JASA ARTICLE

An explainable context-adaptive fusion and expert-in-the-loop evaluation framework for underwater sonar image classification

Kamal Basha S,[1,a)] Anukul Kiran B,[1,b)] Athira Nambiar,[1,c)] Suresh Rajendran,[2,d)] and Sooraj K Ambat[3,e)]

[1]*Department of Computational Intelligence, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu 603203, India*

[2]*Department of Ocean Engineering, Indian Institute of Technology Madras, Chennai, Tamil Nadu 600036, India*

[3]*Naval Physical and Oceanographic Laboratory, Defence Research and Development Organisation, Kochi, Kerala 682021 India*

ABSTRACT:
Sonar image interpretation is critical for identifying submerged objects in underwater exploration. However, conventional deep learning models often fail to capture sonar-specific features such as acoustic shadows and highlights, and they typically operate as black boxes, limiting both performance and trust. To address these challenges, we propose an explainable sonar image classification system that fuses specialized classifiers for shadows, highlights, and general features through a context-adaptive fusion mechanism. The system further incorporates expert-in-the-loop evaluation via the Augmented QUality Assessment for eXplainability (AQUA-X) framework to ensure interpretability and trust. The contributions of this study are threefold: first, the development of a novel Context-Adaptive Fusion Framework (CAFF) that integrates Naive, ShadowNet, and HighlightNet classifiers using an attention-based fusion mechanism; second, the incorporation of Gradient-weighted Class Activation Mapping, SHapley Additive exPlanations, and Local Interpretable Model-agnostic Explanations to provide detailed interpretability of sonar-specific features; and third, expert validation through AQUA-X, enabling comparative assessment and refinement of explanations and identification of the most suitable explainable artificial intelligence techniques for sonar interpretation. Extensive analysis on both real and S3Simulator+ datasets further supports the robustness of the proposed framework. Together, CAFF and AQUA-X promote trustworthy and transparent artificial intelligence solutions for real-world sonar applications. © *2026 Acoustical Society of America*. https://doi.org/10.1121/10.0042386

(Received 4 August 2025; revised 9 January 2026; accepted 12 January 2026; published online 18 February 2026)

[Editor: James F. Lynch] Pages: 1583–1598

I. INTRODUCTION

Sound navigation and ranging (Sonar) (Cervenka and De Moustier, 1993) is a critical imaging modality for the detection and classification of underwater objects, particularly in low-visibility or turbid environments where optical systems fail. Unlike optical imagery, sonar images are shaped by three key acoustic phenomena: acoustic highlights, acoustic shadows, and acoustic noise (Grzadziel, 2023). Highlights appear as bright regions in side-scan sonar (SSS) imagery, caused by strong acoustic impedance difference between underwater objects and the seafloor (Blondel, 2010). These differences lead to high-intensity backscatter, often revealing details about an object's shape or orientation. Acoustic shadows, by contrast, are low- or zero-intensity (anechoic) regions that occur when objects obstruct the sonar pulse, preventing ensonification of the seafloor behind them. The shape and extent of the shadow depend on object height, sonar altitude, and the grazing angle, making shadows useful for estimating object geometry and relative positioning. However, their low intensity causes many conventional image-processing pipelines to misinterpret them as background or noise (Sarno Filho *et al.*, 2025).

Despite the semantic richness of shadows and highlights, most deep learning models developed for sonar analysis are adapted from optical-domain pipelines and thus fail to leverage the physics-based characteristics of sonar data (Hu *et al.*, 2021; Manning, 2002). Shadow regions, in particular, are often overlooked or improperly modeled, leading to suboptimal classification performance and limited generalization (Nguyen *et al.*, 2019). Moreover, sonar datasets are typically scarce, confidential, and costly to acquire to develop robust models. In defence and safety-critical applications, where decisions must be transparent and verifiable, black-box artificial intelligence (AI) systems (Von Eschenbach, 2021) are inadequate. This underscores the necessity of context-aware, explainable, and human-in-the-loop AI systems that align with expert reasoning and are robust and reliable in real-world sonar operations.

To address the aforementioned limitations, we propose a novel Context-Adaptive Fusion Framework (CAFF) to

[a)]Email: c58527@srmist.edu.in
[b)]Email: c54351@srmist.edu.in
[c)]Email: athiram@srmist.edu.in
[d)]Email: sureshr@iitm.ac.in
[e)]Email: sooraj.npol@gov.in

 © 2026 Acoustical Society of America

enhance both classification accuracy and model interpretability in sonar image analysis. The CAFF integrates three specialized classifiers: (i) Naive Classifier using combined features, (ii) ShadowNet modeling acoustic shadow regions, and (iii) HighlightNet focusing on high-intensity highlight regions. These classifiers are dynamically fused using an adaptive mechanism that selects the most relevant feature stream based on the context of the input image. This architecture allows the system to intelligently emphasize shadow-based or highlight-based cues, effectively mimicking the reasoning process of domain experts.

Further, to assess the performance of different explainable AI (XAI) approaches, a novel expert-in-the-loop framework, that is, Augmented QUality Assessment for eXplainability (AQUA-X) is presented in this paper, one of the first of its kind. AQUA-X enables domain experts to systematically evaluate and enhance model explanations through a dedicated trustworthiness assessment panel, providing a structured mechanism for human-guided explainability evaluation. To further ensure interpretability, XAI methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju *et al.*, 2017), SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017), and Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro *et al.*, 2016) are incorporated, each providing complementary insights through gradient-based saliency, Shapley value attribution, and local perturbation analysis, respectively, as detailed in the comprehensive XAI review by Minh *et al.* (2022). To support consistent evaluation, a structured XAI taxonomy and statistical validation scheme is presented via expert scores as well as the inclusion of metrics from the open source Quantus toolbox (Hedström *et al.*, (2023). This design supports the operational transparency goals emphasized by Defense Advanced Research Projects Agency (Gunning and Aha, 2019), and promotes the development of trustworthy AI systems for real-world sonar applications. The key contributions of the paper are the following:

- CAFF, a novel model for sonar image classification, focusing on shadow and highlight regions with interpretable attention maps;
- HighlightNet and ShadowNet, two mask-guided domain-specific classifiers using spatio-channel attention to specialize in highlight and shadow analysis;
- AQUA-X, an expert-in-the-loop evaluation framework for assessing the trustworthiness of explanations produced by Grad-CAM, SHAP, and LIME; and
- XAI Taxonomy and Metrics, a structured protocol combining expert feedback and quantitative metrics to evaluate explanation quality in sonar imaging.

A. Related work

1. Sonar image classification

Recent advances in sonar image classification tackle challenges such as data scarcity, noise, and domain shift. A contrastive adaptation of domain augmentation (CADA) is introduced to improve generalized zero-shot learning (GZSL) in side-scan sonar (SSS) classification (Jia *et al.*, 2025). A comprehensive overview of deep learning robustness in underwater robotics, highlighting vulnerabilities such as adversarial attacks and the simulation-to-reality gap (Aubard *et al.*, 2025) [Marine-PULSE (Du *et al.*, 2025b), a specialized dataset for underwater structures with transfer learning]. S3Simulator dataset (Kamal Basha and Nambiar, 2025) developed using Segment Anything Model and computer-aided design-based simulation for synthetic sonar data, demonstrating their utility in underwater object classification. Several real-world datasets have been developed, such as SeabedObjects-KLSG (Huo *et al.*, 2020) and SCTD (Zhang *et al.*, 2022), covering a range of targets like wrecks, mines, and planes. Yang *et al.*, (2023) employed diffusion models for realistic sonar image synthesis, whereas Du *et al.*, (2023) demonstrated superior classification performance using GoogleNet and Residual Network (ResNet101). Kernel-based extreme learning machines applied with principal component analysis for sonar classification (Zhu *et al.*, 2017), demonstrating alternative model effectiveness. Earlier works mainly emphasized feature extraction and classification accuracy, whereas CAFF advances these efforts by integrating domain-specific interpretability through HighlightNet and ShadowNet, thereby incorporating physics-based cues in sonar imagery.

2. XAI techniques for sonar data

XAI has become crucial in modern AI systems, particularly in high-stakes domains such as medicine and defence, where transparency, accountability, and trust are essential. State-of-the-art XAI tools such as Grad-CAM (Selvaraju *et al.*, 2017) provides visual explanations by highlighting class-relevant regions in input images using gradients from the final CNN layers. LIME (Ribeiro *et al.*, 2016) explains model predictions by locally approximating them with interpretable models, whereas SHAP (Lundberg and Lee, 2017) leverages game-theoretic principles to deliver consistent and reliable feature attributions. Both are widely adopted for *post hoc* model interpretability. The applicability of XAI for grayscale sonar imagery has been examined, showing that perturbation-based methods are more effective than backpropagation-based techniques because of modality-specific constraints (Richard *et al.*, 2025). In another study, machine learning models were used to classify passive sonar signals, with LIME employed to enhance interpretability, demonstrating strong performance using random forest classifiers (Manapaka *et al.*, 2024).

II. METHODOLOGY

In this section, the proposed explainable CAFF and expert-in-the-loop evaluation framework for underwater sonar image classification is described. Figure 1 presents a high-level overview of the proposed framework. Figure 2 details the internal architecture of CAFF, integrating the

https://doi.org/10.1121/10.0042386

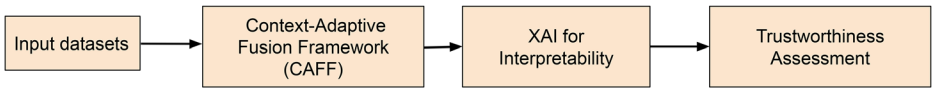

FIG. 1. High-level overview of the proposed explainable context-adaptive fusion and expert-in-the-loop evaluation framework, illustrating the main modules and their interactions. (i) CAFF, (ii) Explainability and Trustworthiness Module, and (iii) overall data flow from simulator and real sonar datasets.

