluções locais e atenção global no mesmo bloco.

D. Trabalhos com dados nacionais

Filho et al. [10] abordam o pré-processamento de sinais de sonar passivo para classificação neural. Fernandes et al. [23] demonstraram viabilidade de aprendizado profundo com dados militares reais da Marinha do Brasil, com resultados competitivos sob baixa SNR. Hummel et al. [24] revisam métodos de aprendizado de máquina para ruído irradiado por navios, destacando a diversidade de treino como determinante da generalização — argumento relacionado aos experimentos de transferência geográfica deste trabalho.

III. MÉTODO

Esta seção descreve o *dataset*, o pré-processamento, as arquiteturas e os protocolos experimentais. Os modelos são avaliados em três cenários complementares: (i) **condições ideais**, nas quais os segmentos do IARA são utilizados sem qualquer adição de ruído externo; (ii) **desempenho sob ruído**, no qual ruído gaussiano sintético ou ruído de fundo real da campanha DC E é somado aos espectrogramas em níveis controlados de SNR; e (iii) **generalização geográfica**, na qual modelos treinados em uma base são avaliados na outra. O código utilizado neste trabalho está disponível em [25].

A. Dataset IARA e Divisão dos Dados

O IARA [7] contém gravações coletadas por observatórios acústicos subaquáticos instalados no leito marinho a 19–28 m de profundidade na Baía de Santos, no âmbito do Projeto de Monitoramento da Paisagem Acústica Submarina da Baía de Santos (PMPAS-BS), conduzido pela Petrobras.

As gravações foram organizadas em *Deployment Campaigns* (DCs), conjuntos de dados coletados sob condições similares de processamento e seleção. O parâmetro CPA (*Closest Point of Approach*) indica a menor distância horizontal entre a embarcação e o hidrofone ao longo de toda a passagem do navio: quanto menor o CPA, mais próxima a embarcação esteve do sensor e, em geral, maior a SNR do sinal registrado. Para cada gravação, uma zona de inclusão define a faixa de distância em que a embarcação deve estar presente, e uma zona de exclusão de 2 km garante que nenhum outro navio contamine o registro.

Foram utilizadas as campanhas DC A (456 gravações, com passagem pelo CPA (*Closest Point of Approach*, ou ponto de maior aproximação) a menos de 250 m do hidrofone), DC B (199 gravações, embarcação a menos de 250 m mas sem passagem pelo ponto de maior proximidade, o que não garante SNR equivalente à DC A), DC C (341 gravações, CPA entre 250–500 m) e DC E (261 gravações de ruído de fundo, sem embarcação na zona de exclusão de 10 km), totalizando 1.257 gravações.

Espectrogramas Mel por Classe — Dataset IARA
(Pipeline B, SR=16kHz, N_Mels=128)

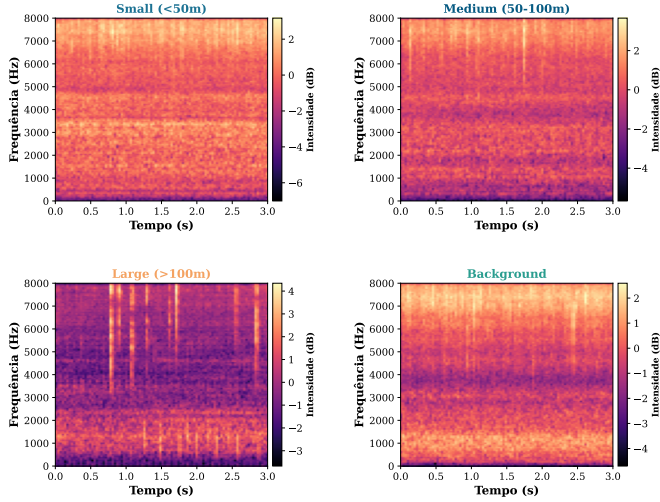


Fig. 1. Espectrogramas Mel das quatro classes do dataset IARA, apresentados com escala de intensidade comum (percentis 2–98) para permitir comparação visual direta. *Small* (<50 m): linhas tonais esparsas de banda estreita; *Medium* (50–100 m) e *Large* (>100 m): harmônicos densos com energia crescente em baixas frequências; *Background*: ruído oceânico estocástico de baixa frequência (DC E). Eixo X: tempo (s); Eixo Y: frequência (Hz).

