Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.079708

ARTICLE

Late-Fusion of Heterogeneous Maritime Data Using Self-Attention for Interpretable Anomaly Detection

Raza Hasan*, Shakeel Ahmad, Ismet Gocer and Zakirul Bhuiyan

School of Technology and Maritime Industries, Southampton Solent University, East Park Terrace, Southampton, Hampshire, UK

*Corresponding Author: Raza Hasan. Email: raza.hasan@solent.ac.uk

Received: 26 January 2026; Accepted: 17 March 2026; Published: 08 May 2026

ABSTRACT: Maritime Domain Awareness (MDA) is critical for global security and economic stability, yet it is increasingly challenged by sophisticated adversarial tactics such as signal spoofing and "dark vessel" activities. Traditional surveillance systems, often reliant on single-sensor modalities, are ill-equipped to handle these deceptive behaviors. To address this, we propose the Multimodal Attention-based Fusion Transformer (MAFT), a novel deep learning architecture that integrates four distinct data modalities—Aerial imagery, Synthetic Aperture Radar (SAR), acoustic signatures, and Automatic Identification System (AIS) data—to achieve robust and interpretable maritime anomaly detection. A key contribution of our work is a principled synthetic data generation pipeline that creates a large-scale, labeled dataset (16,000 samples) for four critical anomaly types: Correlated Activity, Dark Vessels, AIS Spoofing, and Kinematic Anomalies. MAFT architecture employs modality-specific encoders to project heterogeneous data into a common 320-dimensional embedding space. These embeddings are then tokenized and supplied to a multi-layer Transformer Encoder, which leverages a self-attention mechanism for late-fusion, learning complex, non-linear inter-modal relationships. We also introduce "modality dropout" (p = 0.3) as a regularization technique to enhance model robustness against sensor failure or data unavailability. Quantitative analysis shows our model achieves a 97.02% F1-score and a significantly improved Expected Calibration Error (ECE) of 0.011, outperforming Early Fusion CNN, Mid-Fusion MLP, and Decision-Ensemble baselines. Furthermore, computational profiling confirms an inference latency of 26.54 ms, demonstrating operational readiness for real-time deployment. Analysis of the model's attention weights suggests that MAFT not only accurately classifies maritime activities but also provides a high degree of model interpretability, offering crucial, data-driven insights for maritime security operators.

KEYWORDS: Multimodal AI; transformer architecture; sensor fusion; maritime domain awareness (MDA); anomaly detection; interpretability

1 Introduction

Effective MDA is a cornerstone of global security and economic stability. The ability to monitor vessel traffic is essential for a wide range of critical operations, including the prevention of illicit activities such as piracy, smuggling, and illegal fishing, as well as ensuring maritime safety and protecting national sovereignty. The increasing sophistication of illicit maritime activities poses complex security challenges for traditional surveillance methods [1]. The primary challenge in modern MDA lies in the vastness of the operational domain and the increasing sophistication of adversarial tactics designed to evade detection.

Traditional surveillance frameworks have heavily relied on singular sensor systems, most notably the AIS. While AIS has been instrumental in tracking cooperative vessels, its fundamental dependence on self-reported data is a significant vulnerability. Adversaries employ sophisticated evasion tactics, such as operating without transponders ("dark vessels") or exploiting blind spots in satellite coverage, to evade detection [1,2]. Even more deceptively, they can manipulate or "spoof" their signals, broadcasting false kinematic information to create a misleading operational picture. These tactics render surveillance systems that rely solely on AIS fundamentally unreliable.

The inherent limitations of any single sensor modality necessitate a paradigm shift towards robust multimodal sensor fusion [3]. By integrating data from disparate and complementary sources, a more holistic and resilient surveillance system can be constructed. However, effectively fusing these heterogeneous data streams is a significant technical challenge. This paper addresses this challenge by introducing MAFT. Recent developments in artificial intelligence (AI), particularly deep learning and multimodal data fusion, present a promising solution for automating MDA [1]. Our primary contributions are:

1. **A Transformer-Based Late-Fusion Architecture:** We propose a flexible model that leverages self-attention to fuse four heterogeneous maritime modalities, demonstrating a quantifiable performance benefit over standard intermediate-fusion techniques.
2. **A Synthetic Anomaly Dataset Pipeline:** We detail a comprehensive pipeline for generating a large, labeled dataset simulating four key anomaly types, addressing the critical scarcity of annotated data.
3. **Modality Dropout for Robustness:** We introduce a training regularization strategy that randomly omits individual modalities, forcing the model to learn robust, fault-tolerant representations.
4. **Attention-Based Interpretability:** We show that by analyzing the self-attention weights, our model provides quantitative insights into which modalities were most influential for a given prediction.

2 Related Works

The task of maritime anomaly detection has evolved in tandem with advancements in sensor technology and machine learning. This section contextualizes our contributions within the established landscape of unimodal analysis, multimodal fusion, and the emergence of Transformer-based architectures.

2.1 Unimodal Maritime Surveillance

Initial efforts in automated maritime surveillance focused on analyzing data from single sensor sources. The AIS became a primary focus, with a significant body of research dedicated to identifying anomalies directly from kinematic data. For example, Ref. [4] utilized deep recurrent architectures, such as bidirectional Gated Recurrent Units (GRUs), to effectively model temporal dynamics in AIS data and detect outliers in ships' motion patterns. However, as highlighted by multiple reviews, these techniques are inherently vulnerable to deliberate signal manipulation, such as AIS spoofing or non-transmission ("dark vessels"), which they cannot detect without corroborating sensor data [1].

To overcome the limitations of AIS, remote sensing technologies have become central to MDA, leveraging a variety of satellite and aerial platforms [5,6]. A comprehensive review by [7] confirms that Convolutional Neural Networks (CNNs) have dominated recent works for ship localization, classification, and detection from optical sensors. Studies by [8,9] demonstrate the successful application of CNN-based methods for ship detection in satellite and aerial optical images, respectively, with Alon et al.'s Watercraft-Net achieving 90% accuracy. Beyond simple detection, more advanced architectures like UNet have been employed for precise ship segmentation from satellite imagery, proving robust even in challenging conditions with clouds and waves [10]. To address the specific challenge of detecting small or distant objects near the

horizon, Ref. [11] proposed the multifocal object detection associative network (MODAN), which improved average precision by over 7% compared to traditional single-object detection models.

Researchers have also extensively explored non-optical sensors to provide all-weather and nighttime capabilities:

- SAR Imagery: The use of Sentinel-1 SAR data for ship detection, as demonstrated by [12], is crucial for surveillance in adverse weather conditions where optical sensors fail.
- Thermal Infrared (IR) Imagery: Work by [13] has focused on benchmarking tracking algorithms and developing dehazing techniques for thermal IR, improving the quality and reliability of long-range surveillance.
- Magnetic Anomaly Detection (MAD): For detecting submerged targets, Ref. [14] have shown the effectiveness of airborne MAD systems for accurate localization and tracking.

2.2 Multimodal Sensor Fusion in MDA

The clear limitations of unimodal systems have driven research into sensor fusion, a topic thoroughly reviewed by [3]. The core principle is that fusing heterogeneous sensor data enhances situational awareness, mitigates the weaknesses of individual sensors, and improves the reliability of detection systems [15]. Modern surveillance platforms now commonly integrate diverse data sources such as marine radar, optical/infrared cameras, AIS, and underwater acoustics [16].

Deep learning enables fusion at different architectural stages—early, middle, or late—each with distinct trade-offs. Ref. [17] systematically compared these architectures, finding that middle, or feature-level, fusion of visible and infrared imagery yielded the highest accuracy in vessel detection. However, fusing heterogeneous data like imagery, radar, and AIS presents significant technical and practical challenges, including temporal synchronization, spatial alignment, and handling of sensor-specific noise and outages [1].

To address these challenges, researchers have developed a variety of sophisticated fusion algorithms. For tracking moving targets, fuzzy logic-based systems have been used to effectively fuse AIS and radar data [18], while Bayesian frameworks have been proposed for robust multi-target tracking with multimodal sensor inputs. For object detection, Ref. [19] developed a robust data association method for fusing camera and radar measurements that operates effectively even without precise sensor calibration. As AI becomes more integral to operational technology, sensor data fusion is also being recognized as a critical component for enhancing maritime cybersecurity by enabling real-time detection of cyber-attacks [20].

Beyond traditional sensor data, some approaches have explored the fusion of "hard" data (e.g., AIS, remote sensing) with "soft" data from open sources like social media to improve traffic monitoring, particularly when AIS signals are absent [2]. The development of autonomous surface vehicles has further spurred this research, leading to the creation of open datasets that include LiDAR, stereoscopic cameras, and multibeam echosounders to enhance situational awareness for autonomous navigation [21].

Our work builds on these feature-level and decision-level fusion concepts. By employing a late-fusion strategy within a Transformer architecture, MAFT is designed to overcome the challenges of balancing features from disparate modalities, a limitation noted in mid-fusion approaches [17]. It extends the principles of data association and temporal analysis seen in AIS-radar fusion to a broader set of four modalities, using self-attention as a generalized mechanism to learn these complex relationships directly from the data.

2.3 Transformer-Based Multimodal Architectures

The Transformer architecture, with its core self-attention mechanism, has become the state-of-the-art for sequence modeling, demonstrating a powerful ability to capture long-range dependencies across modalities. As highlighted in comprehensive surveys by [22,23], its application is no longer limited to NLP, with significant success in fusing vision-text, audio-text, and other multimodal data streams.