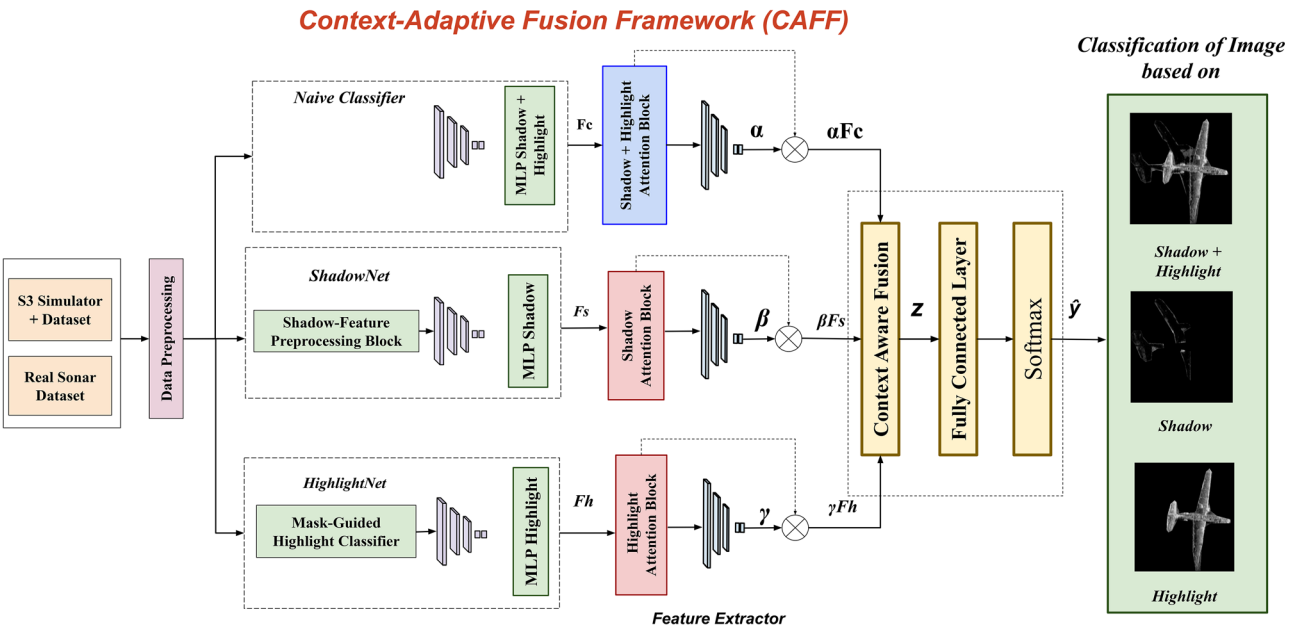

FIG. 2. Detailed architecture of the CAFF showing the Naive Classifier, ShadowNet, HighlightNet, and Context-Aware Fusion module for underwater sonar image classification.

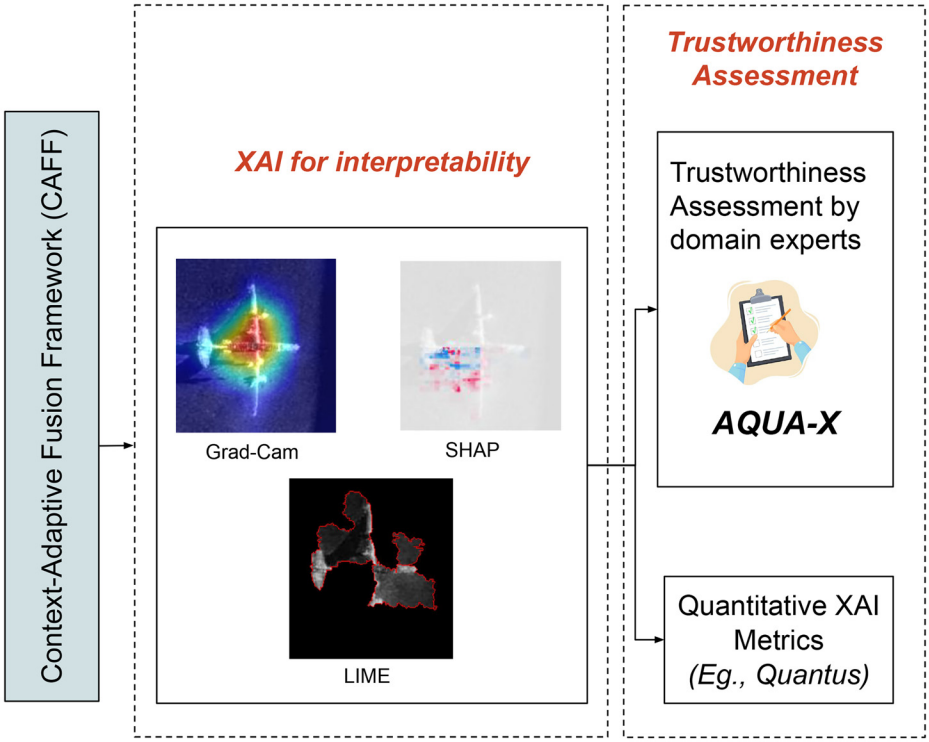

FIG. 3. Explainability and Trustworthiness Module integrating XAI techniques (Grad-CAM, SHAP, and LIME) with expert-in-the-loop and quantitative trustworthiness evaluation (AQUA-X and Quantus).

TABLE I. Comparison of classifier branches in the proposed CAFF model.

Model	Input type	Attention	Purpose	Backbone
Naive	Raw sonar image	Channel	Overall contextual cues	DenseNet-121
ShadowNet	Shadow mask (lightness–chroma–hue)	Spatial	Low-intensity shadow modeling	EfficientNet-B0
HighlightNet	CA-CFAR mask	Spatio-channel	High-intensity highlight modeling	DenseNet-161

Naive Classifier, ShadowNet, HighlightNet, and the Context-Aware Fusion Module. Figure 3 depicts the combined Explainability and Trustworthiness Module, which uses Grad-CAM, SHAP, and LIME for interpretability, and AQUA-X and Quantus for trust evaluation. Section II A details the dataset and preprocessing pipeline. Section II B presents the CAFF, including the design of NAIVE Classifier, ShadowNet, HighlightNet, and the Context-Aware Fusion Module. Section II D introduces the Explainability Module, incorporating Grad-CAM, SHAP, and LIME. Finally, Sec. II E describes the trustworthiness evaluation frameworks based on AQUA-X and Quantus, which enables validating and scoring the quality of generated explanations.

A. Dataset and preprocessing

The study utilizes a combination of real and synthetic sonar images sourced from SeabedObjects KLSG, Sonar Common Target Detection (SCTD) (Zhang *et al.*, 2022), and S3Simulator+ datasets (Kamal Basha and Nambiar, 2025). The KLSG and SCTD datasets primarily include ship, plane, and seabed classes, whereas S3Simulator+ extends coverage to ship, plane, mine, seabed, and others. To ensure uniformity, all images were converted from grayscale to three-channel red–green–blue with 8-bit precision, resized to 224×224 pixels, and normalized across real and synthetic domains. On average, approximately 3000–4000 images per class were included, totaling around 25 000 images. Data augmentation techniques (Shorten and Khoshgoftaar, 2019), such as random flipping, rotation, and contrast adjustment, were applied to enhance generalization.

B. CAFF

The CAFF as shown in Fig. 2 is designed to enhance sonar image classification by capturing region-specific acoustic cues. It comprises three parallel classifier branches: (i) Naive Classifier, which operates on the full input image without any spatial masking; (ii) ShadowNet, trained on acoustic shadow regions; and (iii) HighlightNet, trained on highlight regions. Although ShadowNet and HighlightNet are guided by attention-weighted masks emphasizing physically meaningful sonar features, the naive branch serves as a generic classifier without any region-specific bias, retaining global image context. A concise comparison of the three classifier branches in terms of input type, attention mechanism, purpose, and backbone network is presented in Table I.

1. ShadowNet: shadow feature classifier

The shadow extraction process (Fig. 4) adapts the methodology from (Silva *et al.*, 2018). It first converts images to a color space where L^* represents luminance, and A^* and B^* (LAB) denote chromaticity components along the green-red, blue-yellow axes, respectively, and subsequently transforms them into the lightness, chroma, hue (LCH) color space to isolate intensity and hue-based features, applies spectral ratio analysis, and uses K-Means clustering and morphological operations to generate a binary shadow mask (pixels labeled 1 for shadows, 0 for non-shadows). As presented in Basha S *et al.* (2025), segmented images with shadow annotations were used to train an EfficientNet-B0 model, which offers a good trade-off between accuracy and computational efficiency. The training incorporates shadow-aware data augmentation, including variations in lighting and occlusions, to enhance model robustness. This ensures the model can handle diverse sonar image conditions effectively.

2. HighlightNet: highlight feature classifier

In parallel with ShadowNet, a novel highlight-specific classifier, named HighlightNet, is also implemented. Specifically, HighlightNet is implemented as a Mask-Guided Highlight Classifier for precise classification of highlights in sonar imagery, as shown in in Fig. 5. The architecture comprises two key modules—(1) Mask Generation and (2) Spatio-Channel Attention.

The first module, Mask Generation, involves generating a binary highlight mask to identify regions of interest in the sonar image. The input sonar image is directly processed using the Cell-Averaging Constant False Alarm Rate (CA-CFAR) algorithm (Dai *et al.*, 2018), which segments areas with high acoustic intensity. This segmentation yields a highlight mask, representing potential underwater targets.

Parallel to mask generation, the input sonar image is passed through DenseNet161 (Huang *et al.*, 2017), a deep convolutional backbone selected for its strong feature reuse and ability to capture hierarchical representations. The extracted features are then refined using a Spatio-Channel Attention block, which enhances the model's focus on informative regions and channels.

a. Spatial attention. To highlight spatially significant regions, the feature map is processed by applying both average pooling and max pooling across spatial dimensions. The resulting two maps are concatenated and passed through a 1×1 convolution to generate a spatial attention map. This map emphasizes regions such as strong acoustic reflections

https://doi.org/10.1121/10.0042386

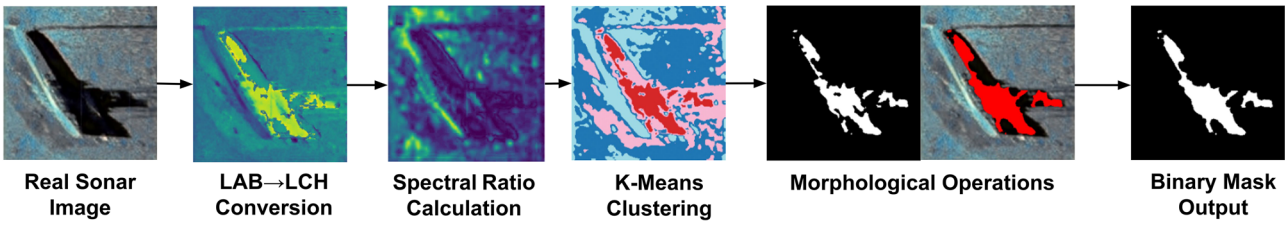

FIG. 4. Visual overview of the Shadow-Feature Preprocessing Block.

and is applied to the original features via element-wise multiplication.

b. Channel attention. To identify the most informative feature channels, global average pooling is applied across each channel. The pooled features are passed through a fully connected layer with sigmoid activation, producing a channel attention vector. This vector is used to reweight the feature channels through element-wise multiplication, enhancing discriminative features like texture and intensity variations.