As quatro classes são definidas pelo comprimento do navio, seguindo a categorização de Snapir et al. [26]: *Small* (<50 m), *Medium* (50–100 m), *Large* (>100 m) e *Background* (ruído oceânico ambiental sem embarcação). Essa abordagem por comprimento foi adotada porque o tipo geral de embarcação não é um indicador confiável de porte na diversidade de navios presente no IARA [7].

As classes apresentam assinaturas acústicas distintas, visíveis nos espectrogramas Mel da Fig. 1, na qual cores quentes indicam alta energia e cores frias, baixa energia. Embarcações *Small* produzem espectrogramas com energia distribuída em bandas horizontais contínuas e poucas linhas tonais tênues, compatíveis com maquinaria de pequeno porte. Embarcações *Medium* e *Large* apresentam coloração progressivamente mais quente nas baixas frequências e, no caso das *Large*, linhas verticais brilhantes compatíveis com pulsos periódicos de maquinaria pesada e cavitação do hélice. A classe *Background* exhibe padrão homogêneo e de baixa intensidade, sem estruturas tonais, representando o ruído oceânico ambiental difuso da Baía de Santos.

A partição dos dados utilizou o identificador da gravação como chave de agrupamento, de modo que todos os segmentos extraídos de um mesmo arquivo permanecessem na mesma partição, impedindo que janelas sobrepostas do mesmo sinal aparecessem simultaneamente em treino e teste (*data leakage*). Reservaram-se 20% das gravações (252 arquivos) como conjunto de teste isolado, mantido inacessível durante todo o treinamento, e os 80% restantes (1.005 arquivos) foram distribuídos em *GroupKFold* com $K = 5$. A partição de teste, estratificada pela classe majoritária de cada gravação, apresenta, por gravação, *Small* (22,8%), *Medium* (19,2%), *Large* (32,6%) e *Background* (25,4%). O *GroupKFold* convencional, por sua vez, não assegura, por si só, a preservação dessas proporções entre os cinco *folds* internos de validação cruzada. Como as gravações de *Background* (DC E) são sistematicamente mais

curtas que as de embarcação, a distribuição por *segmento* — unidade sobre a qual a avaliação é de fato conduzida — difere da distribuição por gravação: *Small* (27,2%), *Medium* (22,1%), *Large* (38,8%) e *Background* (11,9%), correspondendo a um limiar de chance ponderado de 28,8%, valor utilizado como referência ao longo deste artigo.

Ressalta-se que os *baselines* de Silva et al. [7] foram obtidos sob protocolo distinto (5×2 *cross-validation* com *grid search*, métrica de acurácia balanceada). As diferenças de acurácia entre os dois trabalhos não podem, portanto, ser atribuídas exclusivamente às arquiteturas avaliadas, e as comparações apresentadas ao longo deste artigo devem ser interpretadas como referência de ordem de grandeza.

B. Pré-Processamento — Pipeline B

Os arquivos .wav foram reamostrados para 16 kHz, submetidos a pré-ênfase ($\alpha = 0,97$) e segmentados em janelas de 3 s com 50% de sobreposição. A pré-ênfase foi mantida para destacar harmônicos de alta frequência do sinal. Cada segmento foi convertido em espectrograma Mel com 128 bandas, FFT de 1024 pontos e salto de 512 amostras, resultando em tensores [1, 128, 94]. A escala Mel foi adotada por comprimir o envelope espectral e destacar harmônicos de banda estreita de maquinaria naval. A normalização *Z-score* por segmento remove a amplitude absoluta, tornando o sistema invariante a variações de ganho ou distância, totalizando 217.258 segmentos.

C. Arquiteturas Avaliadas

A CNN Baseline, com três blocos de convolução 2D, *BatchNorm*, ReLU e *MaxPooling* ($\approx 127k$ parâmetros), serve como referência mínima de desempenho neste trabalho. A CRNN estende essa base com duas camadas LSTM unidirecionais empilhadas ($\approx 3,82M$ parâmetros), consistente com a configuração de Liu et al. [15] descrita na Seção II.

O Transformer-DWC divide o espectrograma Mel em segmentos espectrais por Convolução Separável em Profundidade (DWC), processados por quatro blocos Transformer de oito cabeças de atenção ($\approx 566k$ parâmetros). Diferente da formulação original de Wang et al. [18] (Seção II), esta implementação usa *Post-LN* em vez de *Pre-LN* e não emprega *positional embedding* — adaptações decorrentes da transposição ao domínio de ruído naval, retomadas na discussão dos resultados.