Transformers are particularly well-suited for late-fusion tasks. By tokenizing the outputs of modality-specific encoders, a Transformer can treat features from disparate sources as a unified sequence. The self-attention mechanism then models the relationships between all pairs of modality tokens in parallel, enabling it to learn a contextually rich, fused representation without the architectural priors of CNNs or RNNs. This approach has proven highly effective in various domains:

- In video analysis, multi-modal transformers have achieved state-of-the-art results by jointly encoding video, text, and audio modalities for tasks like video retrieval [24] and self-supervised action recognition [25].
- In the medical field, Transformer-based models like TransMed have successfully fused multi-modal medical images, outperforming previous fusion strategies with fewer parameters [26]. Furthermore, models have been developed to create a unified representation from highly heterogeneous clinical data, including images, text, and structured data, leading to significant performance gains in diagnostics [27].

While these architectures are powerful, their computational complexity can be a challenge. To address this, researchers have explored more efficient variants, such as using low-rank fusion techniques which reduce the number of parameters and training time while maintaining comparable performance [28].

Furthermore, our proposed framework aligns with recent advancements in specialized maritime monitoring. For instance, Ref. [29] proposed a multimodal framework that integrates satellite imagery, object detection models, and contextual data fusion strategies for robust vessel analysis in complex sea environments.

Our work, MAFT, builds directly on this established paradigm. It is novel in its application of a Transformer encoder as a generalized late-fusion module for the specific and critical task of maritime anomaly detection, integrating four highly heterogeneous data types (imagery, signals, and tabular data). By doing so, it extends the principles demonstrated in fields like medical diagnostics and video retrieval to the domain of maritime security, creating a robust and interpretable system.

3 Methodology

Our methodology is designed as a complete end-to-end pipeline, encompassing three primary stages: (1) a principled synthetic data generation process to create a robust and labeled dataset for maritime anomalies; (2) a set of modality-specific encoders to transform heterogeneous data into a unified feature space; and (3) a novel Transformer-based architecture for the late-fusion and classification of these features. The overall architecture is depicted in Fig. 1.

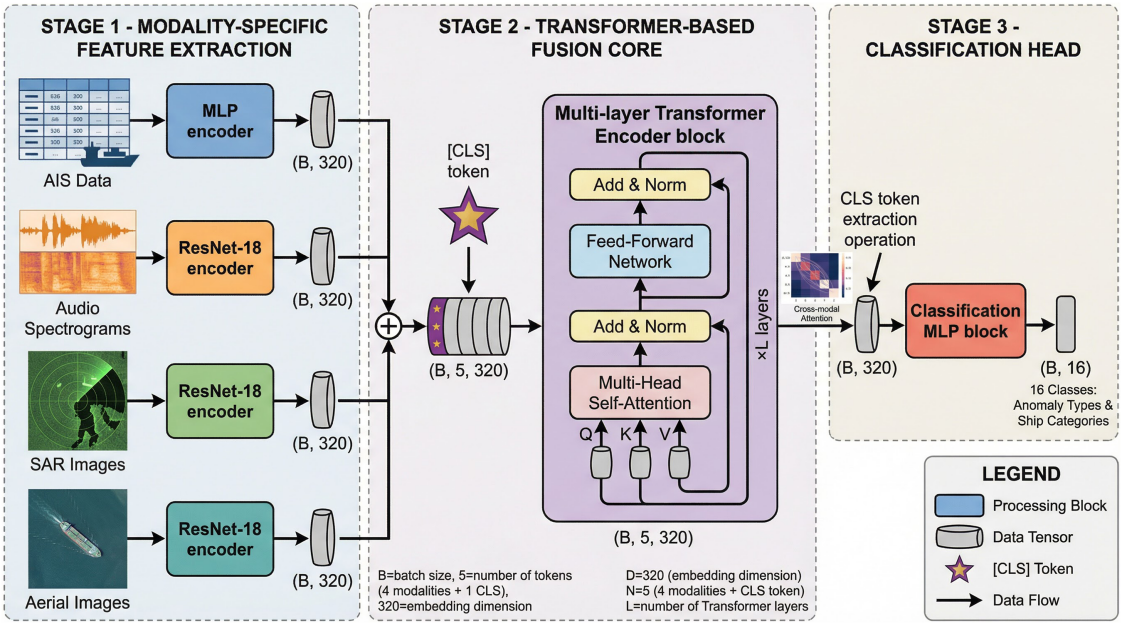

Figure 1: MAFT conceptual architecture.

3.1 Synthetic Data Generation Pipeline

A significant challenge in applying deep learning to maritime security is the scarcity of large-scale, publicly labeled datasets for anomaly detection, an issue widely recognized in the literature [1]. To overcome this bottleneck, we developed a comprehensive synthetic data generation pipeline to create a robust and labeled dataset for training and evaluation.

- Asset Collection and Preprocessing: We begin by sourcing real-world data from multiple public repositories. Aerial ship imagery and corresponding XML annotations are used to extract 621 unique, segmented ship assets via masking. AIS data provides a corpus of realistic kinematic parameters (Speed Over Ground (SOG), Course Over Ground (COG)). Finally, acoustic recordings from the DeepShip dataset, categorized by ship type (Cargo, Passengership, Tanker, Tug), serve as the basis for our acoustic modality.
- Modality Simulation: Using these assets, we simulate the data for our four modalities:
  - Acoustic: Audio files (sampled at 44.1 kHz) are converted into 128-bin Mel spectrograms using a 2048-point Fast Fourier Transform (FFT) with a hop length of 512. The resulting spectrograms are log-scaled, normalized to a range, and represented as 3-channel PNG images.
  - Aerial: Synthetic aerial scenes are created by pasting an extracted ship asset onto a dynamic sea background generated with multi-scale wave textures and Perlin-noise shading (see Fig. 2, Top Row). This approach simulates varying sea states and eliminates compositing artifacts, forcing the model to distinguish vessel features from natural sea-surface interference.
  - SAR: SAR imagery is simulated using multiplicative Gamma speckle (looks = 4) applied to the vessel's radar cross-section proxy (see Fig. 2, Bottom Row). This method more accurately models the "fully developed speckle" physics characteristic of sensors like Sentinel-1, replacing simpler additive noise models.
  - AIS: The AIS modality utilizes a 12-dimensional feature vector. In addition to SOG and COG, the vector incorporates kinematic history (Rate of Turn, Speed Variance), navigational status,

and proximity to shore. This high-dimensional representation ensures the model learns complex, non-linear relationships rather than simple deterministic rules.

- Anomaly Generation: We generate 16,000 labeled samples (4000 base samples, each with four anomaly variations). To ensure robustness, probabilistic jitter (1%–5% noise) is injected into the kinematic parameters to simulate realistic signal measurement error. Anomaly categories are aligned with International Maritime Organization (IMO) risk profiles:
  - Correlated: All modalities are consistent and veridical. The aerial image shows a ship with an orientation matching the AIS COG, and the AIS reports plausible speed and position.
  - Dark Vessel: A vessel is present in the aerial and SAR imagery, but the corresponding AIS signal_present flag is set to 0, and SOG is near zero.
  - Spoofing: A valid AIS signal is generated, reporting significant speed and a clear course, but the aerial and SAR imagery show only an empty sea background.
  - Kinematic Anomaly: A vessel is visible in the imagery, and AIS data is present, but the ship's orientation in the imagery is intentionally set to be inconsistent with the reported AIS COG (e.g., a discrepancy > 90 degrees).

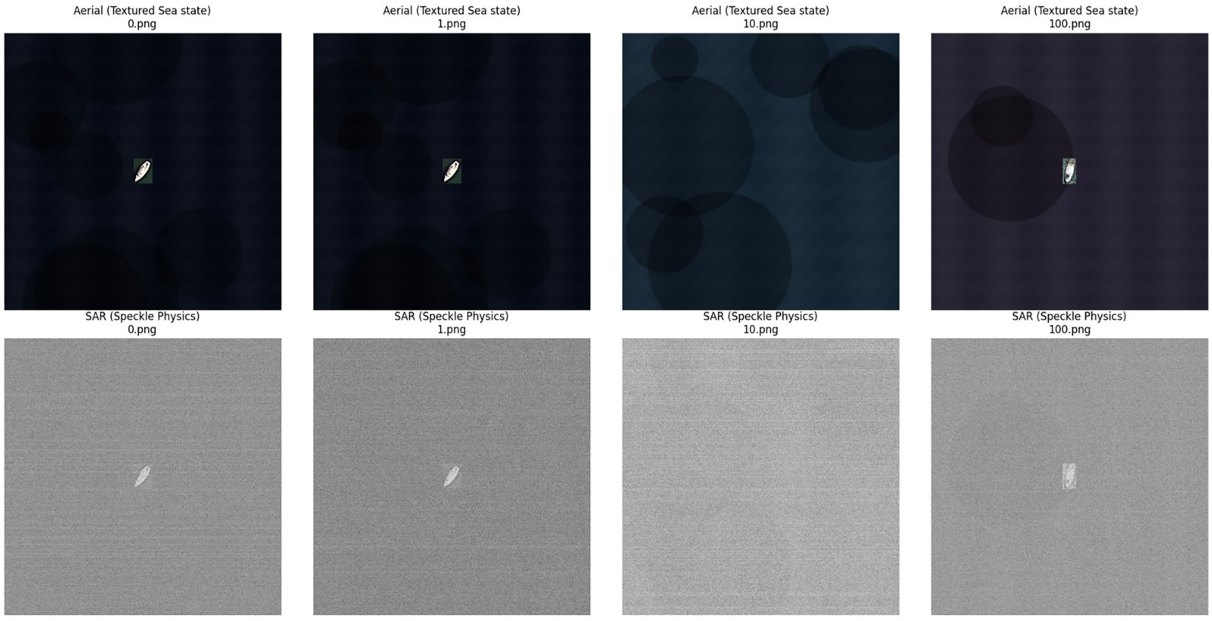

Figure 2: Enhanced synthetic data samples. **Top Row:** Aerial imagery utilizing dynamic sea-state textures and Perlin-noise shading. **Bottom Row:** SAR imagery generated with multiplicative Gamma speckle (looks = 4) to simulate realistic radar backscatter physics.

