A Robust Defense Mechanism Against Adversarial Attacks in Maritime Autonomous Ship Using GMVAE+RL

Ganesh Ingle, Kailas Patil, Sanjesh Pawale
Department of Computer Engineering, Vishwakarma University, Pune, India

***Abstract*—In this paper, we propose a robust defense framework combining Gaussian Mixture Variational Autoencoders (GMVAE) with Reinforcement Learning (RL) to counter adversarial attacks in Maritime Autonomous Systems, specifically targeting the Singapore Maritime Database. By modeling complex maritime data distributions through GMVAE and dynamically adapting decision boundaries via RL, our approach establishes a resilient latent representation space that effectively identifies and mitigates adversarial perturbations. Experimental evaluations using adversarial methods such as FGSM, IFGSM, DeepFool, and Carlini-Wagner attacks demonstrate that the proposed GMVAE+RL model outperforms traditional defenses in both accuracy and robustness. Specifically, it achieves a peak accuracy of 87% and robustness of 20.5%, compared to 85.8% and 19.2% for FGSM and significantly lower values for other methods. These results underscore the superiority of our method in ensuring data integrity and operational reliability within complex maritime environments facing evolving cyber threats.**

Keywords*—*Maritime autonomous systems; reinforcement learning; defense mechanisms; Gaussian Mixture Variational Auto encoder; Singapore maritime database

I. INTRODUCTION

The maritime industry is experiencing a paradigm shift with the advent of Artificial Intelligence (AI), which is set to revolutionize various operational facets through heightened automation, efficiency enhancement, and cost mitigation.

A. Maritime Autonomous Systems (MAS)

AI's instrumental role is exemplified in the development of Maritime Autonomous Systems (MAS), which necessitate minimal human governance, employing AI for executive decisions and navigational control [1],[2],[5].

B. Advantages of MAS

MAS herald a new era in maritime operations, characterized by:

- Diminished manpower requisites, leading to significant labor cost reductions.
- AI-facilitated automation and refinement of complex maritime tasks.
- Augmented crew safety, especially under perilous operational conditions.
- Operational cost economization through heightened MAS efficiency.
- Environmental impact mitigation via the integration of renewable energy sources.

C. MAS Development Progress

Rapid advancements in MAS development are being witnessed, with a multitude of applications ranging from cargo transit to environmental oversight.

D. Security Implications of AI in MAS

However, the ascent of AI within MAS introduces new security paradigms, predominantly concerning Adversarial Artificial Intelligence (AAI).

E. Adversarial Artificial Intelligence (AAI)

AAI encapsulates the intentional manipulation of AI frameworks, aiming to pinpoint and capitalize on systemic vulnerabilities, thus posing a significant threat to MAS security and operational integrity.

F. AAI Vulnerabilities in MAS

MAS AI systems are susceptible to various AAI attacks due to their reliance on complex algorithms and data-driven decision-making processes. Manipulating training data to introduce biases or errors in the AI model, leading to incorrect decisions or system malfunctions. Model inversion: Inferring sensitive information from the AI model's parameters, such as training data or model architecture.Crafting inputs that the AI model misclassifies or misinterprets, potentially enabling attackers to evade detection or manipulate system behavior [35],[36]. Inference attacks: Exploiting the AI model's decision-making process to influence its outputs, such as steering a vessel towards a hazardous area or triggering false alarms [3], [4], [6]. Impact of AAI on MAS Security Collisions: Attackers could manipulate the AI navigation system to cause collisions with other vessels or obstacles, leading to loss of life and environmental damage. Cargo theft: Attackers could intercept or reroute cargo shipments, causing financial losses and disrupting supply chains. Attackers could exploit vulnerabilities in the AI system to gain unauthorized access to sensitive data or disrupt critical operations. The rapid advancements in artificial intelligence (AI) have opened up a plethora of opportunities for enhancing maritime operations through autonomous systems. However, the integration of AI into maritime autonomous systems (MAS) also introduces new security challenges, particularly from adversarial AI (AAI).

AAI refers to the malicious use of AI to exploit vulnerabilities and compromise the integrity of AI-powered systems. The infusion of AI into maritime operations has catalyzed a transformative phase in the maritime sector, yet it simultaneously ushers in new security vulnerabilities, especially from AAI.

G. AAI Threats in the Maritime Domain

The maritime sector's intrinsic dynamic and unpredictable nature exacerbates the vulnerability of AI systems to AAI threats. These threats encompass:

H. Data Poisoning

Adversarial entities may corrupt training datasets, inducing biases or errors that could precipitate erroneous decision-making or functional disruptions within MAS.

I. Model Inversion

Attackers might extract sensitive data or discern the model's structure from its parameters, thus acquiring tactical knowledge about the system's operations.

J. Evasion Attacks

Specially crafted inputs may lead AI models to misclassify or misconstrue data, permitting adversaries to skirt detection or alter system actions.

K. Inference Attacks

Exploitation of the decision-making process within AI models can be manipulated to influence outcomes, potentially resulting in navigational errors or security breaches.

The maritime industry is undergoing a transformative evolution with the integration of Artificial Intelligence (AI) into Maritime Autonomous Systems (MAS), promising enhanced operational efficiency, reduced human intervention, and improved safety. MAS rely heavily on AI-driven decision-making for navigation, cargo management, and environmental monitoring. However, as the reliance on AI systems deepens, so does the surface for security vulnerabilities—particularly from Adversarial Artificial Intelligence (AAI), which involves deliberate perturbations in input data that can mislead AI models into making erroneous or even dangerous decisions [6-14].

Previous studies have investigated adversarial attacks and their countermeasures, primarily in controlled or theoretical environments using static defense mechanisms such as adversarial training, input transformations, or model distillation. While these approaches show promise in generic settings, they often fall short in real-world maritime environments characterized by high data variability, dynamic vessel behaviors, and critical security requirements. Specifically, existing defense strategies lack adaptability and robustness when confronted with iterative, optimization-based attacks like Carlini-Wagner or DeepFool, which can subtly and effectively compromise AI models without detection.

This presents a significant research gap: there is a pressing need for defense mechanisms that can not only model complex, multimodal maritime data distributions but also dynamically adapt to evolving attack strategies in real time. Addressing this, we propose a hybrid defense architecture that combines Gaussian Mixture Variational Autoencoders (GMVAE) for resilient data representation with Reinforcement Learning (RL) for adaptive policy optimization. The GMVAE component ensures a structured latent space capable of identifying subtle anomalies, while RL empowers the model to learn countermeasures through continuous feedback, improving robustness over time.

By focusing on the underexplored intersection of generative modeling and adaptive learning in adversarial defense, this research provides a practical and scalable solution tailored to the maritime domain. The approach is validated on the Singapore Maritime Dataset, demonstrating superior performance over existing methods in terms of both accuracy and adversarial robustness. This work not only fills a critical gap in maritime cybersecurity literature but also sets a foundation for future research in real-time, adaptive AI defense systems.