O Conformer [22], proposto originalmente para reconhecimento de fala, processa o sinal em cada bloco por meio de quatro operações encadeadas: camadas *feed-forward* (FF) nas extremidades filtram e reescalam as representações; a atenção multi-cabeça (MHSA) captura dependências de longo alcance entre regiões do espectrograma; e a convolução *depthwise* 1D de *kernel* 31 extrai padrões locais entre instantes adjacentes no eixo temporal (FF→MHSA→DWC 1D→FF, $\approx 1,56M$ parâmetros).

Todas as arquiteturas foram treinadas sob configuração idêntica, de modo a isolar o efeito estrutural de cada modelo: função de perda de entropia cruzada não ponderada, otimizador Adam com taxa de aprendizado de 1×10^{-3} e *weight decay* de 1×10^{-4} , agendador *CosineAnnealingLR* ($T_{\max}$ = número de épocas), *batch* de 64 amostras, *dropout* de 0,3 e 30 épocas, sem busca em grade de hiperparâmetros e com

semente fixa ($seed = 42$). Os pesos correspondentes à melhor época de validação de cada *fold* foram retidos para avaliação.

D. Protocolos de Avaliação

Para avaliar a resistência dos modelos ao ruído, dois tipos de interferência foram considerados: gaussiano sintético e oceânico real (DC E). Em ambos os casos, o nível de SNR é controlado pelo fator α :

$$\alpha = \sqrt{\frac{P_{\text{sinal}}}{P_{\text{ruído}} \cdot 10^{\text{SNR}/10}}}, \quad P = \frac{1}{N} \sum_{i=1}^N x_i^2 \quad (1)$$

onde P é a potência média do tensor do espectrograma Mel (*feature-level SNR*), e o sinal resultante é:

$$\tilde{x} = x_{\text{sinal}} + \alpha \cdot x_{\text{ruído}} \quad (2)$$

Para o ruído gaussiano sintético, $x_{\text{ruído}}$ é amostrado de $\mathcal{N}(0, 1)$, garantindo distribuição uniforme em todas as frequências. Para o ruído oceânico real, segmentos aleatórios da campanha DC E são utilizados como $x_{\text{ruído}}$, capturando as características espectrais efetivas da Bacia de Santos. A combinação é realizada por adição direta dos espectrogramas no domínio da frequência. Trata-se de uma aproximação computacionalmente eficiente, que desconsidera a fase dos sinais e corresponde a uma soma incoerente de potências; a validação da mistura no domínio temporal constitui etapa futura para homologação operacional. Os níveis avaliados foram: limpo, 20, 10, 5, 0 e -5 dB.

Nos testes, considera-se acerto exclusivamente quando o modelo identifica a classe original do navio (*Small*, *Medium* ou *Large*). Classificar um segmento de navio como *Background* constitui um Falso Negativo, situação em que o sistema reporta ruído onde há um navio real. Para o SisGAaz, este é o erro operacionalmente mais crítico, pois corresponde a uma embarcação que atravessa a área monitorada sem ser detectada.

Para a generalização geográfica, foram consideradas duas direções de transferência. Na direção IARA $\rightarrow$ ShipsEar, os modelos treinados no IARA (4 classes) foram avaliados em dados do ShipsEar [8]. Como o ShipsEar organiza as gravações em 11 categorias por tipo de embarcação, estas foram mapeadas para as 4 classes do IARA por comprimento estimado: *Small* (A, F, G, I); *Medium* (B, C, E); *Large* (D, H, J); *Background* (NO).

Na direção inversa ShipsEar $\rightarrow$ IARA, modelos treinados no ShipsEar foram avaliados em dados do IARA. Como o modelo original gera 11 pontuações de saída (*logits*), a pontuação de cada grupo do IARA foi definida pela máxima pontuação entre as subclasses correspondentes do ShipsEar. Ressalta-se que essa direção foi avaliada apenas para a CNN Baseline e o Transformer-DWC; a avaliação bidirecional é, portanto, parcial, e as comparações entre as duas direções devem ser interpretadas com essa restrição.