Although the highlight mask is not directly input into this attention module, it plays a key supervisory role during training. An auxiliary mean squared error (MSE) loss is used to guide both the spatial and channel attention maps to align with the binary highlight mask. This ensures that the attention mechanisms consistently focus on true highlight regions in the sonar image.

This Spatio-Channel Attended Feature Map serves as the final attention-refined representation, which is subsequently passed through a multi-layer perceptron (MLP) that transforms the features for final decision-making. The output of the MLP is then fed into a Highlight Attention Block, which generates the final prediction by weighting the semantic contribution of highlight-specific features. Importantly, although the highlight mask is not directly input to the attention modules, it supervises them during training by enforcing alignment between the Final Attention Map and the Highlight Mask through the MSE loss. This ensures that the model's learned attention is grounded in physically meaningful highlight regions.

C. Context-aware fusion module

In complex visual environments, a single model often struggles to generalize across diverse contextual challenges such as shadows, overexposed highlights, and ambiguous textures. To address this, we propose an adaptive fusion model that leverages specialized sub-networks, each trained to handle specific visual conditions while integrating their complementary knowledge using a learned attention mechanism. Our architecture consists of three parallel branches: Naive Classifier, ShadowNet, and HighlightNet. During training, these branches operate independently and in parallel, each learning complementary feature representations. During inference and fusion, they act collaboratively by combining their feature maps through the adaptive attention mechanism to produce a robust classification decision.

Feature encoding and alignment is done by giving an input image I; each branch extracts a feature map as follows:

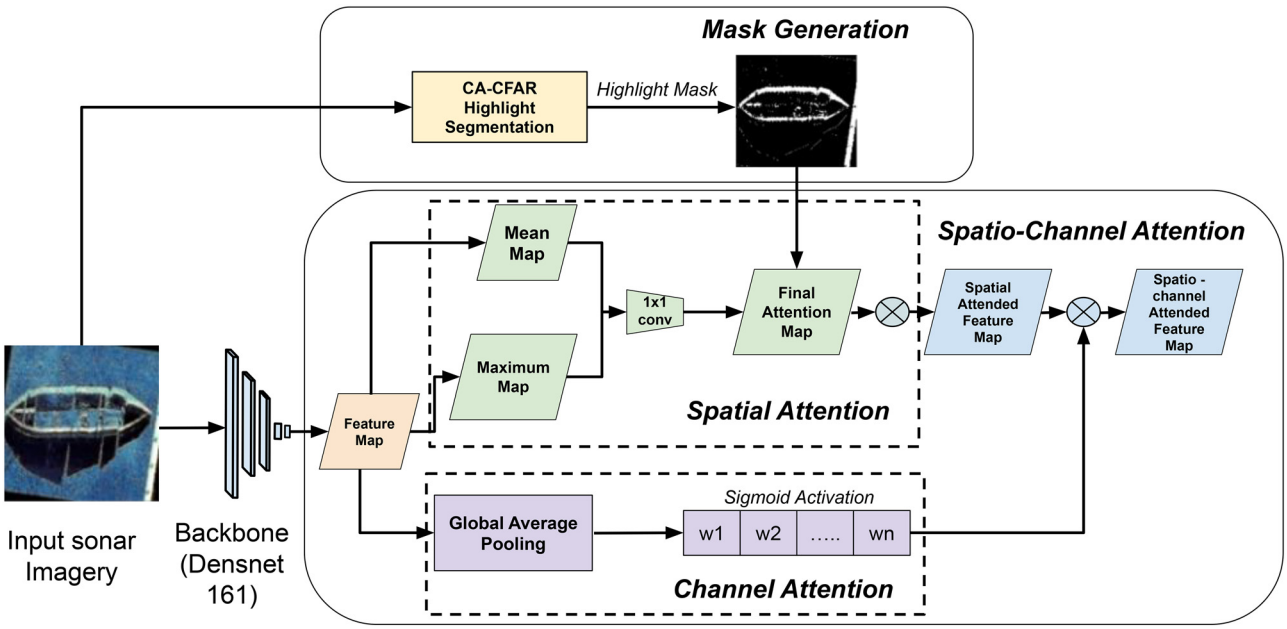

FIG. 5. Mask-Guided Highlight Classifier architecture.

$$F_c = \Phi_c(I), \quad F_s = \Phi_s(I), \quad F_h = \Phi_h(I). \tag{1}$$

Here, Φ_c, Φ_s, and Φ_h denote the feature extractors of the Naive Classifier, ShadowNet, and HighlightNet, respectively, applied to the input image I. The resulting features F_c, F_s, and F_h represent the semantic representations extracted by each branch. To ensure consistency in dimensionality and enable effective fusion, a 1×1 convolution is applied to project all features into a common embedding space, reducing each branch output from $7 \times 7 \times 2048$ to a uniform size of $7 \times 7 \times 512$,

$$F_x = \mathrm{Conv}_{1\times1}(f_x), \quad x \in \{c, s, h\}. \tag{2}$$

When performing the spatial attention mechanism to enhance region-specific sensitivity and suppress noisy activations, we apply a gated attention mechanism for each feature map F_x. The attention score at each location (i,j) is computed as

$$\tilde{e}_{i,j}^{(x)} = \lambda \mathbf{u}^T(\tanh(F_x(i,j)) \odot \sigma(F_x(i,j))). \tag{3}$$

These scores are normalized using a softmax across all spatial locations as

$$A_x(i,j) = \frac{\exp(\tilde{e}_{i,j}^{(x)})}{\sum_{(p,q)} \exp(\tilde{e}_{p,q}^{(x)})}, \quad x \in \{c, s, h\}. \tag{4}$$

The output attention map A_x highlights semantically relevant regions within the modality-specific context. Descriptor computation is computed by a global descriptor, which is obtained by weighted average pooling using the attention map,

$$\alpha = \sum_{i,j} A_c(i,j) \cdot F_c(i,j), \quad \beta = \sum_{i,j} A_s(i,j) \cdot F_s(i,j),$$
$$\gamma = \sum_{i,j} A_h(i,j) \cdot F_h(i,j). \tag{5}$$

These vectors encode the most informative visual cues from each stream, conditioned on spatial attention. Adaptive attention weights, which determine the influence of each descriptor in the final decision, are computed by first evaluating the ℓ_2 norms of the descriptors (α, β, γ) and then normalizing them to obtain context-aware importance scores as follows:

$$\overline{\alpha} = \frac{\|\alpha\|_2}{\|\alpha\|_2 + \|\beta\|_2 + \|\gamma\|_2},$$
$$\overline{\beta} = \frac{\|\beta\|_2}{\|\alpha\|_2 + \|\beta\|_2 + \|\gamma\|_2},$$
$$\overline{\gamma} = \frac{\|\gamma\|_2}{\|\alpha\|_2 + \|\beta\|_2 + \|\gamma\|_2}. \tag{6}$$

These weights adaptively highlight the most confident stream, giving dynamic control to the fusion mechanism. Final fusion and prediction is obtained by a fused representation by taking a weighted sum of all descriptors as follows:

$$z = \overline{\alpha} \cdot F_c + \overline{\beta} \cdot F_s + \overline{\gamma} \cdot F_h. \tag{7}$$

The final class prediction is generated via a fully-connected (FC) layer and softmax activation as

$$\hat{y} = \mathrm{Softmax}(Wz + b), \tag{8}$$

where W and b are learnable parameters. One significant benefit of this design is its modularity; each stream can be pre-trained independently, and the model can still make reasonable predictions even if one stream fails. Moreover, the use of explicit attention maps and norm-based confidence allows for intuitive interpretability and potential visualizations of which modality dominates in different scenarios.

D. Explainability module

To enhance interpretability, three complementary XAI methods are used: LIME (Ribeiro *et al.*, 2016), SHAP (Lundberg and Lee, 2017), and Grad-CAM (Selvaraju *et al.*, 2017), which are incorporated into the system. These techniques were selected for their ability to provide complementary forms of explanation. LIME generates local, instance-based interpretations, SHAP provides global feature-importance attribution based on Shapley values, and Grad-CAM produces class-specific visual heatmaps highlighting discriminative regions.

E. Trustworthiness assessment via AQUA-X and Quantus

1. AQUA-X assessment framework

To complement algorithmic interpretability with human reasoning, a structured evaluation protocol AQUA-X is implemented (Fig. 6). This framework enables domain experts to assess the quality of model-generated explanations across multiple dimensions of trustworthiness. Evaluations are conducted for all three explainability methods (Grad-CAM, LIME, SHAP), with the objective of measuring alignment between model explanations and human expectations in sonar interpretation.

Each explanation is evaluated using three human-centric criteria: Localization Accuracy (rated on a scale from 1 to 5), Completeness (a binary yes/no assessment), and Ease of Interpretation (also rated on a 1–5 scale) (Table II). Higher scores reflect stronger alignment with expert knowledge, greater coverage of relevant regions, and clearer visual explanations. To test explanation robustness under real-world and degraded conditions, an augmented validation set is utilized, as shown in Fig. 7. Approximately 120 sonar images were randomly sampled to ensure balanced coverage across the four evaluated classes: ship, plane, mine, and human. These samples were used to evaluate the reliability of the explainability methods under the AQUA-X protocol. Real sonar data were employed for the ship, plane, and human classes, whereas synthetic data were used for the

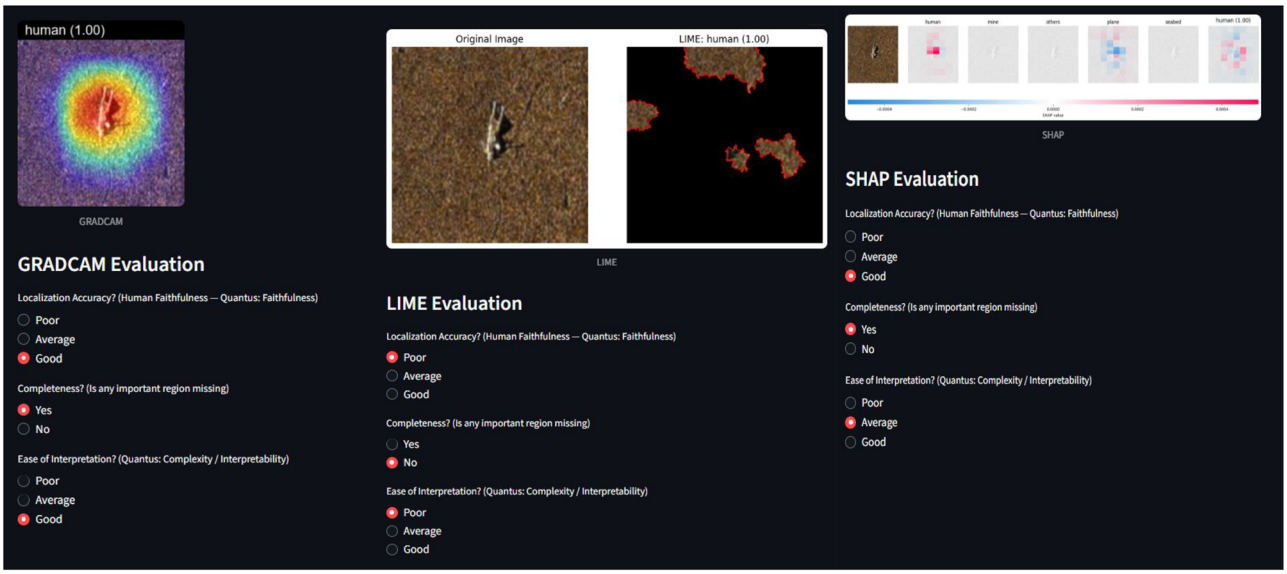

FIG. 6. AQUA-X; expert assessment of sonar explanations.

mine category. A single domain expert conducted the evaluations to maintain consistency, and future work will involve multiple raters to assess inter-expert agreement. Each image was assessed using the same three criteria to determine the most reliable explainer under adverse conditions. The AQUA-X evaluation interface is implemented as a graphical user interface tool and is publicly available at https://github.com/bashakamal/AQUA-X.