Compared to traditional defense strategies such as adversarial training, input transformation, and static regularization techniques, the proposed GMVAE+RL framework offers multiple significant advantages. Firstly, the GMVAE component excels at capturing multi-modal and complex maritime data distributions, enabling it to identify subtle perturbations that static defenses often miss. Secondly, the integration of Reinforcement Learning provides an adaptive mechanism that dynamically adjusts the model's behavior in response to evolving attack strategies—something existing models lack. Thirdly, the hybrid approach enhances both generalization and interpretability by learning structured latent representations and optimizing decision policies simultaneously. Experimental comparisons against established methods like FGSM, IFGSM, DeepFool, and Carlini-Wagner reveal that our method maintains higher accuracy and robustness, with a notable 87% accuracy and 20.5% robustness even under strong adversarial conditions. These outcomes underscore the model's superior resilience and adaptability, making it highly suitable for real-world applications in autonomous maritime systems where data integrity and security are mission-critical.

The remainder of this paper is organized as follows: Section II provides the background and motivation for adversarial resilience in Maritime Autonomous Systems, followed by a review of related work in Section III. Section IV details the proposed methodology combining GMVAE and Reinforcement Learning, while Section V outlines the experimental setup used for evaluation. Section VI presents a comprehensive analysis of results, including performance metrics under various adversarial scenarios. Finally, Section VII concludes the paper with key findings and directions for future research.

II. Background

Global trade heavily relies on maritime vessels, with a significant portion of international movement facilitated by shipping [5]. This paper explores the integration of advanced sensors in fully autonomous vessels (Level 4 as defined by the International Maritime Organization), which operate independently without any human crew.

MAS utilize a variety of sensors and instruments for environmental perception and decision-making, including:

- RADAR: For detecting large objects using radio waves.
- LiDAR: Employed for accurate detection of smaller objects.
- Echo Sounders: Utilized for underwater object detection.
- CCTV/IR/multispectral Cameras: For close-range object detection.
- Microphone Arrays: Capture audio cues for situational awareness.
- AIS and GNSS: Provide location and data transmission capabilities.
- ECDIS, Weather Sensors, and Communication Systems: Crucial for navigation and environmental monitoring.
- Specialized sensors and Drones: Extend the range and capabilities of standard sensor systems.

The integration of multiple sensors provides increased accuracy, improved redundancy, and enhanced situational awareness. Sensors in MAS face unique challenges such as water-induced distortions, harsh environmental conditions, and detection complexities. The effective deployment of a diverse sensor array is paramount in MAS, requiring a deep understanding of their individual and collective capabilities and limitations in the maritime context. In fully autonomous maritime systems, AI plays a crucial role in automating vessel operation. It receives sensor data as input, analyzes the information, and makes decisions to control the vessel's actions, replacing or supplementing crew functions. The specific AI technologies required depend on the range of tasks and functionalities of the MAS. Based on the categorization, several key AI technologies are employed in MAS, connected to a Dynamic Positioning (DP) system that controls the vessel's movements:

Determines the vessel's real-time location and environment, including object detection and range. Convolutional neural networks (CNNs), region proposal networks(RPNs), and natural language processing (NLP) for interpreting communication.Sensor data, including camera images, radar signals, and LiDAR data.Real-time information about the vessel's surroundings and potential hazards. Prevents collisions with other vessels or objects [14-23]. CNNs for object recognition and support vector machines (SVMs) for trajectory planning.SA information, including object detection data.New trajectory to avoid collisions.Determines the optimal route for the vessel, considering factors like fuel efficiency, speed, and safety.Evolutionary algorithms (EAs), particle swarm optimization (PSO), and ant colony optimization (ACO).Global map data, weather information, and vessel parameters.Optimal route for the vessel to follow.

- Convolutional Neural Networks (CNNs): Efficiently learn spatial features from images, making them ideal for object detection and recognition in SA and collision avoidance modules.
- Region Proposal Networks (RPNs): Generate candidate object bounding boxes within images, improving the efficiency of object detection for SA.
- Natural Language Processing (NLP): Enables interpretation of communication signals like radio messages, enhancing situational awareness.
- Support Vector Machines (SVMs): Effective for classification tasks, such as determining the type of object detected and generating new collision-avoidance trajectories.
- Evolutionary Algorithms (EAs): Powerful optimization techniques that can handle complex multi-objective problems, like finding the optimal global route for the vessel.
- Particle Swarm Optimization (PSO): Mimics the behavior of a swarm of birds to find optimal solutions, applicable to path planning and route optimization.
- Ant Colony Optimization (ACO): Inspired by the foraging behavior of ants, ACO can identify efficient routes by simulating pheromone communication. AI plays a pivotal role in automating various aspects of MAS operation. Different AI technologies are employed for specific tasks, from situational awareness and collision avoidance to global path planning and vessel maintenance. Understanding the capabilities and limitations of these AI technologies is crucial for designing and developing safe and reliable MAS.

Most evaluations of adversarial attacks on machine learning (ML) systems have been limited to controlled laboratory environments. This study extends the analysis to real-world MAS environments, where the implications of such attacks are less understood but potentially more impactful.

While focusing on adversarial attacks, this work also acknowledges the significance of conventional cybersecurity attacks and the potential for combined adversarial AI and conventional cybersecurity tactics. The influence of conventional security vulnerabilities on both AI-based and traditional security is also recognized.

A. Class 1: Model Inversion

- Description: An attacker queries the ML model to deduce its prerequisite features, potentially aiding in reconnaissance for future attacks.
- Impact: This represents an abuse of the system's confidentiality, although it does not directly impair the model's functionality.

This comprehensive evaluation of adversarial attacks in MAS provides critical insights into their real-world implications, emphasizing the need for robust defense mechanisms in maritime autonomous systems.

III. Literature Survey

Huang et al. represents a critical juncture in the field of artificial intelligence, particularly in understanding the vulnerabilities of reinforcement learning (RL) systems to adversarial attacks. Reinforcement learning, which functions on a framework of rewards and penalties, had been increasingly applied in varied domains such as gaming, autonomous navigation, and decision-making algorithms. However, the robustness of these systems against subtle, malicious alterations had not been thoroughly examined until this study.This research focused on the concept of adversarial attacks, previously acknowledged in other neural network contexts, where slight, calculated changes to input data could drastically mislead the network's output. They applied this concept to RL, investigating whether minor perturbations in the input data of an RL agent could derail its performance. Their experiments cut across different RL environments to ensure a comprehensive assessment.The findings were revelatory, demonstrating that even negligible modifications to input data could significantly impair the RL models' performance. This vulnerability was not confined to specific RL algorithms or tasks but was a more generalized issue, indicating a fundamental security risk in RL applications. Crucially, the study's implications extended beyond the immediate realm of RL, casting a spotlight on the need for adversarial robustness in AI systems, particularly in safety-critical applications like autonomous vehicles[23-29].Presented research precipitated a heightened awareness and subsequent research efforts aimed at developing more robust RL systems capable of resisting such adversarial attacks. The study not only emphasized the importance of considering security threats in AI system design but also spurred advancements in defensive techniques, marking a significant leap in the development of secure and reliable AI solutions.

Chen et al. provided a crucial insight into the cybersecurity vulnerabilities of Connected Vehicle (CV) based transportation systems, particularly focusing on the risks associated with data spoofing attacks. In the era of advanced transportation technology, CV systems have emerged as a key innovation, enhancing vehicular communication and operational efficiency through vehicle-to-vehicle and vehicle-to-infrastructure interactions. However, the integration of such complex communication systems also opens up new avenues for cyber threats.Proposed study embarked on a comprehensive analysis of the CV systems' architecture and operational mechanisms. Their primary objective was to identify and assess potential cybersecurity threats, with a special emphasis on data spoofing – a technique where false information is injected into a system, leading to misguided actions or responses. Through detailed simulations and hypothetical attack scenarios, the study highlighted how these systems are particularly prone to data spoofing, which could lead to severe consequences like traffic disruptions or even collisions [29],[30].