IV. RESULTADOS

A. Condições Ideais

A Tabela I apresenta os resultados no conjunto de teste isolado (252 gravações), obtidos sem *grid search*. A CRNN alcançou a maior acurácia entre as arquiteturas avaliadas

(66,14%). Os valores de Silva et al. [7] são reproduzidos apenas como referência de ordem de grandeza: além do protocolo experimental distinto, correspondem a acurácia balanceada, enquanto os resultados deste trabalho correspondem a acurácia simples, métricas que divergem sistematicamente sob distribuição não uniforme de classes. Entre as arquiteturas aqui avaliadas, CRNN e Conformer diferem menos de 1 p.p., com desvios padrão sobrepostos; ambas devem ser consideradas equivalentes em condições ideais, sendo a escolha determinada pelo desempenho nos cenários seguintes.

TABELA I. ACURÁCIA NO CONJUNTO DE TESTE ISOLADO — PIPELINE B

Modelo	ACC (%)	Std (%)
<i>Referência — protocolo distinto[†]</i>		
MLP — Silva et al. [7]	67,48	—
CNN — Silva et al. [7]	64,99	—
<i>Este trabalho — Pipeline B</i>		
CRNN	66,14	$\pm 1,17$
Conformer	65,23	$\pm 0,53$
Transformer-DWC	63,84	$\pm 1,04$
CNN Baseline	63,18	$\pm 1,24$

B. Desempenho sob Ruído

A Tabela II apresenta os resultados por nível de SNR (Fig. 2); a coluna “Limpo” reproduz a Tabela I sob protocolo idêntico de 5 *folds*, para referência da degradação relativa.

TABELA II. ACURÁCIA (%) POR NÍVEL DE SNR

Modelo	Lim.	Gaussiano			Real DC E		
		5dB	0dB	-5dB	5dB	0dB	-5dB
CNN	63,18	49,86	36,22	28,73	51,54	39,18	29,42
	$\pm 1,24$	$\pm 4,06$	$\pm 9,74$	$\pm 6,29$	$\pm 1,88$	$\pm 3,96$	$\pm 3,08$
CRNN	66,14	57,34	45,47	39,10	58,25	46,39	32,52
	$\pm 1,17$	$\pm 1,92$	$\pm 2,52$	$\pm 1,07$	$\pm 0,97$	$\pm 0,50$	$\pm 2,64$
DWC	63,84	50,64	39,42	31,89	54,15	43,45	33,69
	$\pm 1,04$	$\pm 1,93$	$\pm 8,20$	$\pm 10,77$	$\pm 1,59$	$\pm 3,05$	$\pm 5,97$
Conf.	65,23	63,04	56,54	45,79	56,50	44,91	29,98
	$\pm 0,53$	$\pm 0,88$	$\pm 1,37$	$\pm 1,86$	$\pm 0,63$	$\pm 1,83$	$\pm 2,63$
Média $\pm$ desvio sobre 5 <i>folds</i> . Limiar de chance ponderado pela distribuição de segmentos: 28,8%.							

Sob ruído gaussiano, o Conformer apresentou a menor degradação: $-2,19$ p.p. em 5 dB (65,23% $\rightarrow$ 63,04%), contra $-13,32$ p.p. da CNN. Em -5 dB atinge 45,79% $\pm 1,86$, contra 39,10% $\pm 1,07$ da CRNN, sem sobreposição de intervalos.

Sob ruído oceânico real (DC E) em -5 dB o quadro se inverte: todas as arquiteturas superam o limiar de chance ponderado (28,8%) com acurácias estatisticamente indistinguíveis — CNN 29,42% $\pm 3,08$, Conformer 29,98% $\pm 2,63$, CRNN 32,52% $\pm 2,64$, Transformer-DWC 33,69% $\pm 5,97$ —, e a acurácia, isoladamente, não discrimina os modelos.

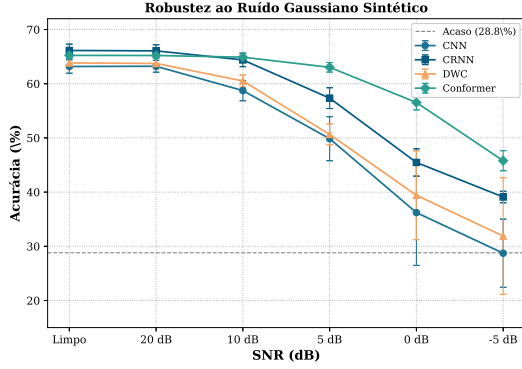
A diferença manifesta-se na estrutura do erro (Tabela III, taxa de Falso Negativo: fração de segmentos de embarcação classificados como *Background*). Em -5 dB o Conformer atinge 40,52% $\pm 6,53$, contra 15,61% $\pm 14,66$ do Transformer-DWC e 7,24% $\pm 5,40$ da CNN; os intervalos do Conformer ([34,0; 47,1]) e do Transformer-DWC ([0,9; 30,3]) não se sobrepõem. Para o SisGAaz a distinção é decisiva: confundir *Medium* com *Large* ainda sinaliza a embarcação; classificá-la como ruído de fundo não. [auditoria]