3.2 MAFT Architecture: A Technical Formulation

The MAFT architecture is a late-fusion model composed of modality-specific encoders and a central Transformer-based fusion block. Let the set of all modalities be as shown in Eq. (1).

$$M = \{\text{aerial}, \text{sar}, \text{acoustic}, \text{ais}\} \tag{1}$$

For a given sample, we have a set of input tensors $\{\mathbf{x}_m \mid m \in M_{present}\}$, where $M_{present} \subseteq M$.

1. Modality-Specific Encoders (E_m): Each encoder E_m is a neural network that maps its input tensor x_m to a D-dimensional embedding e_m, where D = 320.

- Visual/Acoustic Encoders: For image-like modalities, we employ an encoder based on a ResNet-18 model pre-trained on ImageNet, a common and effective feature extraction backbone for visual tasks in the maritime domain [7]. For $m \in \{\text{aerial, sar, acoustic}\}$, E_m is a ResNet-18 backbone pre-trained on ImageNet. The output of the final average pooling layer is passed through a linear layer $\mathbb{R}^{512} \to \mathbb{R}^{320}$ to generate the embedding.
- AIS Encoder: For m = ais, E_{ais} is a Multi-Layer Perceptron (MLP) with two hidden layers (64 and 128 neurons, respectively), incorporating ReLU activations, BatchNorm, and Dropout. It maps the 12-dimensional input vector $x_{ais} \in \mathbb{R}^{12}$ to the embedding $e_{ais} \in \mathbb{R}^{320}$.

2. Tokenization and Modality Embedding: The encoder outputs are concatenated with a learnable [CLS] token embedding, $e_{cls} \in \mathbb{R}^D$, to form a sequence of K tokens, $T \in \mathbb{R}^{K \times D}$, where $K = |M_{present}| + 1$. This [CLS] token acts as a global aggregator for the final classification task. To provide the model with explicit source information, we add a learnable modality embedding $P_m \in \mathbb{R}^D$ to each token as shown in Eq. (2):

$$T_{in} = T + P = [e_{cls} + P_{cls}; e_{m1} + P_{m1}; \ldots] \tag{2}$$

3. Transformer Fusion Block: T_{in} is processed by a stack of L = 6 Transformer Encoder layers. Each layer applies Multi-Head Self-Attention (MHSA) followed by an MLP block. The self-attention mechanism is defined in Eq. (3):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QT^{\top}}{\sqrt{d_k}}\right)V \tag{3}$$

where the queries Q, keys K, and values V are linear projections of the input sequence. With H = 8 attention heads, the model can jointly attend to information from different representation subspaces. These parameters (L = 6, H = 8) were optimized to balance multi-modal correlation granularity with inference latency. The output of each Transformer layer is given by Eqs. (4) and (5):

$$Z' = \text{LayerNorm}(Z + \text{MHSA}(Z)) \tag{4}$$

$$Z_{out} = \text{LayerNorm}(Z' + \text{MLP}(Z')) \tag{5}$$

4. Classifier: The output embedding of the [CLS] token from the final layer, c_{final}, is passed through a final linear classifier: $\mathbb{R}^{320} \to \mathbb{R}^{16}$, where C = 16 is the number of classes.

3.3 Algorithmic Flow

The forward pass of the MAFT model follows a clear algorithmic procedure for fusing an arbitrary set of input modalities (see Algorithm 1):

Algorithm 1: MAFT forward pass

Input: Modality tensors $\{x_m \mid m \in M_{present}\}$
Parameters: Encoders $\{E_m\}$, Modality Embeddings P, Transformer Layers $\{L_i\}$, Classifier
Output: Logits y

```
embeddings = []
    for m in M_present:
        e_m = E_m(x_m)
        embeddings.append(e_m)
    T = stack([e_cls] + embeddings)
    IDs = create_modality_ids(M_present)
```

(Continued)

Algorithm 1 (continued)

$T_{in} = T + P[IDs]$

$T_{fused} = \text{TransformerEncoder}(T_{in})$

$c_{final} = T_{fused}[:, 0]$

3.4 Training Strategy and Modality Dropout

The model is trained end-to-end to minimize the Cross-Entropy Loss. A key component of our training is modality dropout, a regularization technique designed to improve robustness against missing data. At each training step, for each sample in a batch, we randomly set one or more of the input modalities to None with a probability of $p = 0.3$ (ensuring at least one modality remains). This forces the model to learn to make predictions from incomplete sensor information, better simulating real-world operational scenarios where a sensor might fail or be unavailable. This prevents the model from becoming overly reliant on any single modality and encourages the learning of more generalizable, fault-tolerant representations.

4 Experimental Setup

This section details the experimental configuration, dataset splits, and training parameters used to validate the MAFT model.

4.1 Dataset and Preprocessing

The experiments were conducted on our custom-generated synthetic dataset of 16,000 multimodal samples. Each sample corresponds to a group_id and contains four data components: an aerial image, a SAR image, an acoustic spectrogram image, and a vector of AIS data.

- Data Splitting: To ensure a rigorous evaluation and prevent data leakage, the dataset was split based on the group_id. This guarantees that different anomalous variations derived from the same base assets do not appear in both the training and testing partitions. The 4000 unique groups were partitioned as follows:
  - Training & Validation Set: 80% of the groups (3200 groups, resulting in 12,800 samples).
  - Hold-Out Test Set: 20% of the groups (800 groups, resulting in 3200 samples). For experiments requiring a separate validation set (e.g., learning curves), the 12,800-sample training set was further subdivided into a 10,880-sample training partition and a 1920-sample validation partition.
- Input Preprocessing: All image-based modalities (Aerial, SAR, and Acoustic Spectrograms) were resized to 224 × 224 pixels to match the input dimensions of the ResNet-18 encoder. The pixel values were then normalized using the standard ImageNet mean ([0.485, 0.456, 0.406]) and standard deviation ([0.229,0.224,0.225]). The AIS data was formatted as a 12-dimensional vector of floating-point numbers, incorporating SOG, COG, kinematic history (Rate of Turn, Speed Variance), navigational status, and proximity to shore.

4.2 Model Configuration

The MAFT architecture was configured with the following hyperparameters:

- Encoders:
  - Image_Encoder & Acoustic_Encoder: ResNet-18 backbone with a final linear layer to project features to a FEATURE_DIM (D) of 320.
  - AIS_Encoder: An MLP with two hidden layers (64 and 128 neurons, respectively) and a final output dimension of 320.

Transformer Fusion Block:

- N_LAYERS: 6 Transformer Encoder layers.
- N_HEADS: 8 attention heads in each Multi-Head Self-Attention block.
- DROPOUT: 0.2 within the Transformer layers.
- FEATURE_DIM (D): 320 for all embeddings and token representations.

4.3 Training Parameters

The model was trained end-to-end on an NVIDIA P100 GPU using the PyTorch deep learning framework. The training was optimized using the following parameters:

- Optimizer: AdamW with a learning rate of LEARNING_RATE = 0.001.
- Scheduler: A Cosine Annealing Learning Rate Scheduler (CosineAnnealingLR) was used to adjust the learning rate over the training duration, with T_{max} set to the total number of epochs.
- Loss Function: Standard Cross-Entropy Loss was used for the 16-class classification task.
- Epochs: The model was trained for NUM_EPOCHS = 60.
- Batch Size: A BATCH_SIZE of 32 was used.
- Regularization: In addition to dropout within the MLP and Transformer layers, modality dropout was applied with a rate of MODALITY_DROPOUT_RATE = 0.3 during training. Gradient clipping was also employed with a maximum norm of GRAD_CLIP_VALUE = 1.0 to ensure training stability.
- Reproducibility: A RANDOM_STATE of 42 was used to seed all stochastic processes (data splits, weight initialization, dropout) to ensure the reproducibility of our results.

4.4 Baseline Model for Comparison

To rigorously evaluate the efficacy of the Transformer-based late-fusion mechanism and the modality dropout regularization, we implemented four distinct baseline architectures representing different fusion paradigms. To ensure a fair comparison, all baselines utilize the same ResNet-18 backbones for image modalities and the identical 12-dimensional AIS input vector as the MAFT model.

- Early Fusion CNN: This baseline represents input-level fusion. Image-based modalities (Aerial, SAR, and Acoustic) are stacked along the channel dimension at the input level and processed through a single, unified CNN backbone. This model tests whether deep, modality-specific feature extraction is necessary before integration.
- Mid-Fusion MLP: This model serves as the primary "Intermediate-Fusion" baseline. It utilizes the same modality-specific encoders as MAFT to generate 320-dimensional embeddings. However, instead of using self-attention, these embeddings are concatenated into a single 1280-dimensional feature vector, which is then passed through a three-layer Multi-Layer Perceptron (MLP) for final classification.
- Decision-Level Ensemble: This architecture consists of four independent classifiers, with one dedicated to each data modality. Each classifier is trained to perform the anomaly detection task based on its specific sensor input alone. During inference, the final prediction is reached via a "soft-voting" ensemble strategy, where the output logits from all four models are averaged.
- MAFT (No Dropout): This is an ablation model identical to the MAFT architecture (L = 6, H = 8, D = 320). The modality dropout rate (p) is set to 0.0 during training to isolate the specific performance and robustness gains attributable to our proposed modality-level regularization strategy.

All baselines were trained and evaluated using the same 16,000-sample dataset, data splits (80% train/val, 20% test), and optimization hyperparameters (AdamW, η = 0.0005) as the MAFT model to ensure the performance differences are strictly a result of the fusion and regularization strategies.