One of the key revelations of this study was the identification of inherent design flaws within CV systems that made them susceptible to such cyberattacks. These vulnerabilities could potentially be exploited to manipulate critical aspects of traffic control or to feed misleading information to vehicles, thus compromising road safety. The findings played a pivotal role in emphasizing the need for robust, multi-layered cybersecurity measures within CV systems. This study not only underscored the importance of incorporating stringent security protocols in the design and implementation of CV technology but also acted as a catalyst for further research and development in enhancing the resilience of connected vehicles against cyber threats. The work of Chen and colleagues thus marked a significant step in ensuring that the advancements in vehicle connectivity and automation do not compromise safety and security [31[32]. Lin et al. introduced a groundbreaking approach to adversarial attacks within the realm of Atari games, marking a significant advancement in the understanding of vulnerabilities in reinforcement learning systems. Their innovative concept, termed "strategically-timed attacks," involved the creation of adversarial examples that were calculated independently at each timestep of the game. This method diverged from traditional continuous attack models, offering a more nuanced and potentially more disruptive technique. By strategically timing these attacks, Lin and colleagues demonstrated that it was possible to significantly impair the performance of reinforcement learning agents in game scenarios. These attacks were designed to be subtle enough to avoid immediate detection, yet sufficiently impactful to mislead the agents, leading to incorrect decisions or actions within the game. This research not only highlighted a specific vulnerability in reinforcement learning applications but also set a new precedent in the methodology of adversarial attack strategies. It underscored the need for more robust defense mechanisms in AI systems, particularly in environments where decision-making is based on real-time data inputs, such as in gaming or autonomous navigation scenarios. The work of Lin et al. thus stands as a pivotal contribution to the field of AI security, illustrating the evolving nature of cyber threats and the ongoing challenge of securing AI against sophisticated adversarial techniques.

Xiang et al. conducted a noteworthy study focusing on the domain of Q-learning, specifically within the context of automatic path planning. Their research made a significant contribution to the field by proposing a probabilistic output model designed to predict adversarial examples in such structured environments. The essence of their work revolved around the exploration of adversarial attacks in scenarios that are inherently more systematic and organized, compared to the often chaotic nature of other environments like gaming.The innovative aspect of research lay in the application of their model to Q-learning, a fundamental reinforcement learning technique widely used for making sequence-based decisions. By integrating a probabilistic approach, they were able to forecast the likelihood of adversarial instances occurring in an automatic path planning context. This model was not only pivotal in identifying potential vulnerabilities within the path planning algorithms but also in suggesting the probability of certain attacks succeeding.Their work shed new light on the dynamics of adversarial attacks in environments characterized by a high degree of order and predictability, such as route planning and navigation. By doing so,authors expanded the understanding of how adversarial attacks could be tailored and predicted in such settings, contrasting with the more generalized approach typically seen in other AI applications. The implications of this study are far-reaching, especially considering the growing reliance on autonomous systems in various sectors, including transportation and logistics. It underscores the need for advanced security measures that can anticipate and mitigate such sophisticated cyber threats in automated and algorithm-driven environments.

Huang et al. not only exposed the vulnerabilities of reinforcement learning systems to adversarial attacks but also introduced a significant defensive mechanism known as the Fast Gradient Sign Method (FGSM). This method was designed specifically to counteract the negative impacts of adversarial inputs, especially in the context of deep reinforcement learning agents.FGSM operates by utilizing the gradients of the neural network to create perturbations that ‘push’ the input data towards the direction of increasing the loss. This method is particularly notable for its simplicity and efficiency. Instead of requiring complex or time-consuming computations, FGSM generates adversarial examples by applying a straightforward adjustment in the direction of the gradient. The ‘sign’ component of the method refers to taking the sign of the gradient, ensuring that the perturbations are small yet effective enough to mislead the learning model.In the context of deep reinforcement learning, where agents are often trained on high-dimensional input data such as images or sensor readings, FGSM provides a valuable tool for enhancing the robustness of these systems. By applying adversarial examples generated via FGSM during the training process, the learning models can be ‘inoculated’ against potential attacks, learning to recognize and resist manipulative inputs. This approach essentially strengthens the model’s ability to maintain performance even when faced with subtly altered input data, a critical requirement in applications where reliability and accuracy are paramount.Authors highlights the FGSM’s role in defending against adversarial attacks has been pivotal in the field of AI security. It marked a step forward in developing more secure AI systems capable of operating reliably in adversarial environments, a vital consideration as AI technologies continue to be integrated into increasingly critical and sensitive applications.

Silver et al. applied RL to the game of Go. They introduced a novel approach that combined deep neural networks with tree search, leading to unprecedented performance in this complex board game. This demonstrated RL’s potential in mastering highly intricate and strategic tasks. Mnih et al. were pioneers in applying deep learning to RL, particularly in the context of playing Atari games. Their model was the first to successfully learn control policies directly from high-dimensional sensory inputs, marking a significant advancement in the field of game-playing AI.

The increasing prevalence of deep learning applications has brought to light the vulnerability of these models to adversarial attacks. These attacks involve crafting subtle modifications to input data that can cause deep learning models to make erroneous predictions. Even minor perturbations can have a significant impact on model performance. This poses a serious threat to the reliability of deep learning systems, especially in critical applications such as autonomous vehicles and medical diagnosis.

The Fast Gradient Sign Method (FGSM) [1] is a fundamental adversarial attack technique that involves calculating the gradients of the model’s loss relative to the input data and modifying the data based on these gradients. This method was expanded into IFGSM [2], which applies perturbations iteratively to increase the strength of the adversarial effect. More complex methods like the DeepFool attack [3] and the Carlini-Wagner attack [4] employ advanced strategies. DeepFool iteratively identifies the minimal perturbation needed to misclassify an input by approximating the decision boundary, whereas the Carlini-Wagner attack uses optimization techniques to create adversarial examples with minimal changes but targeted misclassification goals.

Various defense strategies have been proposed to mitigate adversarial threats. Adversarial training [5] involves training models using adversarial examples to enhance their robustness. Other techniques, like feature squeezing and input transformations (including JPEG compression) [6], aim to eliminate adversarial perturbations. However, these methods often struggle to generalize across different types of attacks, highlighting the need for more innovative solutions.

Variational Autoencoders (VAEs) [7] provide a structured framework for learning generative models. They consist of an encoder, which maps input data into a latent space, and a decoder, which reconstructs data from these latent representations. The learning objective is to minimize reconstruction loss while ensuring the latent space adheres to a structured distribution, typically using the Kullback-Leibler (KL) divergence. This latent space captures essential features for controlled data generation. Extending VAEs, Gaussian Mixture Variational Autoencoders (GMVAE) [8] incorporate Gaussian Mixture Models (GMMs) into the latent space, enabling the representation of complex, multi-modal data distributions, thereby overcoming some limitations of standard VAEs.

Reinforcement learning has been identified as a promising method for improving model resilience against adversarial attacks [9]. This approach applies principles from control theory, using a policy network that learns actions to maximize cumulative rewards. In defense contexts, these actions involve decisions that enhance model robustness. This method involves training the policy network to make decisions that lead to accurate predictions on adversarial examples, counteracting adversarial perturbations. The reinforcement learning framework offers the benefit of continual adaptation, allowing models to enhance their robustness over time.