A inversão de comportamento entre os dois tipos de ruído — a arquitetura de melhor desempenho sob gaussiano é a de maior Falso Negativo sob ruído real — é o achado mais relevante deste cenário. A decomposição em -5 dB é

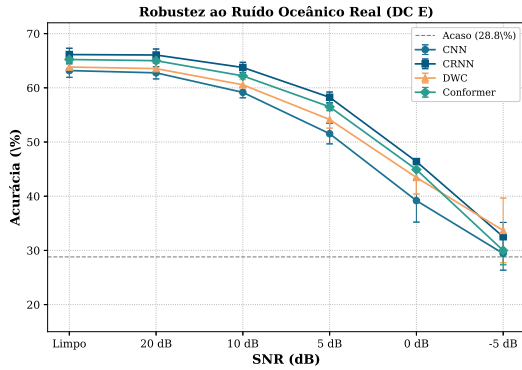
TABELA III. TAXA DE FALSO NEGATIVO (%) — EMBARCAÇÃO CLASSIFICADA COMO *BACKGROUND* [auditoria]

Modelo	Lim.	Gaussiano			Real DC E		
		5dB	0dB	-5dB	5dB	0dB	-5dB
CNN	1,60	1,70	4,63	0,80	2,69	4,90	7,24
		$\pm 0,55$	$\pm 2,39$	$\pm 7,83$	$\pm 1,44$	$\pm 1,26$	$\pm 3,08$
CRNN	0,80	0,49	0,14	0,00	4,24	13,68	28,34
		$\pm 0,22$	$\pm 0,28$	$\pm 0,19$	$\pm 0,00$	$\pm 0,44$	$\pm 1,93$
DWC	0,99	0,61	12,00	20,10	3,07	5,72	15,61
		$\pm 0,39$	$\pm 1,07$	$\pm 23,71$	$\pm 39,93$	$\pm 0,76$	$\pm 2,43$
Conf.	1,45	0,61	0,17	0,02	6,38	18,12	40,52
		$\pm 0,25$	$\pm 0,15$	$\pm 0,14$	$\pm 0,03$	$\pm 1,22$	$\pm 3,26$

Média $\pm$ desvio sobre 5 *folds*. Menor é melhor.



(a) Ruído gaussiano sintético



(b) Ruído oceânico real (DC E)

Fig. 2. Acurácia (%) por nível de SNR, com associação cor-modelo idêntica nos dois painéis. (a) Sob ruído gaussiano, o Conformer mantém desempenho estável até 5 dB. (b) Sob ruído oceânico real DC E, todas as arquiteturas superam o limiar de chance ponderado (28,8%, linha tracejada) em -5 dB, com acurácias estatisticamente indistinguíveis; a diferença manifesta-se na taxa de Falso Negativo (Tabela III), não na acurácia.

compatível com essa inversão: no Conformer, 40,5% dos erros convergem para *Background* contra 35,2% de confusão entre portes — único modelo em que a convergência predomina; no Transformer-DWC a proporção inverte-se (15,6% contra 50,5%), assim como na CRNN (28,3% contra 42,6%) e na CNN (7,2% contra 61,8%). Uma hipótese compatível envolve o *Patch Embedding*: no Transformer-DWC ele agrupa regiões do espectrograma antes de processá-las, o que poderia atenuar ruído estocástico de baixa frequência; no Conformer, a convolução *depthwise* 1D percorre o eixo temporal sem esse agrupamento, podendo propagá-lo. Trata-se, contudo, de hipótese: os quatro modelos diferem simultaneamente em capacidade, tipo de convolução, blocos *feed-forward*, normalização e codificação posicional, de modo que os resultados indicam tendência, não mecanismo isolado — confirmação exigiria estudo de ablação e análise de saliência.