Following expert scoring from the AQUA-X panel, quantitative analyses compare the explanation methods and identify observed trends in their performance. A one-way analysis of variance (ANOVA) (Stähle and Wold, 1989) test is used to assess whether the average localization accuracy scores for Grad-CAM, SHAP, and LIME differed significantly. Tukey's honest significant difference (HSD) *post hoc* test (Hilton and Armstrong, 2006) identified which specific method pairs showed significant variation, such as whether SHAP outperformed LIME. For the ease of interpretation scores, ANOVA) and Tukey HSD were again applied to evaluate cognitive usability across methods. For the binary completeness ratings (yes/no), a chi-square test (Connelly, 2019) is conducted to examine whether any explainer consistently failed to highlight semantically important regions.

TABLE II. Evaluation protocol used in AQUA-X for explanation quality scoring.

Criterion	Description
Localization accuracy	Measures the alignment between highlighted regions and expert-defined object locations (reflecting human faithfulness).
Completeness	Assesses whether any critical regions or features are missing from the explanation.
Ease of interpretation	Evaluates how intuitively the explanation can be understood by a human observer (reflecting explanation clarity and cognitive load).

This evaluation provided a quantitatively informed trustworthiness profile for each XAI method, offering insights that go beyond visual inspection and supporting the development of more interpretable and operationally reliable sonar AI systems.

2. Quantitative saliency metrics via Quantus

For the quantitative salience metrics, the Quantus (2026) library is also utilized, which provides a comprehensive suite of benchmarking tools for interpretability. The Top-K Intersection score is used to evaluate localization performance, with higher scores indicating better alignment between explanation heatmaps and relevant semantic regions.

III. EXPERIMENTAL SETUP

This section outlines the experimental configuration adopted to train, evaluate, and interpret the proposed Context-Adaptive Fusion and XAI Sonar Interpretation System.

A. Model training and implementation details

All models in CAFF were implemented using PYTORCH. The Naive Classifier employed a DenseNet-121 backbone, the ShadowNet used an EfficientNet-B0 architecture, and the HighlightNet adopted a DenseNet-161 backbone. All networks were initialized with ImageNet pre-trained weights and fine-tuned on the sonar dataset using transfer learning, which provided faster convergence and improved generalization compared to training from scratch. Training is conducted on a workstation equipped with an NVIDIA RTX 4000 Ada Generation GPU (NVIDIA Corporation, Santa Clara, CA) graphics processing unit with 20 GB of memory. Cross-entropy loss is used as the objective function, and optimization is performed using the Adam optimizer with an initial learning rate of 1×10^{-4}. Each model is trained for

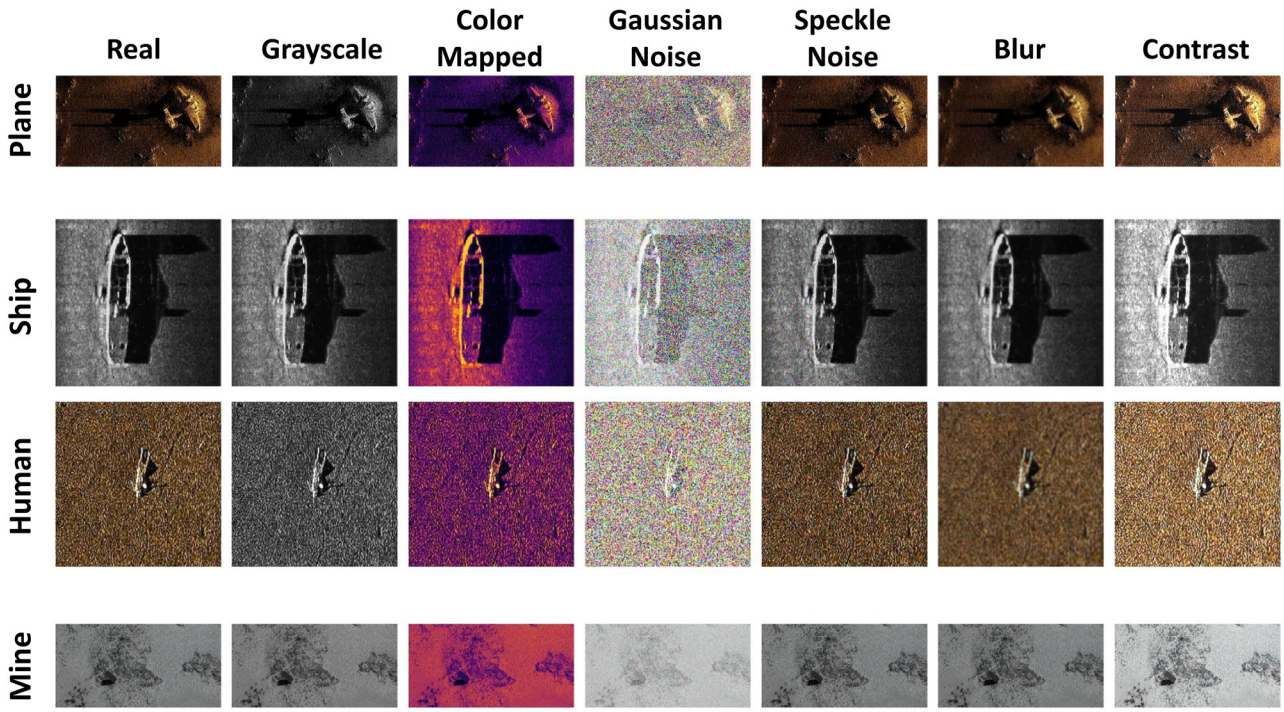

FIG. 7. Augmented sonar samples used in AQUA-X to evaluate the robustness and trustworthiness of XAI methods.

100 epochs using a batch size of 32, and the best-performing checkpoint is selected based on validation accuracy.

The dataset consisted of a mixture of real and simulated sonar images. A fixed 70:30 split was applied for training and validation, as the focus of this study was on evaluating generalization between real and synthetic domains rather than performing cross-validation.

B. Evaluation metrics

The performance of the proposed system is evaluated across two dimensions: classification accuracy and explanation quality. Classification accuracy is used as the primary metric to measure the effectiveness of the classifiers within the CAFF, including Naive, ShadowNet, and HighlightNet classifiers, and a confusion matrix is used for class-wise analysis.

Explanation quality is assessed through both quantitative metrics and expert-driven evaluation. Quantitative analysis is conducted using the Quantus (2026) library, where the Top-K Intersection metric is applied to quantify the overlap between saliency maps and semantically relevant regions, as described in Sec. II E 2. Expert evaluation is performed using the AQUA-X protocol (Sec. II E 1) in which Grad-CAM, SHAP, and LIME explanations are rated by domain experts based on localization accuracy, completeness, and ease of interpretation.

IV. EXPERIMENTAL RESULTS

In this section, the performance of the proposed CAFF and XAI Sonar Interpretation System is evaluated across classification and explainability dimensions. Quantitative and qualitative classification outcomes are reported in Secs. IV A 1 and IV A 2, respectively. Explainability analyses, including Grad-CAM, SHAP, LIME, and Most Relevant First/Least Relevant First (MoRF/LeRF) perturbation studies, are presented in Sec. IV A 2. Quantitative explanation metrics are provided in Table IV, and expert-based evaluation via the AQUA-X protocol is discussed in Sec. IV C.

A. Classification results

1. Quantitative results

The performance of each classifier within the CAFF is summarized in Table III. Corresponding class-wise confusion matrices are presented in Fig. 8 for a more detailed analysis. The overall accuracy of the Context-Adaptive Fusion model is higher compared to that of the individual classifiers. This demonstrates that the attention-guided integration of features from the Naive Classifier, ShadowNet, and HighlightNet effectively enhances classification performance by focusing on more informative regions of the input.

From the confusion matrices in Fig. 8, we observe that the Naive Classifier [Fig. 8(a)] exhibits moderate

TABLE III. Classification accuracy of different classifiers.

Classifier	Accuracy (%)
Naive Classifier	92.24
Shadow Classifier	85.98
Highlight Classifier	94.86
Context-Adaptive Fusion	96.75

https://doi.org/10.1121/10.0042386

classification performance, with visible misclassifications across several diagonal and off diagonal elements. ShadowNet [Fig. 8(b)] slightly underperforms, especially in classes affected by shadowed regions (human is misclassified as seabed by 70.7%, and others is misclassified as seabed by 31.4%), where confusion is more prominent. HighlightNet [Fig. 8(c)] improves performance by reducing errors in highlight-rich classes but still exhibits certain off diagonal misclassifications (human is misclassified as ship by 15.5% and as seabed by 7.6%). In contrast, the Context-Adaptive Fusion model [Fig. 8(d)] displays a much stronger diagonal dominance with minimal confusion between classes. The sharpness of the diagonal in its confusion matrix reflects more confident and correct predictions, validating the effectiveness of combining spatial cues from shadow and highlight features. These improvements illustrate that the fusion mechanism is not only quantitatively superior but also qualitatively more consistent in distinguishing between complex sonar classes. The fusion model benefits from complementary information across multiple domains, enabling better generalization and robustness in real-world underwater environments.