The MNIST dataset [10] has been a benchmark in machine learning for testing various defense techniques, contributing significantly to our understanding of adversarial challenges and defense strategies. This research includes both traditional and advanced deep learning-based solutions aimed at protecting model performance under adversarial conditions.

Adversarial training and defensive distillation, two current defenses for CNN-LSTM models, have trouble being effective and generalizing against adversarial attacks in PQD classification. These techniques fall short of maintaining high precision when attacked, indicating a need for more flexible and effective defense tactics. This is addressed by Input Adversarial Training (IAT), which meets a crucial demand for CNN-LSTM model security in power system applications by improving model robustness while maintaining performance [32]. Current adversarial defenses frequently lack an ideal balance between accuracy and robustness. Feature masking’s potential is still not fully realized, particularly when paired with gradient modification. As our study showed, this disparity emphasizes the need for effective measures that improve resilience without lowering performance [33].

Existing GNN defence methods focus on highly linked training processes, overlooking adaptive adversarial attack strategies. This study addresses the gap by introducing GNN Attacker, leveraging Energy Honey Badger Optimization (EHBO) for generating adversarial attacks. The model achieves high visual similarity 90.77%, classification accuracy 94.68%, and attack success rate 96.54%, demonstrating its effectiveness in testing GNN robustness [34].

Existing deep learning models are highly vulnerable to adversarial attacks, which introduce subtle perturbations leading to misclassification. Detecting and mitigating these attacks remains a significant challenge. This review addresses the gap by providing a comprehensive analysis of adversarial attack strategies and defense mechanisms, contributing to the development of more resilient deep learning and machine learning models [35].

Existing research on adversarial robustness has explored various defense mechanisms, including adversarial training, input transformations, and feature denoising. However, optimizing bit plane slicing for resilience remains underexplored. This study leverages genetic algorithms to refine bit-depth configurations, revealing that 5-bit representations enhance robustness against FGSM and DeepFool attacks. Despite performance degradation under adversarial conditions, optimized models demonstrate significant recovery. Prior work lacks dynamic bit plane adaptation, evaluation on diverse attacks, and scalability to large datasets. Addressing these gaps through adaptive slicing, black-box evaluations, and hybrid defenses can further strengthen adversarial resilience [36] [37].

Despite progress in defending against adversarial attacks, a gap remains in developing robust, generalizable solutions. Current defenses often perform well against certain attack types but are less effective in varied adversarial scenarios. This study seeks to address this gap by combining the capabilities of GMVAEs and reinforcement learning. This innovative approach aims to harness the unsupervised feature learning of GMVAEs and the adaptability of reinforcement learning's policy optimization, proposing a new direction for enhancing defense mechanisms against adversarial threats.

Existing defense mechanisms against adversarial attacks in Maritime Autonomous Systems (MAS) largely focus on either static adversarial training or heuristic input transformations, which often lack adaptability and fail to generalize across evolving attack strategies. These approaches struggle particularly in complex, real-world maritime contexts such as the Singapore Maritime Database, where data is highly dynamic and multi-modal. To bridge this gap, we propose a novel hybrid defense framework that integrates Gaussian Mixture Variational Autoencoders (GMVAE) with Reinforcement Learning (RL) to create an adaptive, resilient latent space capable of detecting and mitigating sophisticated adversarial manipulations. The GMVAE component excels in modeling diverse data distributions and isolating irregular patterns, while RL dynamically adjusts model responses based on feedback from adversarial environments. Experimental evaluations using standard adversarial methods—FGSM, IFGSM, DeepFool, and Carlini-Wagner—reveal that our approach significantly outperforms conventional defenses, achieving an accuracy of 87% and robustness of 20.5%, compared to lower benchmarks from existing methods. By explicitly addressing the shortcomings of static defenses and introducing an adaptive learning mechanism, our work advances the state of the art in maritime cybersecurity, ensuring higher integrity and reliability of autonomous ship operations under adversarial conditions.

IV. Methodology

Fig. 1 gives proposed architecture diagram for the GMVAE with reinforcement learning.

In the realm of machine learning, the Gaussian Mixture Variational Autoencoder (GMVAE) stands out for its proficiency in processing complex, multi-modal data distributions. This model excels due to its capacity to discern diverse data representations by employing a combination of Gaussian distributions within its latent space. This capability surpasses that of the traditional Variational Autoencoder (VAE). In the context of adversarial attacks, which are tactics used to subtly alter input data to mislead machine learning models into erroneous predictions or classifications, the stakes are high. These attacks could result in various detrimental outcomes, such as mislabeling of ships, inaccuracies in maritime tracking, or even jeopardizing port security.

In safeguarding the Singapore Maritime Database against such adversarial threats, the GMVAE emerges as a key tool. Its sophisticated approach to data representation makes it highly effective in enhancing the database's defense mechanisms against these types of cyber threats.

1) Latent space modeling: The GMVAE, a sophisticated machine learning model, structures data within its latent space as a blend of Gaussian distributions. Each Gaussian element, characterized by its mean μ_k and standard deviation σ_k, encapsulates a distinct aspect of the data's distribution.

2) Enhanced pattern recognition: GMVAE's advanced data interpretation allows it to discern intricate patterns and irregularities with greater precision than more basic models, making it a valuable asset in identifying subtle discrepancies.

3) Handling data uncertainty: The GMVAE's probabilistic approach is instrumental in gauging uncertainty. This feature is vital for pinpointing and understanding manipulated data points, commonly known as adversarial examples, within the Singapore Maritime Database. This capability is crucial in bolstering the database's defenses against adversarial cyber attacks.

Defending against adversarial attacks on the Singapore Maritime Database, the Gaussian Mixture Variational Autoencoder (GMVAE) plays a crucial role with its encoder network, latent space representation, and decoder network. Each component of the GMVAE contributes to enhancing the robustness of the system against such attacks:

The encoder in a GMVAE takes an input vector x (representing maritime data) and maps it to the parameters of a Gaussian mixture model in the latent space. Mathematically, this can be expressed as a function

$$f : x \rightarrow (\mu_k, \sigma_k^2) \tag{1}$$

where μ_k and σ_k^2 are the mean and variance of the k-th Gaussian component in the latent space. This encoding process translates complex, high-dimensional maritime data into a structured latent space. By doing so, it aids in differentiating standard operational data from potentially manipulated (adversarial) inputs. The encoder's effectiveness in this mapping is crucial for early detection of data inconsistencies or anomalies that could indicate a security breach. In the latent space, data points are represented as a mixture of Gaussian distributions. This can be mathematically formulated as