O Transformer-DWC, embora não apresente o maior desvio

padrão em condições ideais ($\pm 1,04\%$, atrás da CNN Baseline com $\pm 1,24\%$, Tabela I), passa a apresentar a maior dispersão sob ruído: $\pm 10,77$ p.p. em -5 dB gaussiano e $\pm 14,66$ p.p. de Falso Negativo no mesmo nível real. Isso é compatível com a instabilidade de gradiente do *Post-LN* descrita por [21] — embora aquele trabalho caracterize instabilidade no início do treinamento, não a variância entre execuções independentes. Verificar se *Pre-LN* e *positional embedding*, conforme [18], reduzem essa variância sem comprometer o desempenho sob ruído real constitui investigação direta.

Independentemente do mecanismo, a divergência entre os cenários sustenta uma conclusão metodológica: testes exclusivamente com ruído gaussiano são insuficientes para caracterizar classificadores acústicos subaquáticos operacionalmente.

C. Generalização Geográfica

A Tabela IV e a Fig. 3 apresentam os resultados de transferência entre bases. O Conformer obteve 41,29% na direção IARA→ShipsEar, superando em 12,87 p.p. a CNN Baseline avaliada neste trabalho (28,42%) e em 10,06 p.p. o *baseline* CNN reportado no trabalho original do IARA ($31,23 \pm 10,97\%$) [7]. Esta última comparação deve ser lida com cautela: além da diferença de protocolo já discutida, o intervalo de confiança reportado no trabalho original é amplo, de modo que a diferença não é estatisticamente conclusiva. Todos os modelos superaram o limiar de chance de 25% nesta direção.

As assinaturas da Bacia de Santos e da Ria de Vigo diferem em temperatura, salinidade e composição da frota, o que explica a transferência apenas parcial observada em todos os modelos. Entre as arquiteturas avaliadas, aquelas que incorporam mecanismos de autoatenção apresentaram os melhores resultados de transferência, o que sugere maior transferibilidade das representações aprendidas; a atribuição desse ganho a um componente arquitetural específico exigiria ablação controlada.

TABELA IV. GENERALIZAÇÃO GEOGRÁFICA — ACURÁCIA (%)

Modelo	IARA→ShipsEar	ShipsEar→IARA
CNN Baseline	28,42	26,63
CRNN	31,58	n.a.*
Transformer-DWC	34,74	31,43
Conformer	41,29	n.a.*
Chance†	25,00	25,00

* n.a.: não avaliado. A direção inversa foi conduzida apenas para dois modelos; constitui trabalho futuro.

† Limiar uniforme de 4 classes. Difere do limiar ponderado de 28,8% da Tabela II, calculado sobre as proporções [auditoria] de classe do IARA, que não se aplicam ao ShipsEar.

V. CONCLUSÕES

Este artigo avaliou sistematicamente CNN, CRNN, Transformer-DWC e Conformer no *dataset* IARA, com três contribuições: (1) o Conformer apresentou a menor degradação sob ruído gaussiano ($-2,19$ p.p. em 5 dB; $45,79\% \pm 1,86$ em -5 dB); (2) sob ruído oceânico real em -5 dB, a degradação não se manifesta na acurácia — estatisticamente indistinguível entre as quatro arquiteturas — mas na taxa de Falso Negativo, que varia de 15,61% (Transformer-DWC) a 40,52% (Conformer), sem sobreposição de intervalos; e (3) o Conformer alcançou 41,29% na transferência

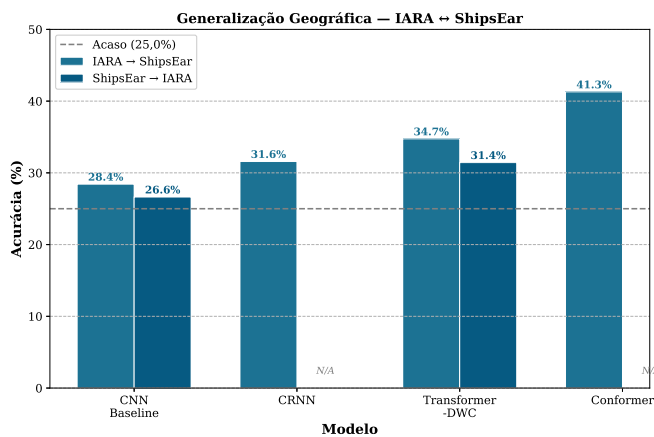


Fig. 3. Acurácia de generalização geográfica por modelo nas direções IARA→ShipsEar e ShipsEar→IARA. A ausência de barra na direção inversa para CRNN e Conformer indica caso não avaliado (n.a.), e não desempenho nulo. O Conformer apresenta o maior valor em IARA→ShipsEar (41,29%) entre as arquiteturas comparadas. [2.5]

IARA→ShipsEar, 10,06 p.p. acima do *baseline* original do IARA, ressaltadas as diferenças de protocolo. A direção ShipsEar→IARA foi avaliada apenas para dois modelos, sendo parcial.