2. Qualitative results

As illustrated in Fig. 9, the attention values corresponding to $\overline{\alpha}$ (Naive Classifier), $\overline{\beta}$ (ShadowNet), and $\overline{\gamma}$ (HighlightNet) vary depending on the image conditions, specifically the presence of shadows and highlights. These attention weights reflect the degree to which each classifier focuses on specific regions of the image (i.e., shadowed or highlighted areas). In Fig. 9(a), where shadows dominate, the attention value for $\overline{\beta}$ (ShadowNet) is significantly higher. Conversely, in Fig. 9(b), where highlight regions are more prominent, the attention weights for both $\overline{\alpha}$ (Naive Classifier) and $\overline{\gamma}$ (HighlightNet) increase. From these observations, it was evident that attention was effectively allocated based on the visual characteristics of the input. The importance of contextual features was adaptively weighted through the proposed fusion strategy, allowing the model to prioritize salient regions relevant to

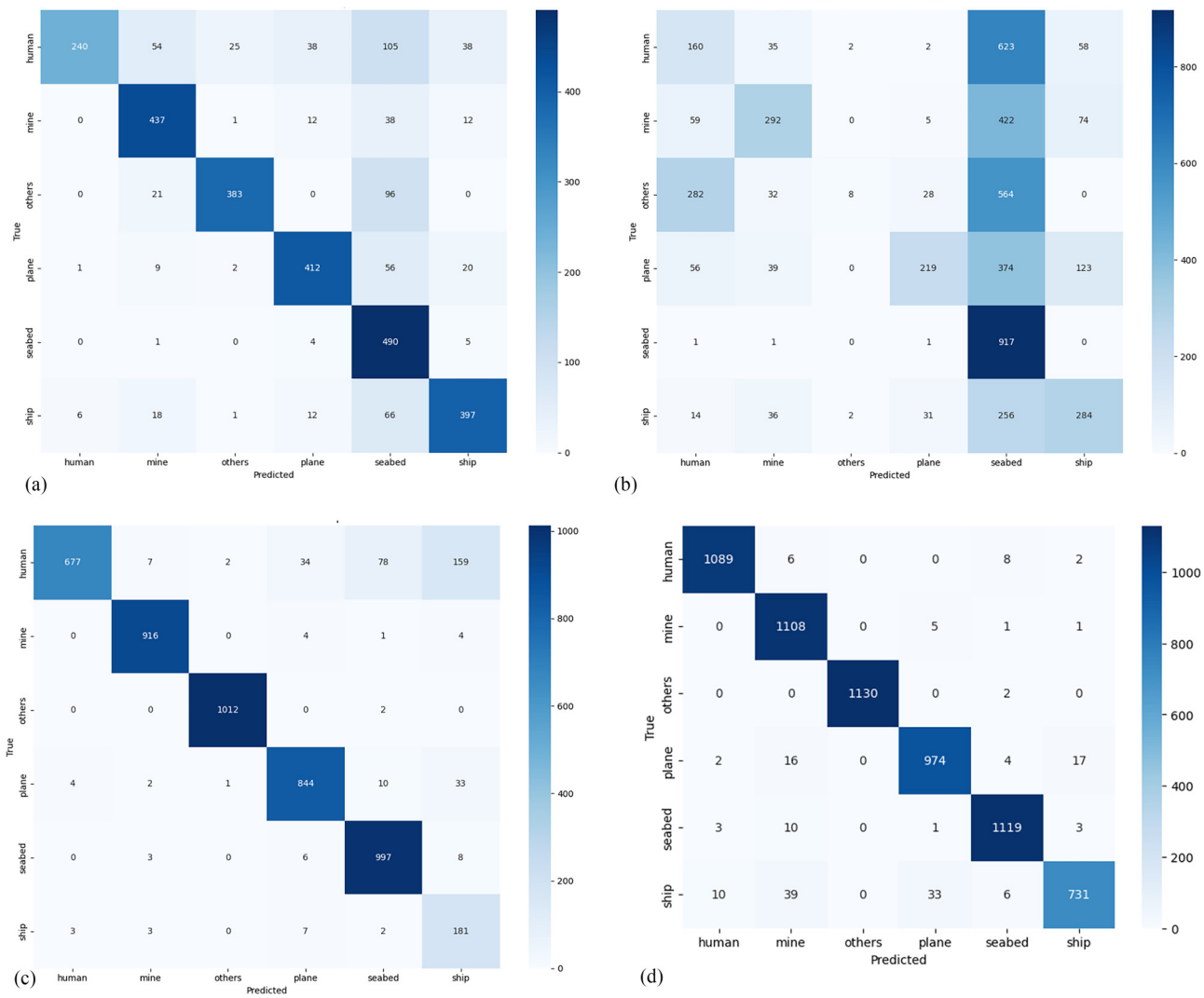

FIG. 8. Confusion matrices for the Naive, Shadow, Highlight, and Context-Adaptive Fusion models: (a) Naive, (b) ShadowNet, (c) HighlightNet, (d) CAFF.

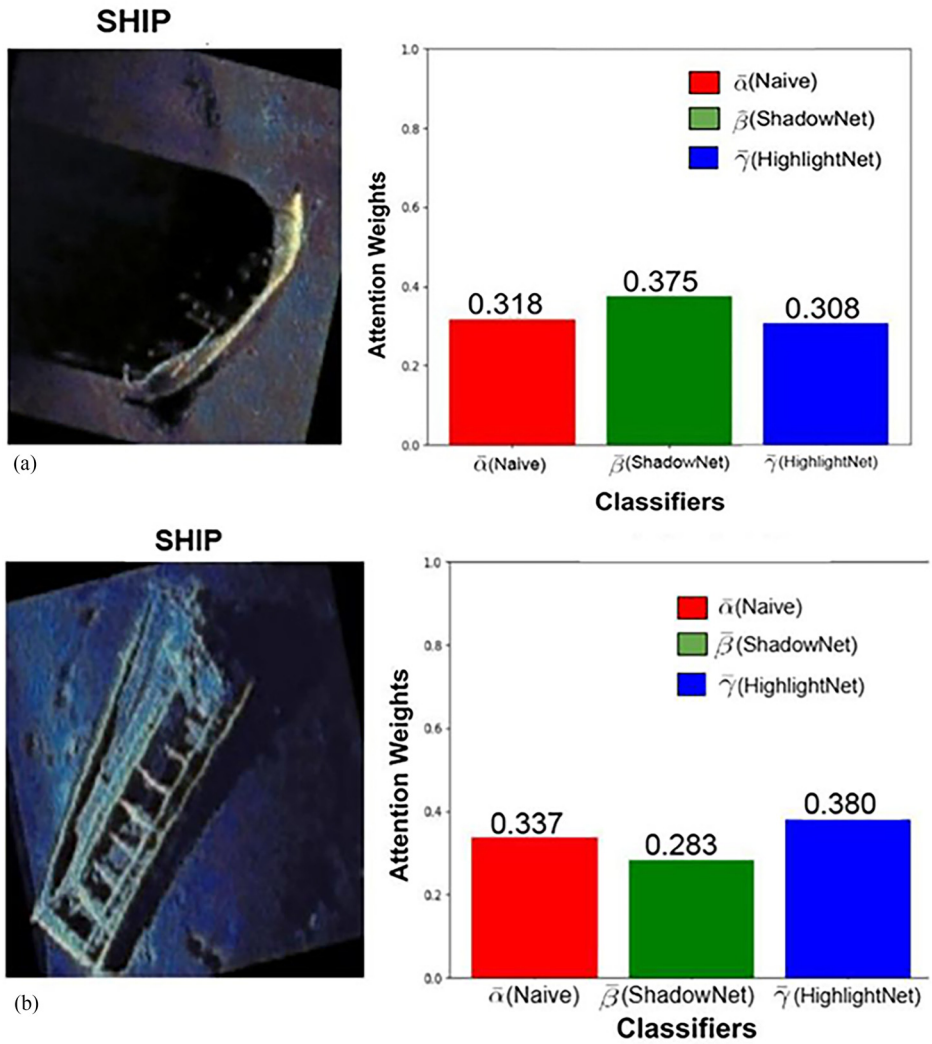

FIG. 9. Attention weights $\overline{\alpha}$, $\overline{\beta}$, and $\overline{\gamma}$ under two image conditions: (a) Ship with dominant shadow regions, (b) ship with dominant highlight regions.

sonar interpretation. This context-adaptive attention fusion significantly contributed to enhanced interpretability and explainability, establishing a pathway toward more transparent model behavior. These qualitative insights form the foundation for the subsequent explainability evaluation, where XAI methods are quantitatively analyzed.

B. Tuning the explainability results

This section presents parameter-tuning and sensitivity analyses for the three explainability methods LIME, SHAP, and Grad-CAM to ensure consistent and faithful visualization quality across experiments.

1. LIME analysis

To assess the effect of segmentation on LIME explanations, alternative algorithms such as Simple Linear Iterative Clustering (SLIC; (Achanta *et al.*, 2012)), Quickshift (Vedaldi and Soatto, 2008), and Felzenszwalb (Felzenszwalb and Huttenlocher, 2004) are integrated into the framework. In sonar imagery, where acoustic highlights and shadows

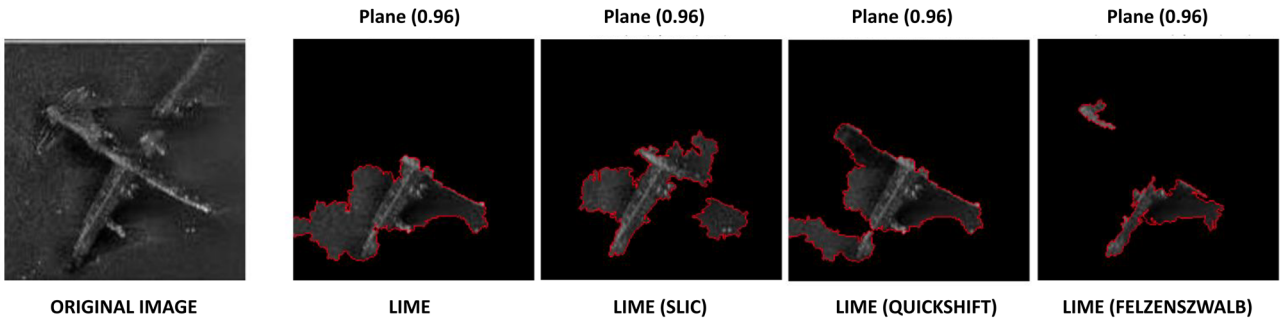

FIG. 10. LIME explanations under different segmentation techniques.