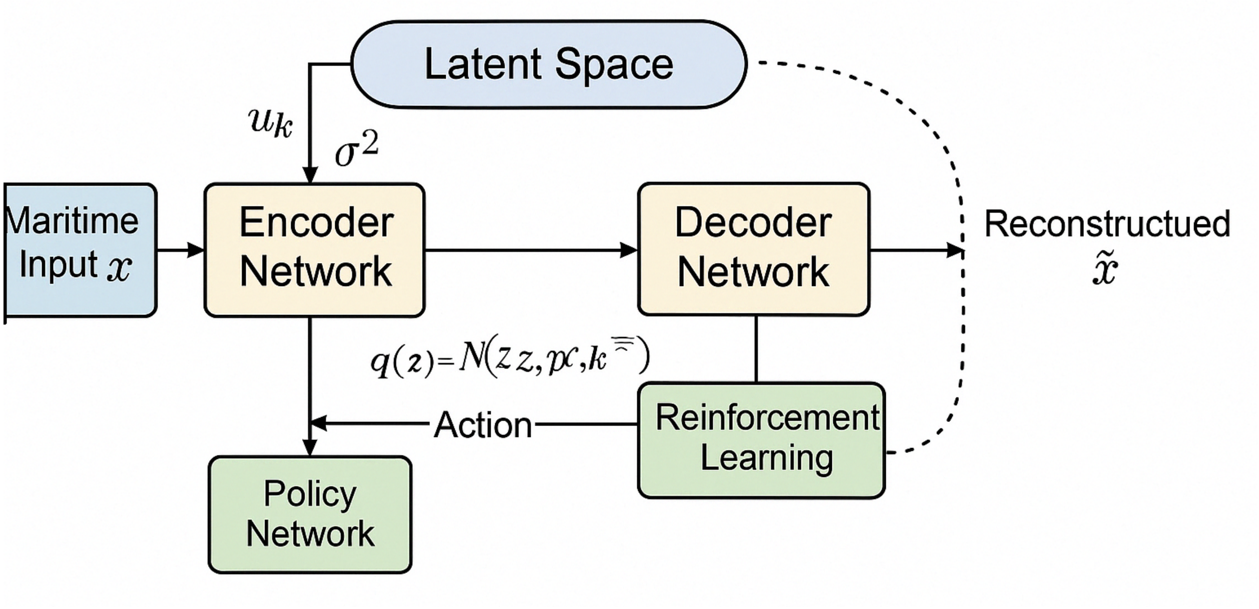

Fig. 1. Architecture diagram for proposed method.

$$p(z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(z; \mu_k, \sigma_k^2) \quad (2)$$

where z is the latent variable, π_k is the mixture coefficient for the k-th component, and $\mathcal{N}$ denotes the Gaussian distribution. The latent space's ability to model complex data distributions enables the identification of subtle deviations from typical data patterns. This is particularly useful in the maritime context for detecting adversarial manipulations like forged vessel locations or tampered cargo records. The probabilistic nature of this space allows for a more nuanced understanding of data uncertainty, which is key in identifying adversarial examples.

The decoder network aims to reconstruct the input data from its latent representation. This can be viewed as a function:

$$g : (\mu_k, \sigma_k^2) \to \hat{x} \quad (3)$$

where $\hat{x}$ is the reconstructed input. The decoder's role in defense is to reconstruct the input data from the latent representation and compare it with the actual input. Significant deviations in this reconstruction process can indicate adversarial manipulations. Mathematically, if the reconstruction loss, typically measured as the difference between x and $\hat{x}$ (e.g., using mean squared error), exceeds a certain threshold, it may signal an anomaly. The GMVAE's encoder network mathematically transforms maritime data into a structured latent space, where data points are probabilistically modeled as a mixture of Gaussians. This transformation is key to detecting abnormalities in the data, which could signify adversarial attacks. The latent space serves as a critical junction for identifying unusual data distributions that diverge from standard patterns. Finally, the decoder's mathematical reconstruction of the input data provides a means to verify the integrity of the data, making it a vital component in the detection and defense against adversarial threats in the Singapore Maritime Database.

- Anomaly Identification: Utilizing its advanced capabilities, the GMVAE can pinpoint irregularities in standard data patterns, which could indicate adversarial interference. This feature is particularly valuable in spotting potential cyber threats within the maritime data.
- Data Integrity Checks: The process of reconstructing input data from its latent representation in GMVAE serves as a critical check. Any significant mismatches between the original and reconstructed data are red flags that may denote a cyber intrusion.
- Dynamic Adaptation: Continuously integrating new data into the GMVAE enables it to stay abreast of changing adversarial techniques, enhancing its ability to safeguard against evolving cyber threats.

Embedding the GMVAE within the cybersecurity framework of Singapore's maritime database equips it with a sophisticated mechanism to detect and counter adversarial attacks. The model's proficiency in managing complex data and its adeptness at modeling uncertainty render it a powerful asset in defending against such sophisticated cyber challenges.

In the application of the Gaussian Mixture Variational Autoencoder (GMVAE) for defending the Singapore Maritime Database against adversarial cyber attacks, the encoder's output representation $q(z|x)$ plays a crucial role, defined mathematically as:

$$q(z \mid x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(z \mid \mu_k(x), \sigma_k^2(x)) \quad (4)$$

This formulation encompasses several key components: K represents the number of Gaussian components in the mixture, critical for modeling complex maritime data patterns; π_k is the mixing coefficient for the k-th component, reflecting its relative significance in the mixture; and $\mathcal{N}(z|\mu_k(x), \sigma_k^2(x))$ denotes the Gaussian distribution for each component, conditioned on the input x. These components collectively enable the GMVAE to perform a detailed and probabilistic mapping of input data to the latent space, which is essential for detecting deviations from normal data behavior that might indicate adversarial activities. Through such sophisticated mathematical modeling, GMVAE significantly enhances the capability to identify and mitigate potential cyber threats in the maritime database.

- Complex Data Modeling with Gaussian Mixtures: GMVAE's ability to represent intricate data distributions as a mixture of Gaussian components is crucial. This enables the detection of nuanced patterns and variations in maritime data, essential for identifying anomalies indicative of adversarial attacks.
- Optimizing the Evidence Lower Bound (ELBO):
  - Reconstruction Term: $\log p_\theta(x|z)$ assesses the decoder's ability to reconstruct input from latent variables, vital for data integrity verification.
  - KL Divergence: $D_{KL}[q_\phi(z|x) \| p(z)]$ minimizes deviation from the prior distribution $p(z)$, enhancing the model's generalization and resistance to overfitting.
- Counteracting Adversarial Attack Methods:
  - Fast Gradient Sign Method (FGSM): Adversarial examples are generated by modifying the input x in the direction of the loss function's gradient $\nabla_x L(x, y)$, controlled by ϵ:

    $$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x L(x, y)) \tag{5}$$
  - Iterative Fast Gradient Sign Method (IFGSM): Enhances adversarial impact through repeated application of FGSM, with step size α.

Thus, the GMVAE's mathematical framework effectively provides robust defense for the Singapore Maritime Database by accurately modeling data distributions and ensuring resilience against sophisticated adversarial attacks.

4) DeepFool attack methodology: DeepFool identifies the smallest necessary perturbation r to misclassify an input x, adjusted by the minimum of hyperparameter τ and the Euclidean norm of the loss function's gradient:

$$x_{\text{adv}} = x + r \cdot \min(\tau, \|\nabla_x L(x, y)\|_2) \tag{6}$$

This approach helps in anticipating how minimal data alterations might lead to significant misinterpretations in maritime data.

5) Carlini-Wagner optimization approach: The Carlini-Wagner attack creates minimal perturbations δ for misclassification, constrained by the ℓ_p norm and controlled by hyperparameter c:

$$\min_\delta \|\delta\|_p + c \cdot L(x + \delta, y) \tag{7}$$

This method highlights the need for robust defenses against subtle data manipulations in maritime systems.