O Transformer-DWC ($\approx 566k$ parâmetros) é o candidato mais promissor para avaliação embarcada nos nós do Sis-GAAz, por combinar a menor taxa de Falso Negativo entre os modelos de atenção sob ruído real com o segundo menor número de parâmetros — um sétimo da CRNN ($\approx 3,82M$), líder em condições ideais. Essa indicação apoia-se exclusivamente em acurácia e parâmetros; latência, memória e desempenho em plataforma-alvo não foram avaliados. O Conformer ($\approx 1,56M$ parâmetros) é adequado a ruído moderado e transferência entre regiões costeiras, ressaltada sua elevada taxa de Falso Negativo sob ruído real, que desaconselha seu uso isolado em vigilância persistente.

Como trabalhos futuros: (i) avaliação com LOFAR e 2-pass *split window*; (ii) completar a avaliação bidirecional de transferência geográfica; e (iii) validação com dados do Instituto de Pesquisas da Marinha (IPqM).

REFERÊNCIAS

- [1] B. G. Storch, R. P. Fernandes, A. A. da Silva, “Estudo comparativo de algoritmos aplicados à navegação de veículos de superfície não tripulados,” in *XXVII Simpósio de Aplicações Operacionais em Áreas de Defesa (SIGE)*, São José dos Campos, SP, Brasil, Sep. 2025.
- [2] R. P. Rocha, R. P. Fernandes, J. A. A. Jr., N. N. de Moura Junior, “Analysis of a regular tetrahedral array for underwater DOA estimation,” in *XLIV Brazilian Symposium on Telecommunications and Signal Processing (SBTrT 2026)*, Salvador, BA, Brasil, Sep. 2026, accepted for publication.
- [3] G. M. Wenz, “Acoustic ambient noise in the ocean: Spectra and sources,” *The Journal of the Acoustical Society of America*, vol. 34, no. 12, pp. 1936–1956, 1962. [Online]. Available: <https://doi.org/10.1121/1.1909155>
- [4] R. J. Urlick, *Principles of Underwater Sound*, 3rd ed. New York, NY, USA: McGraw-Hill, 1983.
- [5] J. Tang, W. Gao, E. Ma, X. Sun, J. Ma, “Deep learning based underwater acoustic target recognition: Introduce a recent temporal 2D modeling method,” *Sensors*, vol. 24, no. 5, p. 1633, 2024. [Online]. Available: <https://doi.org/10.3390/s24051633>
- [6] L. C. F. Domingos, P. E. Santos, P. S. M. Skelton, R. S. A. Brinkworth, K. Sammut, “A survey of underwater acoustic data classification methods using deep learning for shoreline surveillance,” *Sensors*, vol. 22, no. 6, p. 2181, 2022. [Online]. Available: <https://doi.org/10.3390/s22062181>