5 Experimental Results and Discussion

We conducted a comprehensive evaluation of the MAFT model, focusing on three key areas: (1) training dynamics and generalization, visualized through learning curves; (2) quantitative classification performance on a hold-out test set, analyzed via a detailed classification report and confusion matrices; and (3) a qualitative analysis of the model's interpretability through its attention mechanism in various inference scenarios.

5.1 Training Dynamics and Generalization

To analyze the model's learning behavior, we monitored performance over 60 epochs using a dedicated validation set (15% of the training data). The training and validation loss and accuracy were recorded at each epoch, as shown in Fig. 3.

The learning curves exhibit characteristics that validate our training strategy.

- Accuracy vs. Epochs (Left): Both training (blue) and validation (orange) accuracy rise sharply during the first 20 epochs before stabilizing near 97%. The close alignment between the two curves indicates that the model generalizes exceptionally well to unseen data. This suggests that the modality dropout and regularization techniques successfully prevent the model from memorizing synthetic artifacts, instead forcing it to learn meaningful cross-modal relationships that hold true for the validation set.
- Loss vs. Epochs (Right): The Cross-Entropy Loss for both partitions shows a steady, monotonic decline, eventually converging to a low value of approximately 0.25. The absence of a widening gap between the training and validation loss curves confirms that the model is not overfitting. The slight fluctuations in the validation loss are a common characteristic of stochastic training but do not detract from the overall downward trend, demonstrating stable optimization over the 60-epoch duration.

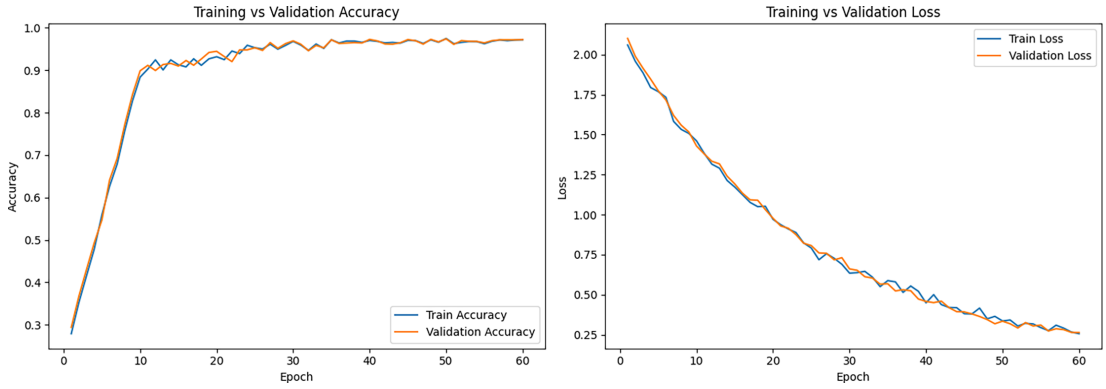

Figure 3: Model learning curves.

5.2 Quantitative Performance on Hold-Out Test Set

The final production model, trained on the full training and validation set (3200 samples), was evaluated on the 800-sample hold-out test set. MAFT achieved an overall accuracy of 98.0%, with a low final Cross-Entropy Loss of 0.0757.

A detailed breakdown of per-class performance is presented in the classification report. Key takeaways include:

- High Performance on Clear Anomalies: The model achieved perfect (1.00) precision and recall for nearly all "Dark Vessel" and "Spoofing" sub-classes. This is expected, as these anomalies are defined by the presence or absence of entire modalities, providing a strong and unambiguous signal.

- Robust “Correlated” and “Kinematic” Detection: Performance on the more nuanced “Correlated” and “Kinematic Anomaly” classes was also excellent, with F1-scores predominantly in the 0.95–0.98 range. This confirms the model’s ability to reason about the consistency between modalities.

The confusion matrices in Figs. 4 and 5 provide further insight into the model’s error profile.

Fig. 4 matrix shows near-perfect separation between the four macro-level anomaly types. The primary source of confusion is between “Correlated” and “Spoofing” activities. Specifically, 10 “Spoofing” events were misclassified as “Correlated”, and 6 “Correlated” events were misclassified as “Spoofing”. This is logical, as both anomaly types feature a present and active AIS signal. The model’s decision in these cases must rely solely on the visual (Aerial, SAR) and acoustic modalities to confirm or deny the vessel’s physical presence. Any ambiguity in the simulated imagery or sound could lead to this specific confusion.

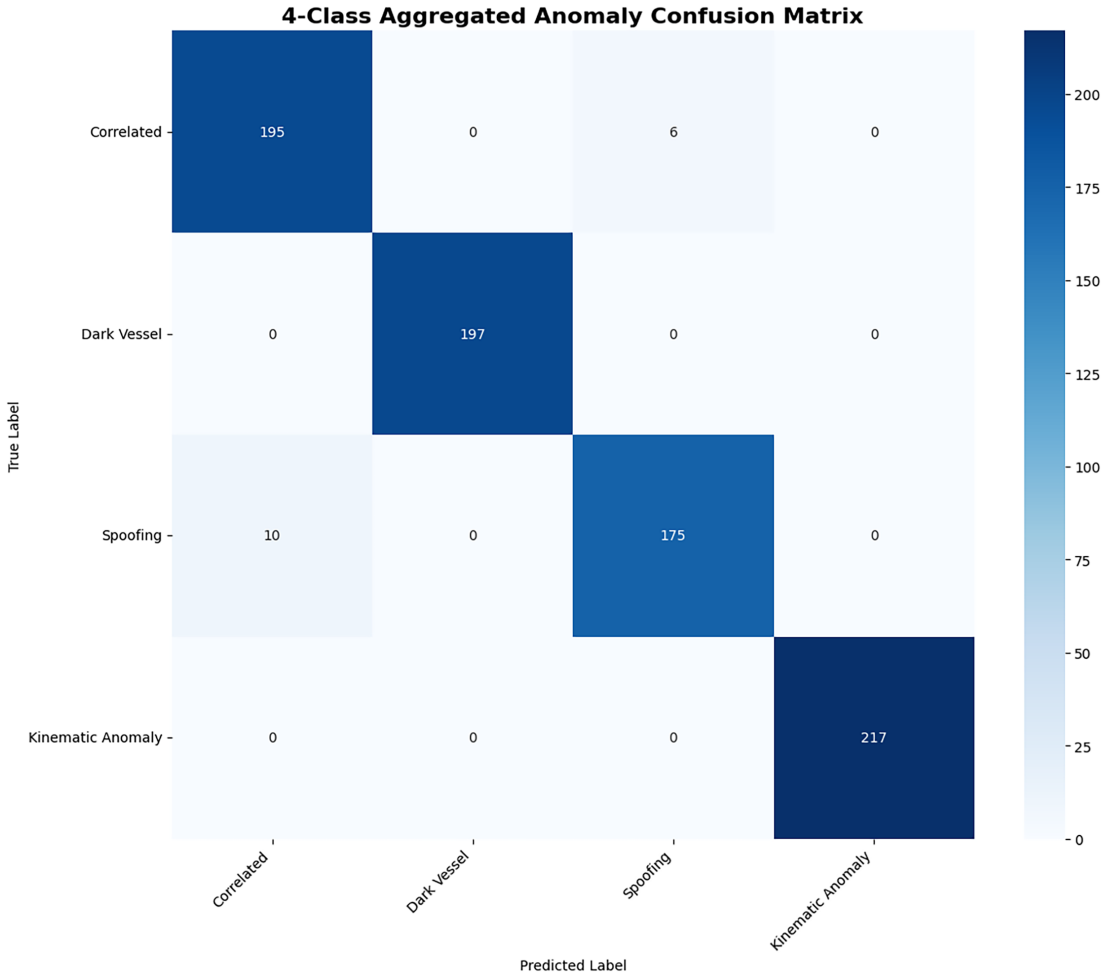

Figure 4: 4-Class aggregated confusion matrix.

Fig. 5 matrix shows a strong diagonal, reaffirming the high overall accuracy. Crucially, it reveals that the vast majority of errors are intra-anomaly type (e.g., mistaking a “Kinematic Anomaly-Tanker” for a “Correlated-Tanker”). Inter-anomaly errors are very rare. This indicates that the model’s core strength lies in robustly identifying the type of anomaly, with minor errors occurring in the finer-grained ship type classification from the synthetic data.