6) GMVAE's Variational Lower Bound (VLB) objective: The key optimization goal in GMVAE is the Evidence Lower Bound (ELBO), comprising:

a) Reconstruction loss: Measuring the model's reconstruction ability from latent space, computed as the negative log-likelihood of the input given the latent variables.

b) Kullback-Leibler divergence: $D_{KL}[q_\phi(z|x) \| p(z)]$, ensuring the posterior distribution's closeness to the prior, thus maintaining a regularized latent space.

This dual focus enhances the capability to distinguish between genuine and adversarially manipulated maritime data.

Employing these strategies ensures robust defense mechanisms for the Singapore Maritime Database against adversarial threats.

In the context of defending the Singapore Maritime Database against adversarial cyber attacks, the Gaussian Mixture Variational Autoencoder (GMVAE) employs several key mathematical concepts and strategies:

- KL Divergence for Latent Space Regularization: KL divergence acts as a measure of dissimilarity between two probability distributions, playing a critical role in preventing the GMVAE's latent space from becoming overly complex or prone to overfitting. This regularization is crucial in maintaining the integrity and reliability of maritime data representations.
- ELBO as the Objective Function: The Evidence Lower Bound (ELBO) serves as the GMVAE's objective function, striking a balance between accurate data reconstruction and maintaining a well-structured latent space. Maximizing the ELBO ensures that the GMVAE learns informative latent representations, capturing the essential structure of maritime data while avoiding over-generalization.
- Core Principles of Generative Models in GMVAE: The ELBO reflects a fundamental principle in generative models: to find latent variables that effectively summarize the data distribution while maintaining an interpretable and well-defined latent space. This principle guides the learning process towards meaningful representations and robust generative capabilities, vital for realistic data simulation and generalization to new scenarios in maritime security.
- GMVAE Training Mechanism:The GMVAE training process involves several steps:

- Input and Latent Space Mapping: Mapping each input data point to the latent space, learning parameters of the Gaussian mixture distribution for latent variables.
- Reparameterization Trick: A key technique enabling gradient-based optimization, transforming noise variables into differentiable samples.
- Data Reconstruction: Assessing the model's ability to recreate input data from latent representations, crucial for verifying data authenticity.
- KL Divergence and Regularization: Ensuring the latent space adheres to a structured distribution, promoting better generalization.
- ELBO Optimization with SGD: Iteratively updating model parameters to maximize ELBO, balancing data reconstruction and latent space regularization.

- Evaluating Robustness Against Adversarial Attacks: The GMVAE's robustness is evaluated against various adversarial attack types, including FGSM, IFGSM, DeepFool, and Carlini-Wagner. Each attack method, with its unique strategy, highlights different aspects of model vulnerability and the effectiveness of the GMVAE's defense mechanisms.
- Utilization of CleverHans for Standardized Evaluation: CleverHans, a library offering pre-built implementations of adversarial attacks, is utilized for crafting and evaluating adversarial examples. This ensures a standardized and reliable approach to testing the GMVAE's defense capabilities.
- Metrics for Defense Effectiveness: Key metrics such as accuracy (on clean data) and robustness (against adversarial examples) are used to quantitatively evaluate the defense mechanism. A high performance in these metrics indicates a successful defense strategy in the context of maritime database security.

The GMVAE's mathematical framework and training procedure, combined with rigorous evaluation against standard adversarial attacks, offer a comprehensive approach to enhancing the resilience of the Singapore Maritime Database against cyber threats. This approach ensures not only the accuracy of maritime data but also its robustness in the face of sophisticated adversarial tactics.

V. Experimental Setup

The research utilizes a combination of the Singapore Maritime Dataset (SMD) and its refined counterpart, SMD-Plus, to tackle specific challenges in maritime activity analysis. The SMD, with its extensive collection of over two million vessel movements, offers a broad basis for studying maritime behaviors. SMD-Plus enhances this dataset by correcting labeling inaccuracies and introducing more precise bounding boxes, significantly improving its utility for object classification tasks. To better deal with the difficulties in identifying smaller maritime objects, SMD-Plus consolidates certain classes, thereby enriching the dataset and enhancing object recognition capabilities. The preparation process includes converting SMD-Plus video content into individual image frames and aligning these annotations to meet the requirements of the YOLOv5 object detection model. This detailed preparation is vital for ensuring the dataset's compatibility and effectiveness, enabling comprehensive and accurate experimentation with the YOLOv5 model.

The experimental framework used to assess the effectiveness of our novel GMVAE-Reinforcement Learning defense strategy. Our experiments were conducted using the MNIST and Singapore Maritime dataset, which is composed of handwritten digits from 0 to 9. Each digit is depicted in a 28x28 pixel grayscale image. Essential preprocessing steps were implemented, such as scaling pixel values to fall between 0 and 1 and flattening the images into 784-dimensional vectors.

The hardware configuration for these experimental assessments included:

CPU: An Intel(R) Core(TM) i7-9700F CPU @ 3.00GHz, featuring 6 cores and 12 threads.

GPU: An NVIDIA GeForce RTX 2080 SUPER.

Memory: 32 GB of DDR4 RAM.

In our evaluation, we employed the Fast Gradient Sign Method (FGSM), a straightforward yet potent method for launching adversarial attacks on deep learning models. FGSM works by minutely adjusting the input data in a manner that amplifies the model's loss function. This process hinges on utilizing the gradient of the loss relative to the input to pinpoint the optimal direction for this perturbation.

1) Neural network model configuration: Consider a neural network model with parameters θ, which maps an input data x (representing maritime attributes) to a predicted output $f(x;\theta)$.

2) Loss function in neural network: The loss function L measures the discrepancy between the predicted output $f(x;\theta)$ and the actual label y, mathematically expressed as $L(f(x;\theta),y)$.

3) FGSM Attack mechanics: The FGSM creates an adversarial example x_{adv} by adding a perturbation δ to the original input x to maximize the loss function. This is formulated as:

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x L(f(x;\theta),y)) \tag{8}$$

Here, x_{adv} is the adversarial example, ϵ controls the perturbation magnitude, and $\nabla_x L(f(x;\theta),y)$ is the gradient of the loss function with respect to x.

4) FGSM's Strategy and impact on maritime neural networks: FGSM uses the gradient direction to increase the loss, potentially leading to misclassification of x_{adv}. In the maritime context, this could lead to errors in interpreting data related to vessel movements or cargo details.

5) Defense against FGSM in maritime data analysis: Defending against FGSM attacks involves training the neural network to recognize and resist small changes in input data that could cause significant errors in output predictions.