- [7] F. O. B. da Silva, J. de Castro Vargas Fernandes, W. S. Filho, J. ao Baptista de Oliveira e Souza Filho, A. Spengler, N. N. de Moura Júnior, “IARA: An underwater acoustic database,” *IEEE Access*, vol. 13, pp. 120 293–120 306, 2025. [Online]. Available: <https://doi.org/10.1109/ACCESS.2025.3585467>
- [8] D. Santos-Domínguez, S. Torres-Guijarro, A. Cardenal-López, A. Pena-Gimenez, “ShipsEar: An underwater vessel noise database,” *Applied Acoustics*, vol. 113, pp. 64–69, 2016. [Online]. Available: <https://doi.org/10.1016/j.apacoust.2016.06.008>
- [9] M. Irfan, Z. Jiangbin, S. Ali, M. Iqbal, Z. Masood, 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. [Online]. Available: <https://doi.org/10.1016/j.eswa.2021.115270>
- [10] W. S. Filho, J. M. de Seixas, N. N. de Moura, “Preprocessing passive sonar signals for neural classification,” *IET Radar, Sonar & Navigation*, vol. 5, no. 6, pp. 605–612, 2011. [Online]. Available: <https://doi.org/10.1049/iet-rsn.2010.0157>
- [11] X. Yin, X. Sun, P. Liu, L. Wang, R. Tang, “Underwater acoustic target classification based on LOFAR spectrum and convolutional neural network,” in *Proc. 2nd Int. Conf. on Artificial Intelligence and Advanced Manufacture*, 2020, pp. 59–63. [Online]. Available: <https://doi.org/10.1145/3421766.3421890>
- [12] X. Ke, F. Yuan, E. Cheng, “Underwater acoustic target recognition based on supervised feature-separation algorithm,” *Sensors*, vol. 18, no. 12, p. 4318, 2018. [Online]. Available: <https://doi.org/10.3390/s18124318>
- [13] R. P. Fernandes J. A. A. Jr., “Underwater target classification with optimized feature selection based on genetic algorithms,” in *XXXVIII Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBTrT 2020)*, Florianópolis, SC, Brasil, Nov. 2020. [Online]. Available: <http://doi.org/10.14209/SBTrT.2020.1570658075>
- [14] S. Yang, A. Jin, X. Zeng, H. Wang, X. Hong, M. Lei, “Underwater acoustic target recognition based on sub-band concatenated mel spectrogram and multidomain attention mechanism,” *Engineering Applications of Artificial Intelligence*, vol. 133, p. 107983, 2024. [Online]. Available: <https://doi.org/10.1016/j.engappai.2024.107983>
- [15] F. Liu, T. Shen, Z. Luo, D. Zhao, S. Guo, “Underwater target recognition using convolutional recurrent neural networks with 3-D mel-spectrogram and data augmentation,” *Applied Acoustics*, vol. 178, p. 107989, 2021. [Online]. Available: <https://doi.org/10.1016/j.apacoust.2021.107989>
- [16] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems 30*, 2017, pp. 5998–6008. [Online]. Available: <https://doi.org/10.48550/arXiv.1706.03762>
- [17] A. Dosovitskiy *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2021. [Online]. Available: <https://doi.org/10.48550/arXiv.2010.11929>
- [18] Y. Wang, J. Xiao, X. Cheng, Q. Wei, N. Tang, “Underwater acoustic signal classification based on a spatial-temporal fusion neural network,” *Frontiers in Marine Science*, vol. 11, p. 1331717, 2024. [Online]. Available: <https://doi.org/10.3389/fmars.2024.1331717>
- [19] S. Feng X. Zhu, “A transformer-based deep learning network for underwater acoustic target recognition,” *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022. [Online]. Available: <https://doi.org/10.1109/LGRS.2022.3201396>
- [20] J. Ren, Y. Xie, X. Zhang, J. Xu, “UALF: A learnable front-end for intelligent underwater acoustic classification system,” *Ocean Engineering*, vol. 264, p. 112394, 2022. [Online]. Available: <https://doi.org/10.1016/j.oceaneng.2022.112394>
- [21] R. Xiong *et al.*, “On layer normalization in the transformer architecture,” in *Proc. 37th Int. Conf. on Machine Learning*, 2020, pp. 10 524–10 533. [Online]. Available: <https://doi.org/10.48550/arXiv.2002.04745>
- [22] A. Gulati *et al.*, “Conformer: Convolution-augmented Transformer for speech recognition,” in *Proc. Interspeech 2020*, 2020, pp. 5036–5040. [Online]. Available: <https://doi.org/10.21437/Interspeech.2020-3015>
- [23] J. C. V. Fernandes, N. N. de Moura Junior, J. M. de Seixas, “Deep learning models for passive sonar signal classification of military data,” *Remote Sensing*, vol. 14, no. 11, p. 2648, 2022. [Online]. Available: <https://doi.org/10.3390/rs14112648>
- [24] H. I. Hummel, R. van der Mei, S. Bhulai, “A survey on machine learning in ship radiated noise,” *Ocean Engineering*, vol. 298, p. 117252, 2024. [Online]. Available: <https://doi.org/10.1016/j.oceaneng.2024.117252>
- [25] L. K. de Oliveira, “Classificação acústica subaquática com redes neurais profundas — código-fonte,” 2026. [Online]. Available: <https://github.com/kfurinaival-bit/acoustic-classification-sisgaaz>
- [26] B. Snapir, T. W. Waive, L. Biermann, “Maritime vessel classification to monitor fisheries with SAR: Demonstration in the north sea,” *Remote Sensing*, vol. 11, no. 3, p. 353, 2019. [Online]. Available: <https://doi.org/10.3390/rs11030353>