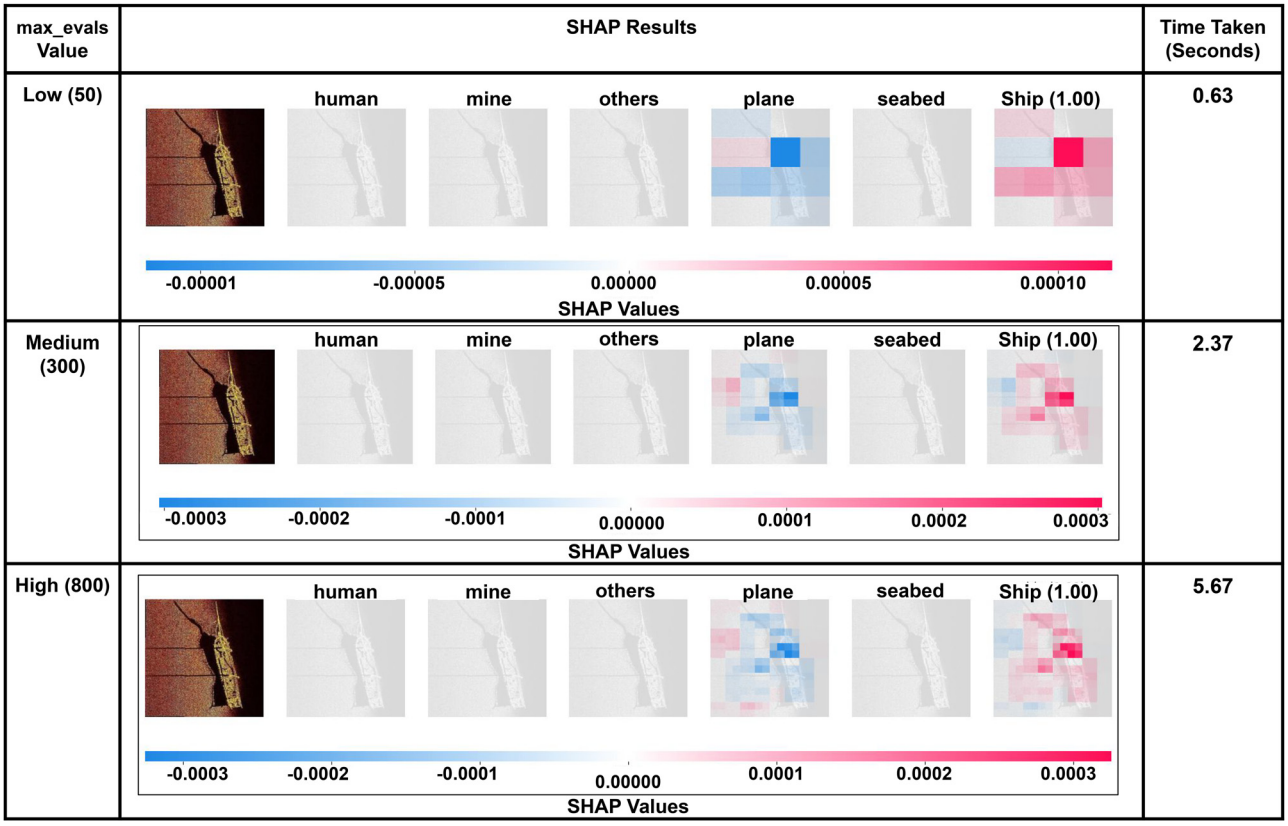

FIG. 11. SHAP analysis on sonar images. Impact of varying max_evals (Low: 50, Medium: 300, High: 800) on attribution quality and computation time.

dominate, default superpixels, perceptually homogeneous clusters of neighboring pixels with similar intensity or texture, often fail to align with meaningful structures. As shown in Fig. 10, segmentation choice directly impacts explanation quality. SLIC offers compact, boundary-aligned regions; Quickshift preserves object shapes; and Felzenszwalb leads to fragmented, less interpretable outputs. These results underscore segmentation granularity as a crucial factor for reliable sonar image explanations. Here, segmentation *choice* refers to the algorithm used (e.g., SLIC, Quickshift, or Felzenszwalb), whereas segmentation *granularity* denotes the number and size of superpixels generated by these methods, which directly influence the level of detail in the LIME explanations. Although the predicted class and confidence (e.g., plane = 0.96) remain unchanged across segmentation choices, the regions highlighted by LIME differ, showing that segmentation primarily influences the localization of explanatory features rather than the classification outcome.

2. SHAP analysis

SHAP parameter ablation (Fig. 11) explores how varying the max_evals parameter defining the maximum number of model evaluations used to approximate Shapley values affects explanation quality and runtime. Three settings low (50), medium (300), and high (800) are tested. Higher values

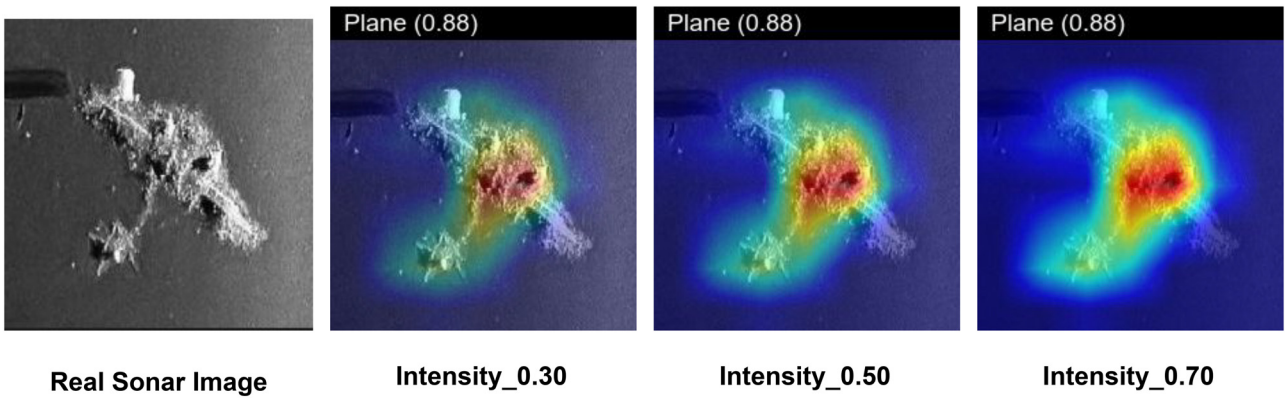

FIG. 12. Grad-CAM ablation on sonar images with varying intensity thresholds (0.3, 0.5, 0.7).

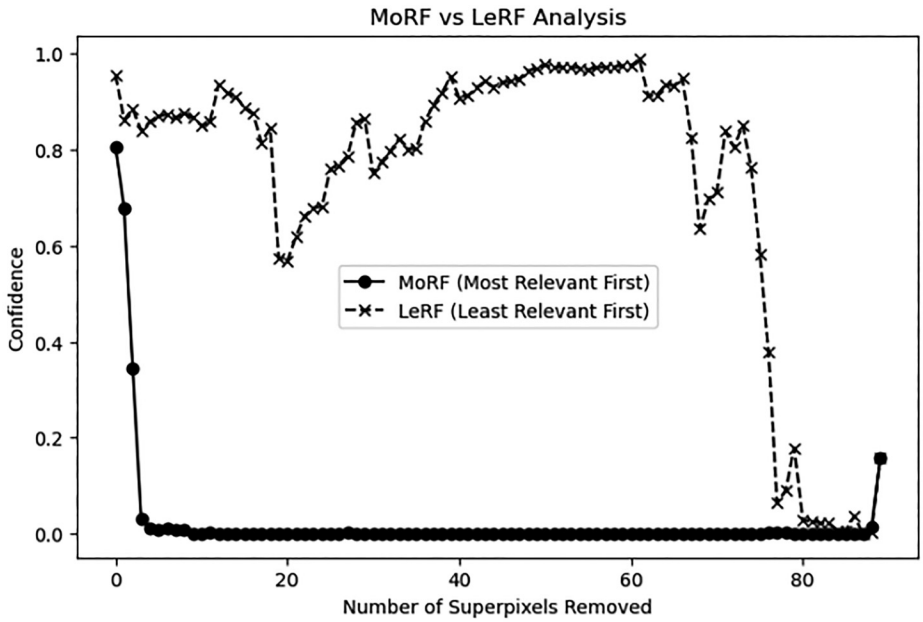

FIG. 13. MoRF and LeRF analysis for sonar image explanations. Reliable explainers show sharp MoRF degradation and stable LeRF responses.

yield more coherent and boundary-aligned explanations but increase computation time (0.63 s to 5.67 s). This highlights a trade-off between fidelity and efficiency, important for real-time sonar interpretation.

3. Grad-CAM analysis

Grad-CAM visualizations for real sonar images are generated using intensity thresholds of 0.3, 0.5, and 0.7 (Fig. 12). Lower thresholds produce diffuse heatmaps,