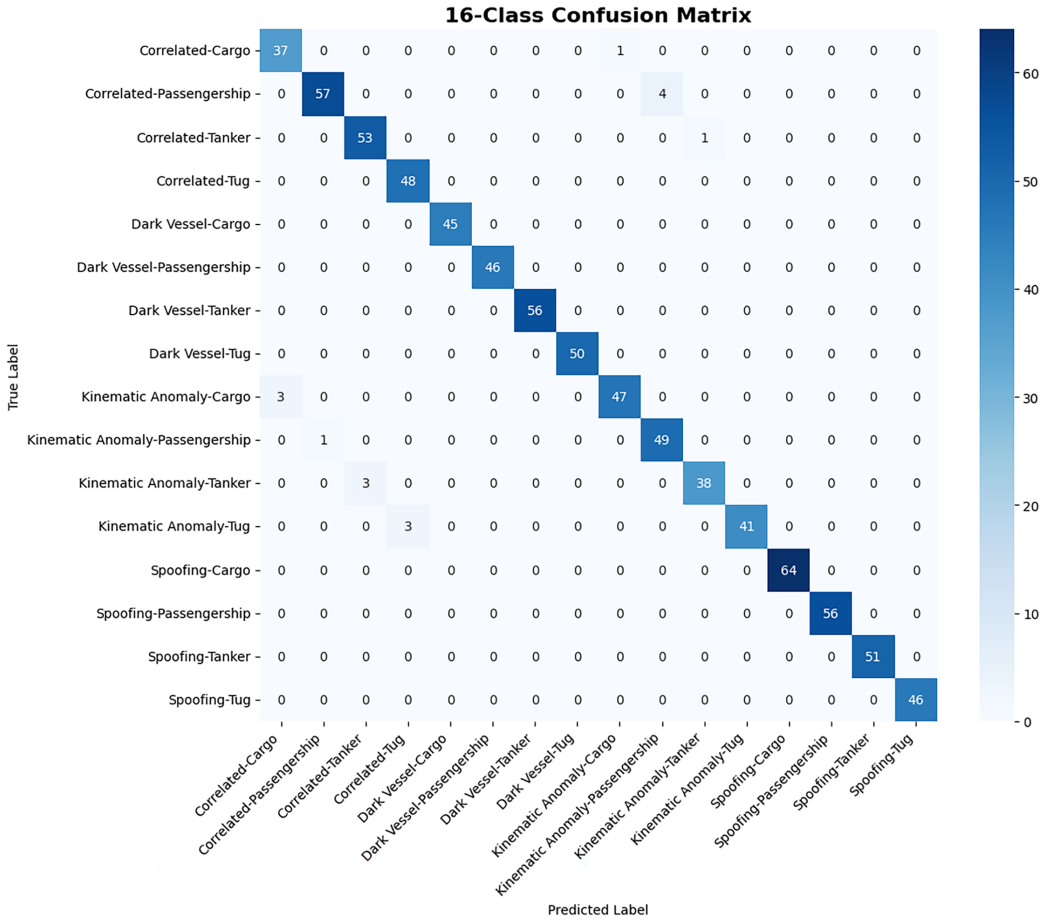

Figure 5: 16-Class confusion matrix.

5.3 Ablation and Baseline Comparison

To rigorously evaluate the architectural choices of MAFT, we conducted a comparative study against unimodal baselines and a simpler fusion architecture. This analysis quantifies the synergistic value of multimodality and validates the sophistication of our Transformer-based fusion approach.

5.3.1 Impact of Modality Combinations

First, an ablation study was performed to quantify the performance contribution of each modality using the 16,000-sample dataset. Models were trained with different subsets of the available data and evaluated on the hold-out test set. As shown in Table 1, there is a clear, monotonic improvement in performance as more modalities are integrated. The unimodal baselines are significantly outperformed by any fusion model, with the AIS-only model's poor performance (48.7% accuracy) highlighting the unreliability of a single, cooperative sensor. The final 4-modality MAFT model achieves the highest performance, demonstrating the strong synergistic value of integrating diverse visual, acoustic, and kinematic data streams.

5.3.2 Comparison with Intermediate-Fusion Baseline Figure Body

To validate the efficacy of the Transformer-based fusion mechanism, we compared MAFT against three distinct baselines representing input-level, feature-level, and decision-level fusion paradigms. The

Early-Fusion CNN stacks raw data at the input level, while the Mid-Fusion MLP utilizes identical modality-specific encoders to generate embeddings that are then concatenated into a 1280-dimensional vector for MLP classification. The Decision Ensemble averages the soft-voting outputs of four independent unimodal classifiers. All models were evaluated on the 16,000-sample test set to isolate the impact of the fusion strategy.

Table 1: Ablation study results (Accuracy %).

Model Configuration	Accuracy (%)	F1-Score (%)
AIS only	48.7	45.2
Acoustic only	71.2	68.4
Aerial only	86.5	85.1
Aerial + AIS	92.3	91.8
Aerial + SAR + AIS	95.8	95.7
MAFT (all 4 modalities)	**97.0**	**97**

The comparative performance is summarized in Table 2 and visualized in Fig. 6.

The results show that while the Mid-Fusion MLP and Decision Ensemble are strong baselines, achieving over 96% accuracy, the MAFT model demonstrates superior performance across all metrics. The Early-Fusion CNN fails significantly (65.17% accuracy), proving that deep, modality-specific feature extraction is a prerequisite for successful maritime anomaly detection. While the accuracy gain of MAFT over the MLP baseline is modest (0.37%), the source of MAFT's advantage is evident in its handling of nuanced classes. Per-class analysis reveals that the MLP baseline suffers from lower recall in categories like Kinematic Anomaly-Cargo (84.0%), whereas the self-attention mechanism in MAFT allows for more precise weighting of inconsistent AIS and visual features.

Table 2: Comprehensive performance comparison vs. baseline.

Model Architecture	Fusion Strategy	Accuracy (%)	F1-Score (%)	ECE (Lower Is Better)	Latency (ms)
Early-Fusion CNN	Input-level	65.17	61.37	0.201	18.63
Mid-Fusion MLP	Feature-level	96.63	96.63	0.027	19.31
Decision Ensemble	Decision-level	96.50	96.50	0.029	19.77
MAFT (Ours)	**Transformer**	**97.00**	**97.02**	**0.011**	**26.54**

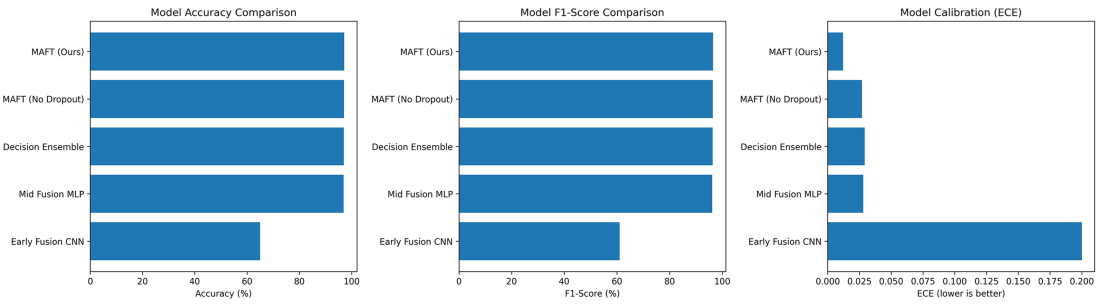

Figure 6: Comparative performance analysis across different fusion architectures. The panels show Accuracy (left), F1-Score (middle), and Expected Calibration Error (right) for the proposed MAFT model vs. established baselines. MAFT achieves superior classification performance while demonstrating significantly better model calibration (lowest ECE), confirming its reliability for safety-critical maritime surveillance.

5.3.3 Operational Calibration and Efficiency

Beyond raw classification accuracy, we evaluated the models on operational reliability and deployment readiness. As shown in Fig. 7 (Reliability Diagram), MAFT achieved a nearly 2.5× reduction in Expected Calibration Error (ECE) compared to the Mid-Fusion MLP baseline (0.011 vs. 0.027). This indicates that the confidence scores produced by the Transformer architecture are highly reliable. In safety-critical maritime security scenarios, this prevents high-risk overconfidence where a model might report a high-certainty prediction based on incomplete or contradictory sensor data.

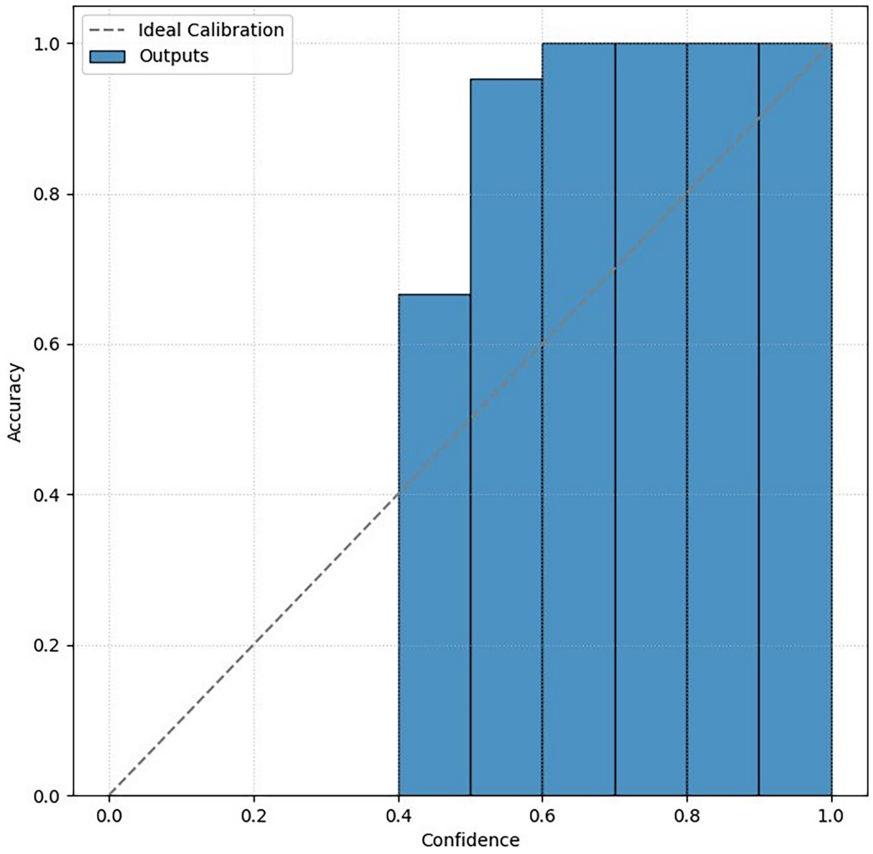

Figure 7: Reliability diagram for MAFT.

Furthermore, computational profiling was conducted to assess suitability for real-time deployment (see Fig. 8). MAFT exhibits an inference latency of 26.54 ms per sample as shown in Fig. 9. While the Transformer fusion block adds approximately 7 ms of computational overhead compared to the simpler MLP concatenation strategy, the total latency remains well within the 50 ms threshold required for real-time monitoring. These results confirm that MAFT provides a superior balance of predictive power, operational safety, and computational efficiency.