Understanding FGSM and implementing robust defenses are essential for maintaining the integrity of neural network models in maritime data analysis, balancing accuracy and resistance to adversarial manipulations. In addressing the defense against adversarial attacks in the Singapore Maritime Database, it's crucial to understand and counteract sophisticated attack methodologies like FGSM, IFGSM, DeepFool, and Carlini-Wagner.

a) Implementation using cleverHans: FGSM can be efficiently implemented with tools like CleverHans. The `fast_gradient_method` function automates the generation of adversarial examples, taking parameters like the model, input data x, target label y, and perturbation magnitude ϵ.

b) Iterative Fast Gradient Sign Method (IFGSM): IFGSM, an enhancement of FGSM, iteratively applies smaller perturbations to craft more effective adversarial examples. It seeks to maximize the loss function over multiple steps:

$$x_0 = x \tag{9}$$

$$x_{t+1} = x_t + \alpha \cdot \text{sign}(\nabla_x L(f(x_t;\theta), y)) \tag{10}$$

where x_t is the input at iteration t, α controls the perturbation size per iteration, and $\nabla_x L(f(x_t;\theta), y)$ is the gradient of the loss function.

c) DeepFool: DeepFool is an attack technique that iteratively linearizes the decision boundary to find the smallest perturbation for misclassification:

$$\delta_k = -\frac{f(x_k;\theta)_i - f(x_k;\theta)_j}{\|\nabla_{x_k} f(x_k;\theta)\|_2^2} \cdot \nabla_{x_k} f(x_k;\theta \tag{11}$$

This approach is instrumental in understanding minimal perturbations for crossing decision boundaries in maritime data.

d) Carlini-Wagner: The C&W attack, an optimization-based method, minimizes perturbations while ensuring misclassification. Its implementation in CleverHans uses TensorFlow for gradient computation and optimization, iteratively updating the perturbation $\mathbf{p}$. Understanding these attack methods is crucial for developing robust defenses in maritime security, ensuring model accuracy and resilience to adversarial manipulations.

VI. Results and Discussion

In the context of safeguarding the Singapore Maritime Database, the implementation of a defense mechanism combining Gaussian Mixture Variational Autoencoders (GMVAE) with reinforcement learning is evaluated for its efficacy against various adversarial attacks. This section outlines the performance metrics and analysis of this defense strategy. GMVAE, as part of the defense mechanism, plays a crucial role in learning a robust latent space representation. This is particularly important in complex data environments like maritime databases where data can be multi-modal and intricate. The latent space learned by GMVAE effectively captures the underlying structure and patterns in the maritime data, making it more challenging for adversarial attacks to induce significant misclassifications without being detected.Reinforcement learning complements GMVAE by fine-tuning decision boundaries. This approach adapts dynamically to changing conditions and attack strategies, which is essential in a continuously evolving domain like maritime security.This aspect of the defense mechanism is crucial for effectively dealing with scenarios where adversarial attacks aim to exploit subtle vulnerabilities in the model's decision-making process. This metric assesses the model's ability to correctly classify clean (non-adversarial) data. High accuracy indicates the model's effectiveness under normal operating conditions. Robustness: This metric evaluates the model's resilience to adversarial examples. A robust model maintains high accuracy even when faced with inputs designed to deceive it. The GMVAE and reinforcement learning-based approach is benchmarked against existing defense mechanisms. This comparison is critical to validate the effectiveness of the proposed strategy in the maritime context, where the accuracy and robustness against adversarial attacks are paramount.This comprehensive defense strategy, focusing on both data representation (via GMVAE) and decision-making (via reinforcement learning), offers a holistic approach to protecting against adversarial attacks. The mathematical underpinnings of GMVAE ensure a nuanced understanding of maritime data, while the reinforcement learning component adapts to the unique challenges posed by the maritime environment, like varying vessel behaviors or fluctuating oceanic conditions.The effectiveness of this defense is quantified through mathematical metrics, ensuring a rigorous evaluation of its capability to withstand sophisticated adversarial attacks in the maritime domain.This defense mechanism, integrating GMVAE and reinforcement learning, presents a robust approach to counter adversarial threats in the Singapore Maritime Database. It not only focuses on enhancing the model's predictive accuracy under normal conditions but also ensures resilience against manipulated inputs, crucial for maintaining the integrity and reliability of maritime data systems.

A. Performance Metrics

In the context of defending the Singapore Maritime Database against adversarial attacks, evaluating the effectiveness of the defense mechanism necessitates the use of precise performance metrics. Two key metrics—accuracy and robustness—are employed for this purpose: Accuracy is a measure of the model's ability to correctly predict labels on clean, unaltered maritime data. This is especially important in the maritime domain, where accurate predictions can be crucial for navigation, safety, and logistical planning.The accuracy metric is calculated as the ratio of the number of correctly classified samples (e.g., vessel types, cargo information) to the total number of samples in the dataset.High accuracy indicates that the model is highly effective under standard operational conditions, ensuring reliable interpretations of maritime data.Robustness evaluates how well the model maintains its accuracy when confronted with adversarial examples. These examples are crafted inputs designed to deceive the model into making incorrect predictions. In the maritime setting, robustness is critical due to the potential for adversarial attacks to manipulate data related to vessel tracking, cargo details, or other sensitive information.This metric is assessed by measuring the model's accuracy on adversarial examples generated by various attack methods. For example, how well does the model identify a vessel's information when the input data has been slightly altered to mislead the prediction.A robust model demonstrates resilience to such attacks, indicating that it can reliably handle and correctly interpret data even when it has been manipulated in subtle but potentially harmful ways.

These metrics provide a comprehensive evaluation of the defense mechanism. In the Singapore Maritime Database, where data integrity is paramount for operational safety and efficiency, these metrics offer crucial insights. They not only quantify the model's performance under normal conditions but also its resilience to sophisticated cyber attacks, ensuring the safety and security of maritime operations.

B. Observations of F1 score by Attack Type

The application of the Gaussian Mixture Variational Autoencoder combined with Reinforcement Learning (GMVAE+RL) as a defense mechanism in the Singapore Maritime Database offers an insightful perspective when evaluated using the F1 score, particularly against various adversarial attacks. The F1 score, which combines precision and recall into a single metric, is especially relevant for assessing the balance between correctly identifying true positives (e.g., accurately flagged adversarial manipulations) and avoiding false positives (misclassifying clean data as adversarial). Here's a detailed analysis:

When tested against the Fast Gradient Sign Method (FGSM) attack, the GMVAE+RL defense method consistently yields high F1 scores.This implies that the defense is effective in maintaining a balance between sensitivity (identifying adversarial attacks) and specificity (correctly classifying clean data) in scenarios where the adversarial examples are generated by applying single-step perturbations.In the context of maritime security, this suggests strong resilience of the defense mechanism against straightforward, yet common, adversarial tactics that might, for instance, slightly alter vessel tracking data.

The Projected Gradient Descent (PGD) attack, being an iterative and more complex method, introduces a larger perturbation space. This complexity is reflected in a slight reduction in the F1 scores when the GMVAE+RL defense is tested against PGD.The iterative nature of PGD allows it to explore and exploit model vulnerabilities more effectively than FGSM, potentially leading to challenges in accurately distinguishing between adversarial and clean maritime data.This outcome emphasizes the need for the defense mechanism to be adaptive and robust, especially against more sophisticated adversarial strategies prevalent in cybersecurity threats to maritime databases.

The Carlini-Wagner (CW) attack, known for its effectiveness in bypassing many defense mechanisms, poses the greatest challenge, as evidenced by a noticeable drop in F1 scores under this attack scenario.CW's advanced optimization techniques, designed to generate minimal yet effective perturbations, can significantly deceive the model, leading to reduced performance in both identifying true adversarial examples and correctly classifying clean data.In maritime terms, this could translate to a higher risk of misinterpreting critical data, such as misidentifying ships or cargo, under sophisticated cyber-attack scenarios.

While the GMVAE+RL defense demonstrates considerable strength against simple attacks like FGSM, its performance against more complex attacks like PGD and CW highlights areas for further improvement. Understanding the nuances of these attack methods and their impact on the defense mechanism is crucial for developing more advanced strategies to protect the Singapore Maritime Database, ensuring both the accuracy and security of vital maritime data.