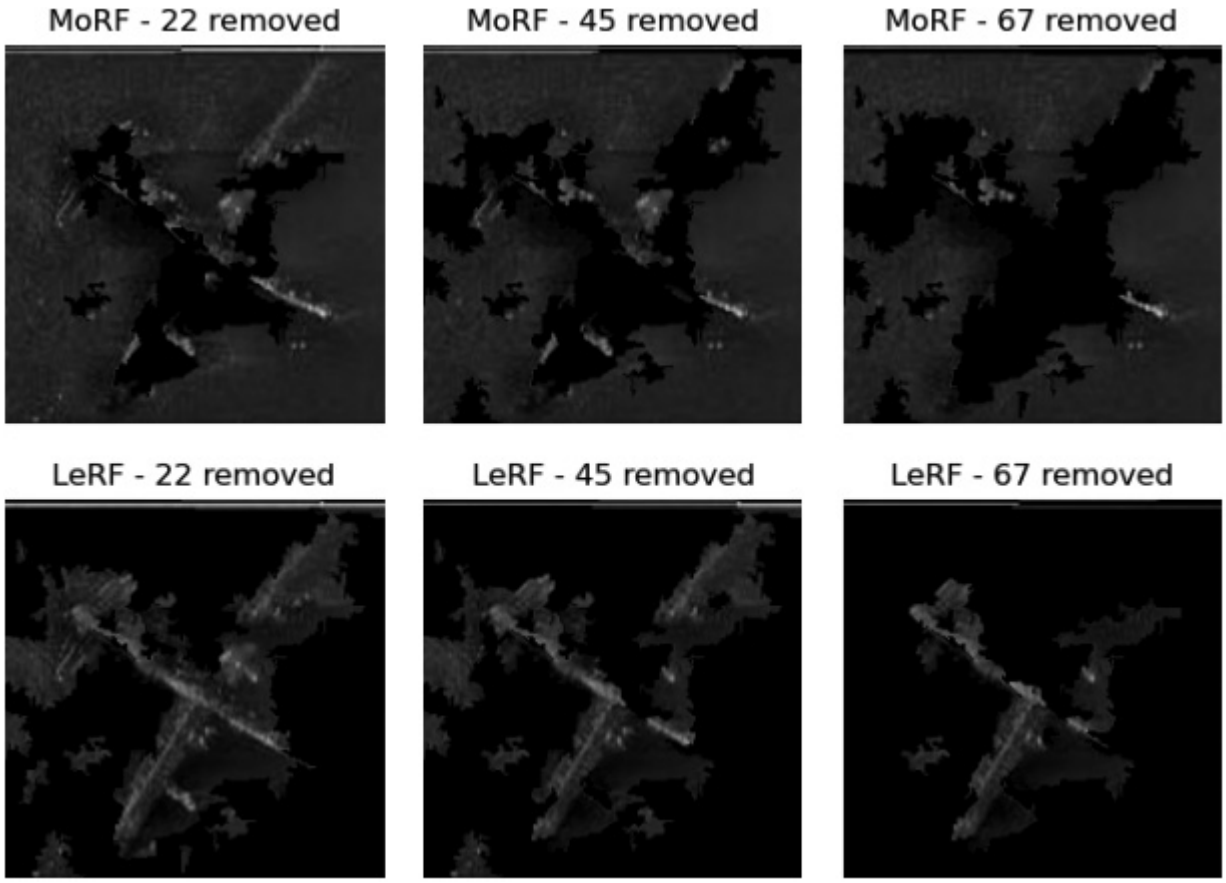

FIG. 14. MoRF measures the drop in model confidence when the most important pixels (as identified by the explainer) are progressively masked. LeRF does the same for the least important pixels.

https://doi.org/10.1121/10.0042386

TABLE IV. Top-K Intersection scores for selected XAI methods.

Explainer	Top-K Intersection Score
IntegratedGradients	0.4532
DeepLiftShap	0.4743
GradientShap	0.4411
LIME	0.0011
KernelShap	0.0790

whereas higher ones enhance localization around object contours and acoustic shadows. The results highlight Grad-CAM's sensitivity to intensity settings and the need for careful tuning to ensure meaningful explanations. These thresholds apply solely to visualization and do not affect expert evaluation in AQUA-X.

4. MoRF and LeRF reliability analysis

To quantitatively assess the faithfulness (Lakkaraju *et al.*, 2023) component of explanation quality within the proposed XAI evaluation protocol, perturbation-based reliability tests are conducted on the explanations generated using LIME, chosen for its compatibility with input-level perturbation analysis. Specifically, the MoRF and LeRF strategies are used. In the MoRF approach, the most salient regions, as identified by the explainer, are progressively masked, leading to a sharp decline in prediction confidence indicating that the model's decision process is closely aligned with the identified explanations. Conversely, in the LeRF approach, the least relevant regions are masked first, during which prediction scores remain relatively stable. As illustrated in Figs. 13 and 14, this contrasting behavior is indicative of the reliability and faithfulness of the generated explanations in the sonar domain.

Table IV presents the Top-K Intersection scores computed using the Quantus (2026) library to evaluate localization performance. DeepLiftShap and IntegratedGradients achieve the highest localization scores, demonstrating strong correspondence between their saliency maps and relevant image regions. In contrast, LIME and KernelShap perform poorly, indicating limited spatial alignment in the sonar context. Although Grad-CAM, LIME, and SHAP were used in the proposed framework, Grad-CAM is not currently supported by Quantus. Additional explainers were included as part of the standardized evaluation suite in the Quantus (2026) library to ensure completeness and reproducibility in benchmarking saliency-based localization.

C. Statistical evaluation of trustworthiness (AQUA-X)

To evaluate the reliability of XAI methods, expert ratings are collected and statistically analyzed using the AQUA-X framework. The three explainers Grad-CAM, SHAP, and LIME are assessed based on Correctness, Ease of Interpretation, and Completeness. Although a one-way ANOVA and Tukey's HSD test are applied to compare trends across explainers, the results are interpreted descriptively because all ratings were provided by a single expert.

Figure 15 illustrates the average human ratings across three dimensions—Correctness, Completeness, and Ease of Understanding—for each XAI method, Grad-CAM, LIME, and SHAP, and revealed the following:

- Grad-CAM emerged as the top-performing method overall. It received the highest scores for both Correctness (~ 4.0) and Ease of Understanding (~ 4.2), indicating that participants found its explanations reliable and easy to interpret. However, it scored relatively low on Completeness (~ 0.4), suggesting that some important information is perceived to be missing in its explanations.
- LIME consistently underperformed across all three dimensions. It had the lowest ratings for Correctness (~ 2.6) and Ease of Understanding (~ 2.7), reflecting users' difficulty in trusting and interpreting its outputs. Although its Completeness score (~ 0.6) is slightly higher

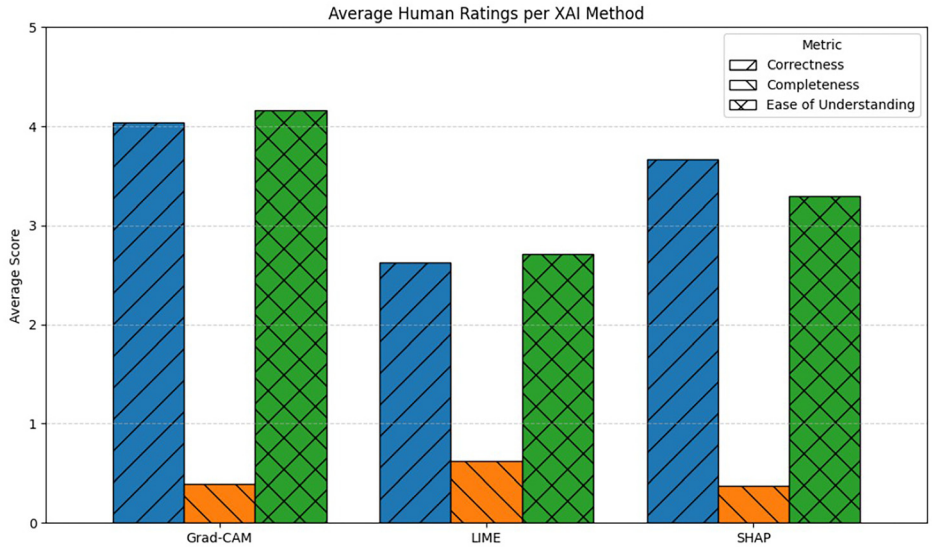

FIG. 15. Average human ratings across three dimensions.

TABLE V. Tukey HSD—pairwise differences (positive = higher first).

Criterion	Group 1	Group 2	Mean diff.[a]	p-value	Significant
Correctness	LIME	Grad-CAM	+1.42	0.0002	Yes
Correctness	SHAP	Grad-CAM	+0.38	0.5144	No
Correctness	LIME	SHAP	−1.04	0.0074	Yes
Ease of understanding	LIME	Grad-CAM	+1.46	0.0001	Yes
Ease of understanding	SHAP	Grad-CAM	+0.88	0.0295	Yes
Ease of understanding	LIME	SHAP	−0.58	0.2023	No

[a]Note: "diff." denotes the mean difference (Group 1 − Group 2).

than those of Grad-CAM and SHAP, the overall perception remained poor.

- SHAP performed moderately well, with respectable scores in Correctness (∼3.7) and Ease of Understanding (∼3.3). Nevertheless, it received the lowest Completeness score (∼0.3), suggesting that users felt key aspects were missing from its explanations.

From the ANOVA results, statistically significant differences are observed across the explainers for both Correctness ($F = 9.31, p = 0.00016$) and Ease of Interpretation ($F = 9.34$, $p = 0.00016$), indicating variability in human perception among the methods. As all images were rated by a single domain expert, these ANOVA results are interpreted descriptively as indicative trends rather than formal inferential outcomes.

Pairwise comparisons using Tukey's HSD test are shown in Table V.

Observations from Table V are the following:

- For Correctness, significantly higher scores are assigned to Grad-CAM and SHAP compared to LIME, with no significant difference observed between Grad-CAM and SHAP.
- For Ease of Understanding, Grad-CAM is rated significantly higher than both LIME and SHAP.

From an AQUA-X evaluation perspective, Grad-CAM stands out as the most suitable XAI method in this domain, offering high faithfulness and interpretability. It is rated highest for its accuracy and visual clarity, despite minor limitations in completeness. SHAP strikes a moderate balance but does not fully capture all relevant aspects. Although LIME provides comparatively more complete explanations, it scores lowest overall because of reduced correctness and poor interpretability. These findings reinforce Grad-CAM's alignment with the requirements of sonar interpretation, offering a reliable trade-off between explanation clarity and model decision fidelity.

V. COMPARISON WITH STATE-OF-THE-ART METHODS

A comprehensive comparison with existing side-scan sonar classification studies is presented in Table VI. Most prior works (Xie *et al.*, 2025; Ge *et al.*, 2021; Du *et al.*, 2023; Du *et al.*, 2025a) rely solely on raw image (RI) features extracted from the SeabedObjects-KLSG dataset, typically covering four object categories: drowning victim (human), aircraft, shipwreck, and seafloor. Although these methods achieve strong performance (94% to 97%), they lack shadow or highlight modeling and are limited in class diversity and interpretability.

In contrast, our CAFF integrates complementary shadow and highlight representations, fused through adaptive attention, and evaluated using explainable AI (XAI) methods including Grad-CAM, SHAP, and LIME. Our model is trained and validated on a combined dataset comprising S3Simulator+, SCTD, and SeabedObjects-KLSG, encompassing six classes (ship, plane, mine, human, seabed, and others) with a significantly higher sample volume. Despite the broader class coverage and increased complexity, CAFF attains an accuracy of 96.75%, comparable to or exceeding prior methods, while offering superior robustness, scalability, and interpretability across both real and synthetic domains.

VI. CONCLUSION

This work presents an explainable Context-Adaptive Fusion and expert-in-the-loop evaluation framework for underwater sonar image classification. The system integrates multiple domain-specific classifiers—Naive Classifier,

TABLE VI. Comparison with existing side-scan sonar image classification methods.