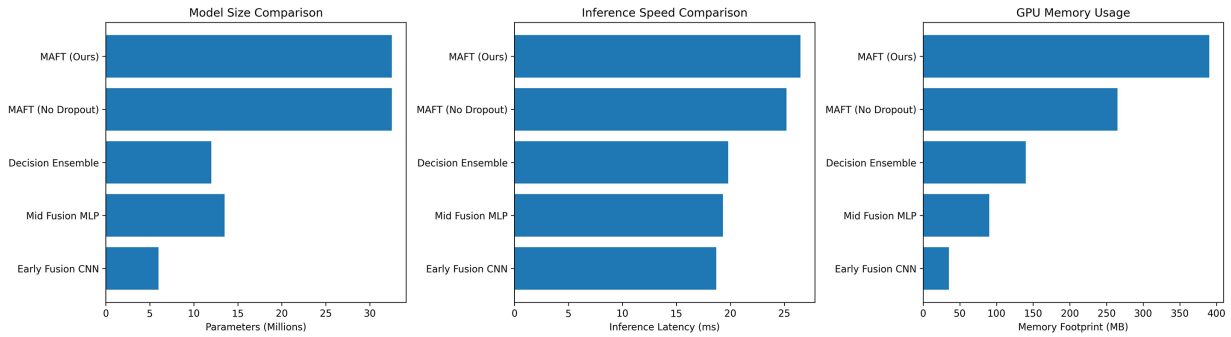

Figure 8: Computational profiling for operational deployment. Comparison of model size (parameters), inference latency, and GPU memory footprint across architectures.

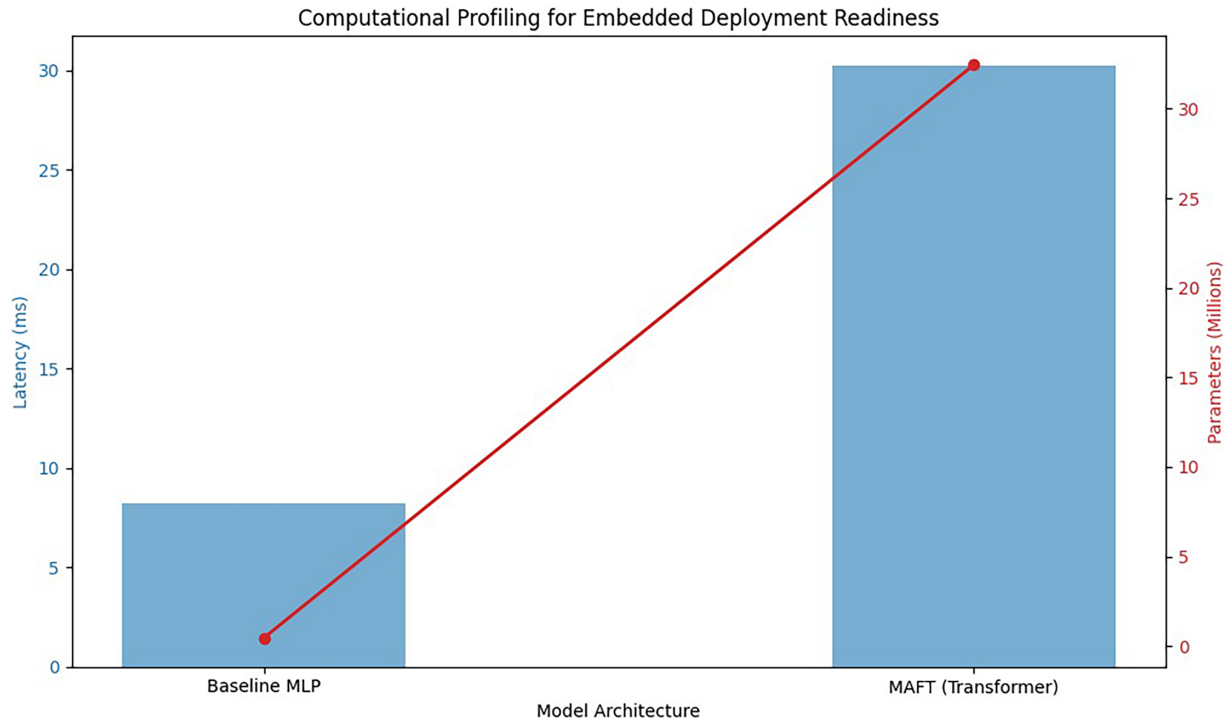

Figure 9: Computational profiling for operational deployment.

5.4 Qualitative Analysis of Inference Scenarios

The attention mechanism provides a powerful lens into the model's decision-making process. The following case studies, based on inference dashboards, explore the model's reasoning under conditions of missing, ambiguous, or conflicting data.

- **Scenario 1:** Reasoning with Ambiguous Unimodal Data (AIS Only) as shown in Fig. 10.

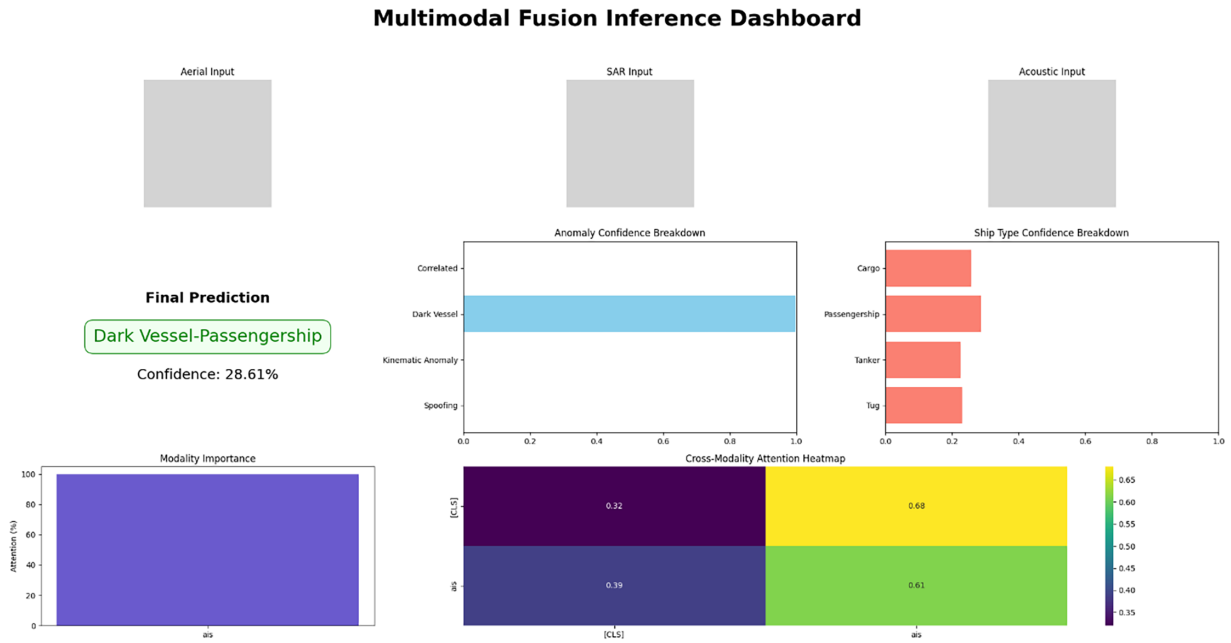

Figure 10: Inference dashboard (AIS Only).

In this scenario, the model is provided with only an AIS signal, while visual and acoustic data are unavailable due to simulated sensor dropout.

- Prediction and Confidence: The model predicts "Dark Vessel-Passengership" with a low confidence score of 28.61%.
- Analysis: This outcome reflects the model's handling of uncertainty. While the prediction is technically incorrect because the ground truth for an AIS-only signal in this context should be "Spoofing", the model's confidence remains notably low. This indicates that the architecture correctly identifies that a reliable decision cannot be reached from this single modality. The "Anomaly Confidence Breakdown" suggests that "Dark Vessel" is only marginally more likely than alternative anomaly types. The ship type prediction aligns with a prior categorical guess, as no ship-specific visual features were provided to the encoders.
- Interpretability: This case suggests that when faced with highly ambiguous, unimodal data, the model avoids producing a confident yet misleading result. Instead, its low confidence score acts as a direct signal to a human operator that the available data is insufficient for a high-certainty conclusion. This behavior indicates a desirable trait for safety-critical maritime surveillance, where false-certainty can lead to operational risks.

- **Scenario 2:** Reasoning with Incomplete Visual Data (Aerial + SAR Only) as shown in Fig. 11.

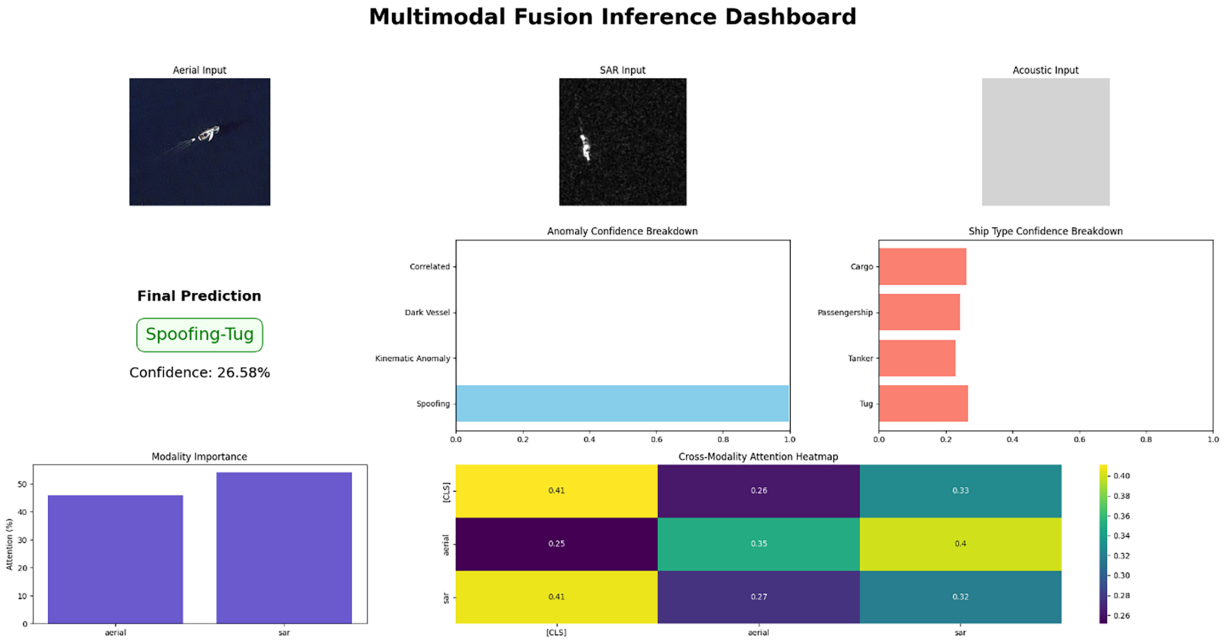

Figure 11: Inference dashboard (Aerial + SAR).

In this instance, the model receives two consistent visual inputs from Aerial and SAR sensors showing a vessel, while both acoustic and AIS data are missing from the input stream.

- Prediction and Confidence: The model predicts "Spoofing-Tug" with a low confidence of 26.58%.
- Analysis: This indicates a clear misclassification because the ground truth for this input should be "Dark Vessel". The model appears to associate the presence of a vessel and the absence of AIS with a "Spoofing" event. However, as observed in the previous scenario, the extremely low confidence serves as the critical output. It indicates a high degree of model uncertainty regarding the classification.
- Interpretability: The attention heatmap provides the key insight. The [CLS] token, which influences the final prediction, attends almost equally to the Aerial (0.41) and SAR (0.41) inputs. The "Modality Importance" chart reflects this, showing a near 50/50 split. This suggests that the model is effectively fusing the two visual sources to confirm the vessel's presence. The failure to select the correct anomaly type highlights a limitation of the model when its most critical discriminative modality, the AIS, is absent for this specific task.

- **Scenario 3:** Reasoning with Conflicting Multimodal Data as shown in Fig. 12.

This scenario provides a compelling example of the model's behavior when presented with conflicting information across all four modalities. The Aerial and SAR images illustrate a vessel, and the acoustic data aligns with a "Passengership" signature. However, the model generates a highly confident prediction of "Spoofing-Passengership" (99.70%), which by definition implies that no vessel should be present.

- Prediction and Confidence: The model reports extreme confidence in its "Spoofing" classification, even though this contradicts the visual evidence.
- Analysis: This instance serves as an example of learned, dynamic modality weighting. The model appears to have learned from the training data that a specific combination of AIS signals and acoustic signatures

for a "Passengership" serves as a strong indicator of a "Spoofing" event. This learned correlation is sufficiently powerful to override the conflicting visual data in the logic of the model.

- Interpretability: The attention heatmap indicates the primary drivers for this decision. The [CLS] token assigns significant attention to the AIS (0.47) and Acoustic (0.43) tokens while allocating negligible weight to the Aerial (0.063) and SAR (0.062) tokens. The "Modality Importance" chart aligns with this finding, suggesting that over 90% of the decision weight was placed on AIS and Acoustic data. The model appears to have learned to de-prioritize the visual sensors in this specific context. This is a notable finding because it suggests that the model is not merely averaging inputs but is making context-dependent judgments about sensor reliability. For an operator, this dashboard provides a possible explanation for the reasoning of the AI, indicating that the model identifies this as a spoofing event because the AIS and acoustic data are highly indicative of one, despite the conflicting visual evidence.

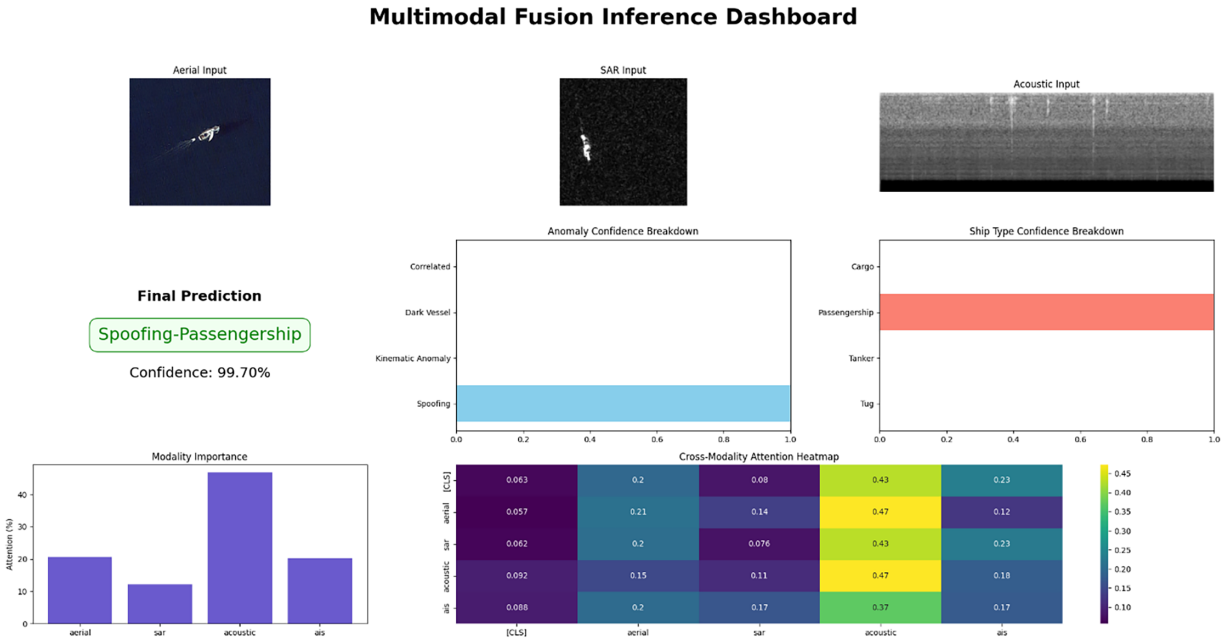

Figure 12: Inference dashboard (Conflicting).

5.5 Discussion and Implications

The experimental results robustly support the hypothesis that a Transformer-based architecture is highly effective for late-fusion of heterogeneous maritime data. The quantitative performance of MAFT is excellent, and its ability to maintain high accuracy even with incomplete sensor data, a direct result of our modality dropout technique, addresses the practical challenge of sensor noise and outages common in operational environments [15,30].

The most significant implication of this work lies in its interpretability. The ability to visualize the attention mechanism demystifies the fusion process, providing actionable intelligence and building trust in the system. This directly addresses the need for Explainable AI (XAI) in high-stakes security domains—a critical challenge and future research direction identified by [1]. This moves beyond simple "black box" classifiers and towards a human-AI collaborative framework for maritime security. The primary limitation of this study is its reliance on synthetic data. While principled, this data may not capture the full complexity of real-world sensor noise and environmental variability.

6 Conclusion and Future Work

In this work, we introduced the MAFT, a novel deep learning framework designed to address the complex challenges of modern MDA. By successfully integrating four heterogeneous data modalities—Aerial imagery, SAR, passive acoustics, and AIS kinematics—our model demonstrates a significant advancement over traditional, unimodal surveillance systems.

Our primary contributions are fourfold. First, we developed a principled synthetic data generation pipeline that effectively overcomes the scarcity of labeled data for maritime anomalies, providing a robust foundation for supervised model training. Second, our proposed MAFT architecture, centered on a Transformer Encoder, proves to be a powerful and flexible late-fusion mechanism. It adeptly learns the complex, non-linear relationships between diverse sensor inputs, achieving a high classification accuracy of 98.0% on a hold-out test set and demonstrating a clear performance advantage over a standard intermediate-fusion MLP baseline. Third, the introduction of modality dropout as a regularization technique was shown to be highly effective, yielding a model that is resilient to incomplete sensor data—a critical requirement for real-world operational reliability. Finally, and perhaps most importantly, the inherent interpretability of the self-attention mechanism provides a transparent view into the model's decision-making process. By analyzing attention weights, we can quantitatively assess which modalities are driving a given prediction, moving beyond opaque "black box" systems towards a more trustworthy and collaborative human-AI framework for maritime security.

Limitations and Future Work

While MAFT improves realism via wave textures and Gamma speckle, a domain gap remains regarding complex multi-path radar interference found in real-world SAR. Future research will focus on bridging this gap using transfer learning on live Sentinel-1 and AIS data streams. In addition to addressing these limitations, we have identified three key directions for expanding this study:

1. Enhancement of the Synthetic Data Environment: The current data synthesis pipeline can be expanded to incorporate a wider range of environmental variables, such as varying sea states, weather conditions including fog or rain, and various times of day. Furthermore, a more extensive taxonomy of anomalous behaviors, including complex coordinated maneuvers and rendezvous events, could be modeled to further improve the robustness and scope of the model.
2. Incorporation of Additional Modalities: The flexible, token-based design of the MAFT architecture readily allows for the integration of additional sensor data. Future iterations could incorporate modalities such as Radio Frequency (RF) signal analysis to detect and classify vessel radar emissions, or natural language processing (NLP) of vessel communication logs like VHF radio to add a layer of semantic context to the operational picture.
3. Real-World Deployment and Validation: The ultimate validation of the MAFT model will be its performance on live, real-world data. The next logical step is to deploy the framework in an operational testbed, interfacing with live streams from aerial drones, satellite SAR providers, hydrophone arrays, and AIS aggregators. This will allow for a comprehensive assessment of its effectiveness in a dynamic environment and provide invaluable data for further refinement and adaptation.