Method and ref.	Dataset	RI	SH	H	Fuse	XAI	Acc. (%)
Ge *et al.* (Ge *et al.*, 2021)	KLSG	Yes	No	No	No	No	97.32
Zhang *et al.* (Du *et al.*, 2023)	KLSG	Yes	No	No	No	No	94.27
Li *et al.* (Du *et al.*, 2025a)	KLSG	Yes	No	No	No	No	95.56
Chen *et al.* (Xie *et al.*, 2025)	KLSG + SCTD	Yes	No	No	No	No	96.80
Basha *et al.* (Basha S *et al.*, 2025)	S3Sim+, KLSG, SCTD	Yes	Yes	No	Yes	No	96.75
Ours (CAFF + XAI)	S3Sim+, SCTD, KLSG	Yes	Yes	Yes	Yes	Yes	96.75

ShadowNet, and HighlightNet— through a fusion mechanism that dynamically adapts to contextual cues such as shadows and highlights. The framework incorporates *post hoc* XAI techniques, including Grad-CAM, SHAP, and LIME, to generate model explanations, which are rigorously assessed using both quantitative metrics [via the Quantus (2026) library] and qualitative expert feedback through the AQUA-X protocol. Grad-CAM consistently demonstrates superior interpretability and correctness, followed by SHAP and LIME. The integration of expert-in-the-loop evaluation ensures that the system's decisions and explanations remain aligned with domain knowledge, thus contributing to the development of reliable and transparent AI systems for sonar-based defence, marine exploration, and autonomous underwater operations. Extensive analysis carried out on both S3Simulator+ and real sonar datasets further validates the model's superior performance and practical relevance. Future work will explore integrating causality-based explainability enabling more trustworthy and intervention-aware explanations. We also plan to incorporate large language models (LLMs) with prompt-based visual-language guidance, refined through human inspector feedback, to support adaptive and semantically meaningful interpretations within AQUA-X.

ACKNOWLEDGMENTS

This work was supported by the Naval Research Board (NRB), Defence Research and Development Organization, Government of India, Grant No. NRB/505/SC/22–23.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The study utilizes a combination of real and synthetic sonar images sourced from SeabedObjects KLSG (Huo *et al.*, 2020), SCTD (Zhang *et al.*, 2022), and S3Simulator+ datasets (Kamal Basha and Nambiar, 2025). The data from SeabedObjects KLSG and SCTD are publicly available. Sample images from the S3Simulator+ dataset are available at https://github.com/bashakamal/AQUA-X, and the full dataset is publicly available at SCTD (https://github.com/MingqiangNing/SCTD).

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. (**2012**). "SLIC superpixels compared to state-of-the-art superpixel methods," IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2274–2282.

Aubard, M., Madureira, A., Teixeira, L., and Pinto, J. (**2025**). "Sonar-based deep learning in underwater robotics: Overview, robustness, and challenges," IEEE J. Oceanic Eng. **50**, 1866.

Basha S, K., Kiran B, A., Nambiar, A., and Rajendran, S. (**2025**). "A novel context-adaptive fusion of shadow and highlight regions for efficient sonar image classification," arXiv:2506.01445.

Blondel, P. (**2010**). *The Handbook of Sidescan Sonar* (Springer Science & Business Media, New York).

Cervenka, P., and De Moustier, C. (**1993**). "Sidescan sonar image processing techniques," IEEE J. Oceanic Eng. **18**(2), 108–122.

Connelly, L. (**2019**). "Chi-square test," Medsurg. Nurs. **28**(2), 127.

Dai, X., Gao, X., Zhou, J., Zhang, S., and Li, W. (**2018**). "Accumulated CA-CFAR process in 2-D for online object detection from sidescan sonar data," IEEE Geosci. Remote Sens. Lett. **15**(2), 304–308.

Du, X., Sun, Y., Song, Y., Chi, W., Dong, L., and Zhao, X. (**2025a**). "Impact of input image resolution on deep learning performance for side-scan sonar classification: An accuracy–efficiency analysis," Remote Sens. **17**(14), 2431.

Du, X., Sun, Y., Song, Y., Dong, L., Tao, C., and Wang, D. (**2025b**). "Recognition of underwater engineering structures using CNN models and data expansion on side-scan sonar images," J. Mar. Sci. Eng. **13**(3), 424.

Du, X., Sun, Y., Song, Y., Sun, H., and Yang, L. (**2023**). "A comparative study of different CNN models and transfer learning effect for underwater object classification in side-scan sonar images," Remote Sens. **15**(3), 593.

Felzenszwalb, P. F., and Huttenlocher, D. P. (**2004**). "Efficient graph-based image segmentation," Int. J. Comput. Vision **59**(2), 167–181.

Ge, Q., Ruan, F., Qiao, B., Zhang, Q., Zuo, X., and Dang, L. (**2021**). "Side-scan sonar image classification based on style transfer and pre-trained convolutional neural networks," Electronics **10**(15), 1823.

Grzadziel, A. (**2023**). "The impact of side-scan sonar resolution and acoustic shadow phenomenon on the quality of sonar imagery and data interpretation capabilities," Remote Sens. **15**(23), 5599.

Gunning, D., and Aha, D. (**2019**). "DARPA's explainable artificial intelligence program," AI Mag. **40**(2), 44–58.

Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., and Höhne, M.-C. (**2023**). "Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond," J. Mach. Learn. Res. **24**(34), 1–11.

Hilton, A., and Armstrong, R. A., "Statnote 6: Post-hoc ANOVA tests," Microbiologist **2006**, 34–36.

Hu, X., Chu, L., Pei, J., Liu, W., and Bian, J. (**2021**). "Model complexity of deep learning: A survey," Knowl. Inf. Syst. **63**(10), 2585–2619.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (**2017**). "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Honolulu, HI (July 21–26), pp. 4700–4708.

Huo, G., Wu, Z., and Li, J. (**2020**). "Underwater object classification in sidescan sonar images using deep transfer learning and semisynthetic training data," IEEE Access **8**, 47407–47418.

Jia, Y., Ye, X., Li, P., and Guo, S. (**2025**). "Contrastive adaptation on domain augmentation for generalized zero-shot side-scan sonar image classification," IEEE Trans. Instrum. Meas **74**, 5021013.

Kamal Basha, S., and Nambiar, A. (**2025**). "S3Simulator: A benchmarking side scan sonar simulator dataset for underwater image analysis," in *Pattern Recognition. ICPR 2024*, Lecture Notes in Computer Science, edited by A. Antonacopoulos, S. Chaudhuri, R. Chellappa, C. L. Liu, S. Bhattacharya, and U. Pal (Springer, Cham, Switzerland), Vol. 15316, pp. 219–235.

Lakkaraju, H., Li, X., Du, M., Chen, J., Chai, Y., and Xiong, H. (**2023**). "M4: A unified XAI benchmark for faithfulness evaluation of feature attribution methods across metrics, modalities, and models," in *Advances in Neural Information Processing Systems (NeurIPS)*.

Lundberg, S. M., and Lee, S.-I. (**2017**). "A unified approach to interpreting model predictions," arXiv:1705.07874.

Manapaka, B. N., Malkapurapu, S. K., and Thadikemalla, V. S. G. (**2024**). "An explainable ML model for multiclass marine vessel classification using passive sonar signals," in *Proceedings of the International Conference on Distributed Computing and Optimization Techniques (ICDCOT)*, Bengaluru, India (March 15–16) (IEEE, New York), pp. 1–6.

Manning, R. (**2002**). "Small object classification performance of high-resolution imaging sonars as a function of image resolution," *Proceedings of the OCEANS'02 MTS/IEEE*, Biloxi, MS (October 29–31) (IEEE, New York), Vol. 4, pp. 2156–2163.

Minh, D., Wang, H. X., Li, Y. F., and Nguyen, T. N. (**2022**). "Explainable artificial intelligence: A comprehensive review," Artif. Intell. Rev. **55**(5), 3503–3568.

Nguyen, H.-T., Lee, E.-H., and Lee, S. (**2019**). "Study on the classification performance of underwater sonar image classification based on convolutional neural networks for detecting a submerged human body," Sensors **20**(1), 94.

Quantus (**2026**). https://github.com/understandable-machine-intelligence-lab/Quantus (Last viewed 30 January 2026).

Ribeiro, M. T., Singh, S., and Guestrin, C. (**2016**). "'Why should I trust you?': Explaining the predictions of any classifier," arXiv:1602.04938.

Richard, G. J., Habonneau, J., Guériot, D., and Le Caillec, J.-M. (**2025**). "Explainable AI methods for underwater mine warfare," IEEE J. Sel. Top. Appl. Earth Observ. **18**, 12755–12772.

Sarno Filho, E. P., Santos, A. D., Simas Filho, E. F., Fernandes, A. C. L., de Seixas, J. M., and de Moura, N. N. (**2025**). "Hilbert–Huang transform with intelligent noise reduction for passive sonar signal processing," IEEE J. Oceanic Eng. **50**, 1387–1402.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (**2017**). "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision* (ICCV), Venice, Italy) (October 22–29), pp. 618–626.

Shorten, C., and Khoshgoftaar, T. M. (**2019**). "A survey on image data augmentation for deep learning," J. Big Data **6**(1), 60.

Silva, G. F., Carneiro, G. B., Doth, R., Amaral, L. A., and de Azevedo, D. F. (**2018**). "Near real-time shadow detection and removal in aerial motion imagery application," ISPRS J. Photogramm. Remote Sens. **140**, 104–121.

Stähle, L., and Wold, S. (**1989**). "Analysis of variance (ANOVA)," Chemom. Intell. Lab. Syst. **6**(4), 259–272.

Vedaldi, A., and Soatto, S. (**2008**). "Quick shift and kernel methods for mode seeking," in *Computer Vision–ECCV 2008. ECCV 2008,* Lecture Notes in Computer Science, edited by D. Forsyth, P. Torr, and A. Zisserman (Springer, Berlin, Germany), Vol. 5305, pp. 705–718.

Von Eschenbach, W. J. (**2021**). "Transparency and the black box problem: Why we do not trust AI," Philos. Technol. **34**(4), 1607–1622.

Xie, B., Zhang, H., and Wang, W. (**2025**). "Side-scan sonar image classification based on joint image deblurring–denoising and pre-trained feature fusion attention network," Electronics **14**(7), 1287.

Yang, Z., Zhao, J., Zhang, H., Yu, Y., and Huang, C. (**2023**). "A side-scan sonar image synthesis method based on a diffusion model," J. Mar. Sci. Eng. **11**(6), 1103.

Zhang, P., Tang, J., Zhong, H., Ning, M., Liu, D., and Wu, K. (**2022**). "Self-trained target detection of radar and sonar images using automatic deep learning," IEEE Trans. Geosci. Remote Sens. **60**, 1–14.

Zhu, M., Song, Y., Guo, J., Feng, C., Li, G., Yan, T., and He, B. (**2017**). "PCA and kernel-based extreme learning machine for side-scan sonar image classification," in *Proceedings of 2017 IEEE Underwater Technology (UT)*, Busan, South Korea (February 21–24) (IEEE, New York), pp. 1–4.