Acknowledgement: During the preparation of this work, the authors used ChatGPT (24 May 2023 version, OpenAI, San Francisco, CA, USA) to assist in language polishing, thereby improving the clarity and readability of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Funding Statement: This project has received funding from the European Union's Horizon Europe research and innovation programme under the Grant Agreement 101168489.

Author Contributions: Raza Hasan conceived of the study, developed the MAFT architecture and drafted the manuscript. Shakeel Ahmad designed the synthetic data generation pipeline and carried out the data modality simulation and formal analysis. Ismet Gocer configured the experimental setup and developed the multimodal fusion inference dashboards. Zakirul Bhuiyan participated in the design of the study, provided technical supervision and helped to draft the manuscript. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets generated and analyzed during the current study were derived from publicly available sources aggregated from the Kaggle platform. The Aerial Imagery and Annotations were sourced from the Ship Detection dataset (https://www.kaggle.com/datasets/andrewmvd/ship-detection). The AIS kinematic data was accessed via the AIS Data for Ships dataset (https://www.kaggle.com/datasets/marsalanakhtar/ais-data-for-ships). The acoustic signatures were obtained from the DeepShip-main repository (https://www.kaggle.com/datasets/vasundharauppuluri/deepship-main). The SAR imagery used for inference visualization was sourced from the SARscope: Unveiling the Maritime Landscape dataset (https://www.kaggle.com/datasets/kailaspsudheer/sarscope-unveiling-the-maritime-landscape). The complete source code for the MAFT pipeline, including the physics-based synthetic data generation scripts, model configurations, and baseline comparison modules, is publicly available at https://github.com/razahasan2000/MAFT.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sahay S, Okur E, Kumar SH, Nachman L. Low rank fusion based transformers for multimodal sequences. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; 2020 Jul 5–10; Online.
2. Kim K, Kim J, Kim J. Robust data association for multi-object detection in maritime environments using camera and radar measurements. IEEE Robot Autom Lett. 2021;6(3):5865–72. doi:10.1109/LRA.2021.3084891.
3. Soldi G, Gaglione D, Forti N, Millefiori LM, Braca P, Carniel S, et al. Space-based global maritime surveillance. Part II: artificial intelligence and data fusion techniques. IEEE Aerosp Electron Syst Mag. 2021;36(9):30–42. doi:10.1109/maes.2021.3070884.
4. Helgesen ØK, Vasstein K, Brekke EF, Stahl A. Heterogeneous multi-sensor tracking for an autonomous surface vehicle in a littoral environment. Ocean Eng. 2022;252(5):111168. doi:10.1016/j.oceaneng.2022.111168.
5. Campos DF, Matos A, Pinto AM. Modular multi-domain aware autonomous surface vehicle for inspection. IEEE Access. 2022;10:113355–75. doi:10.1109/ACCESS.2022.3217504.
6. Hadzagic M, Isabelle M, Kashyap N. Hard and soft data fusion for maritime traffic monitoring using the integrated Ornstein-Uhlenbeck process. In: Proceedings of the 2020 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA); 2020 Aug 24–29; Victoria, BC, Canada. p. 98–105.
7. Melillos G, Themistocleous K, Danezis C, Michaelides S, Hadjimitsis DG, Jacobsen S, et al. Detecting migrant vessels in the Cyprus region using Sentinel-1 SAR data, Online. In: Proceedings of the Counterterrorism, Crime Fighting, Forensics, and Surveillance Technologies IV; 2020 Sep 21–25; Online.
8. Gabeur V, Sun C, Alahari K, Schmid C. Multi-modal transformer for video retrieval. In: Proceedings of the 16th European Conference on Computer Vision—ECCV 2020; 2020 Aug 23–28; Glasgow, UK. p. 214–29.
9. Akbari H, Yuan L, Qian R, Chuang WH, Chang SF, Cui Y, et al. VATT: transformers for multimodal self-supervised learning from raw video, audio and text. Adv Neural Inf Process Syst. 2021;34:24206–21.
10. Bauk S, Kapidani N, Luksic Z, Rodrigues F, Sousa L. Review of unmanned aerial systems for the use as maritime surveillance assets. In: Proceedings of the 2020 24th International Conference on Information Technology (IT); 2020 Feb 18–22; Zabljak, Montenegro. p. 1–5.
11. Soldi G, Gaglione D, Forti N, Di Simone A, Daffina FC, Bottini G, et al. Space-based global maritime surveillance. Part I: satellite technologies. IEEE Aerosp Electron Syst Mag. 2021;36(9):8–28. doi:10.1109/maes.2021.3070862.
12. Han X, Wang YT, Feng JL, Deng C, Chen ZH, Huang YA, et al. A survey of transformer-based multimodal pre-trained modals. Neurocomputing. 2023;515(2):89–106. doi:10.1016/j.neucom.2022.09.136.

13. Sithiravel R, Balaji B, Nelson B, McDonald MK, Tharmarasa R, Kirubarajan T. Airborne maritime surveillance using magnetic anomaly detection signature. IEEE Trans Aerosp Electron Syst. 2020;56(5):3476–90. doi:10.1109/TAES.2020.2973866.
14. Yoon S, Jalal A, Cho J. MODAN: multifocal object detection associative network for maritime horizon surveillance. J Mar Sci Eng. 2023;11(10):1890. doi:10.3390/jmse11101890.
15. Talpur K, Hasan R, Gocer I, Ahmad S, Bhuiyan Z. AI in maritime security: applications, challenges, future directions, and key data sources. Information. 2025;16(8):658. doi:10.3390/info16080658.
16. Karki S, Kulkarni S. Ship detection and segmentation using Unet. In: Proceedings of the 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT); 2021 Feb 19–20; Bhilai, India. p. 1–7.
17. Maganaris C, Protopapadakis E, Doulamis N. Outlier detection in maritime environments using AIS data and deep recurrent architectures. In: Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments; 2024 Jun 26–28; Crete, Greece. p. 420–7.
18. Potamos G, Stavrou E, Stavrou S. Enhancing maritime cybersecurity through operational technology sensor data fusion: a comprehensive survey and analysis. Sensors. 2024;24(11):3458. doi:10.3390/s24113458.
19. Salloum H, Sutin A, Sedunov N, Sedunov A, Kadyrov D. Low-cost multimodal integrated marine surveillance system. In: Proceedings of the 2022 IEEE International Symposium on Technologies for Homeland Security (HST); 2022 Nov 14–15; Boston, MA, USA. p. 1–6.
20. Roheda S, Krim H, Riggan BS. Robust multi-modal sensor fusion: an adversarial approach. IEEE Sens J. 2021;21(2):1885–96. doi:10.1109/JSEN.2020.3018698.
21. Teixeira E, Araujo B, Costa V, Mafra S, Figueiredo F. Literature review on ship localization, classification, and detection methods based on optical sensors and neural networks. Sensors. 2022;22(18):6879. doi:10.3390/s22186879.
22. Alon AS, Macalisang J, Reyes RC, Sevilla RV, Belga GD. Watercraft-net: a deep inference vision approach of watercraft detection for maritime surveillance system using optical aerial images. In: Proceedings of the 2020 IEEE 7th International Conference on Engineering Technologies and Applied Sciences (ICETAS); 2020 Dec 18–20; Kuala Lumpur, Malaysia. p. 1–5.
23. Liu W, Liu Y, Gunawan BA, Bucknall R. Practical moving target detection in maritime environments using fuzzy multi-sensor data fusion. Int J Fuzzy Syst. 2021;23(6):1860–78. doi:10.1007/s40815-020-00963-1.
24. Gaglione D, Soldi G, Meyer F, Hlawatsch F, Braca P, Farina A, et al. Bayesian information fusion and multitarget tracking for maritime situational awareness. IET Radar Sonar Navig. 2020;14(12):1845–57. doi:10.1049/iet-rsn.2019.0508.
25. Xu P, Zhu X, Clifton DA. Multimodal learning with transformers: a survey. IEEE Trans Pattern Anal Mach Intell. 2023;45(10):12113–32. doi:10.1109/tpami.2023.3275156.
26. Thombre S, Zhao Z, Ramm-Schmidt H, Vallet Garcia JM, Malkamaki T, Nikolskiy S, et al. Sensors and AI techniques for situational awareness in autonomous ships: a review. IEEE Trans Intell Transport Syst. 2022;23(1):64–83. doi:10.1109/tits.2020.3023957.
27. Farahnakian F, Heikkonen J. Deep learning based multi-modal fusion architectures for maritime vessel detection. Remote Sens. 2020;12(16):2509. doi:10.3390/rs12162509.
28. Wang Y, Rajesh G, Raajini XM, Kritika N, Kavinkumar A, Shah SBH. Machine learning-based ship detection and tracking using satellite images for maritime surveillance. J Ambient Intell Smart Environ. 2021;13(5):361–71. doi:10.3233/ais-210610.
29. Bakirci M. Advanced ship detection and ocean monitoring with satellite imagery and deep learning for marine science applications. Reg Stud Mar Sci. 2025;81(2)):103975. doi:10.1016/j.rsma.2024.103975.
30. Al Mansoori A, Swamidoss I, Almarzooqi A, Sayadi S. An investigation of various dehazing algorithms used on thermal infrared imagery for maritime surveillance systems. In: Proceedings of SPIE Conference on Target and Background Signatures VII; 2021 Sep 13–18; Online.