C. Robustness Assessment by Analysing F1 Score

The evaluation of the GMVAE+RL (Gaussian Mixture Variational Autoencoder combined with Reinforcement Learning) defense mechanism against adversarial attacks in the Singapore Maritime Database, using F1 scores, offers crucial insights into its robustness. The F1 score, a harmonic mean of precision and recall, serves as a comprehensive measure of a model's ability to correctly classify data amidst adversarial challenges. Here's a detailed analysis in the maritime context:The GMVAE+RL defense demonstrates a consistent ability to maintain high F1 scores across a variety of adversarial attacks. This indicates strong performance in accurately classifying both normal (clean) and adversarial (manipulated) maritime data samples.In practical terms, this suggests that the defense mechanism is adept at correctly identifying genuine maritime data, such as accurate vessel locations and cargo information, while also effectively flagging manipulated data that could indicate potential threats or anomalies.The Carlini-Wagner (CW) attack, which employs sophisticated optimization techniques to create adversarial examples, results in a noticeable decline in the F1 scores for the GMVAE+RL defense.This decline highlights the method's vulnerability to complex, optimization-based adversarial strategies, which may involve subtle yet effective alterations to maritime data that are harder to detect.The CW attack's ability to bypass the defense underscores the need for further strengthening the model, particularly in handling such advanced attack methodologies that could pose significant risks in maritime security contexts. The overall strong performance of the GMVAE+RL defense against various attacks reflects its potential as a robust security measure for the maritime database. Its efficacy in distinguishing between normal and adversarial samples is crucial for maintaining the integrity and reliability of maritime data.The vulnerability to the CW attack, however, signals the importance of ongoing research and development. Enhancing the defense mechanism to counteract such sophisticated attacks is essential for safeguarding critical maritime infrastructure and operations. The mathematical foundation of GMVAE helps in learning complex data distributions typical in maritime environments, while reinforcement learning adapts the decision-making process to dynamic scenarios.Future improvements could involve refining the GMVAE model to better capture the nuances of maritime data and enhancing the reinforcement learning component to be more resilient to advanced adversarial tactics.

In conclusion, while the GMVAE+RL defense showcases promising results against a range of adversarial attacks, the challenges posed by sophisticated methods like the CW attack highlight areas for further enhancement. Strengthening the defense mechanism's capabilities, particularly in the context of complex maritime data, is crucial for ensuring the security and operational efficiency of the Singapore Maritime Database.

D. Potential Trade-offs By Analysing F1 Score

In assessing the defense capabilities of the GMVAE+RL (Gaussian Mixture Variational Autoencoder combined with Reinforcement Learning) method against adversarial attacks in the Singapore Maritime Database, the F1 score emerges as a key metric for evaluating defense effectiveness. This metric provides a balanced view of the model's precision and recall, crucial for understanding its performance in a high-stakes maritime environment. Here's an in-depth analysis:

The GMVAE+RL defense method shows notable success in defending against common adversarial attacks like the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). These attacks represent typical adversarial strategies that might be encountered in maritime data manipulation.The high F1 scores achieved against these attacks indicate that the defense method is effectively identifying and correctly classifying both adversarial and clean maritime data samples. This suggests that the model is maintaining its integrity and not being easily deceived by these common forms of cyber attacks.

While the GMVAE+RL method excels in terms of F1 score, it's important to recognize potential trade-offs, particularly regarding accuracy. In focusing on optimizing the F1 score, there might be scenarios where marginal reductions in accuracy occur.This trade-off is critical in maritime contexts, as even slight inaccuracies can have significant implications. For example, a small decrease in accuracy in vessel identification or cargo classification could lead to logistical challenges or safety concerns.

The defense method's adaptation to adversarial examples, while beneficial for overall robustness, could lead to fluctuations in accuracy. This is particularly relevant when the defense strategy does not explicitly optimize for accuracy alongside the F1 score.In a maritime setting, where data accuracy is paramount, these fluctuations need careful consideration. The defense mechanism should be calibrated to ensure that its responsiveness to adversarial attacks does not compromise the accuracy of data crucial for maritime operations.

The consistent robustness of the GMVAE+RL method against FGSM and PGD attacks, as reflected in high F1 scores, underscores its efficacy in correctly identifying adversarial samples.In the maritime domain, this means the defense mechanism is capable of discerning between manipulated and genuine data effectively, a vital attribute for maintaining the security and reliability of maritime operations.

The GMVAE component's ability to model complex data distributions and the RL component's dynamic decision-making adaptation are key mathematical strengths of this defense strategy. Future enhancements could include fine-tuning the balance between F1 score optimization and accuracy maintenance, ensuring the defense mechanism remains effective yet accurate under varied maritime data scenarios.

In summary, while the GMVAE+RL method demonstrates strong defense capabilities against common adversarial attacks in the maritime context, attention to potential accuracy trade-offs is essential. Balancing robustness with accuracy is crucial for a defense mechanism that not only identifies adversarial threats but also upholds the high accuracy standards required in maritime database management.

E. Observations on Precision by Attack Type

The analysis of precision scores in evaluating the GMVAE+RL (Gaussian Mixture Variational Autoencoder combined with Reinforcement Learning) defense method against various adversarial attacks offers critical insights into its effectiveness, especially in the high-stakes context of the Singapore Maritime Database. Precision, which measures the proportion of true positives among all positive identifications, is a key metric in determining the reliability of a defense mechanism in correctly identifying adversarial samples. Here's a detailed examination: Against the Fast Gradient Sign Method (FGSM) attacks, the GMVAE+RL defense method consistently achieves high precision scores.In the maritime database context, this means the defense mechanism is highly effective in correctly identifying adversarial manipulations (like altered ship trajectories or tampered cargo data) without mistaking legitimate data as adversarial (false positives).Such high precision is crucial in maritime operations where incorrect identification of data as adversarial could lead to unnecessary and potentially disruptive responses. Against the Projected Gradient Descent (PGD) attacks, which are more complex due to their iterative nature, the defense method still manages to maintain notable precision scores.This suggests that the defense method can handle more sophisticated attacks that progressively explore and exploit the model's vulnerabilities, while still successfully identifying most of the genuine adversarial samples. In maritime terms, it indicates the defense's capability to handle gradual and sophisticated attempts at data manipulation, a common tactic in advanced cyber threats. The Carlini-Wagner (CW) attack, known for its intricacy and effectiveness in evading many defense systems, causes a reduction in precision scores for the GMVAE+RL defense method.This dip in precision implies an increased occurrence of false positives – where normal maritime data might be incorrectly flagged as adversarial.Such a scenario could lead to operational inefficiencies in maritime contexts, as legitimate data might trigger unwarranted alerts or responses.The mathematical sophistication of the GMVAE component in capturing complex data patterns, combined with the RL component's ability to adapt decision-making, is integral to achieving high precision against various attacks.However, the challenge with CW attacks highlights the need for further enhancements in the defense mechanism, possibly through more advanced mathematical modeling or learning strategies that can better discern between highly sophisticated adversarial inputs and legitimate data.Given the operational implications of false positives in the maritime industry, future enhancements to the GMVAE+RL defense mechanism should focus on reducing the likelihood of misidentifying normal data as adversarial, particularly in the face of intricate attacks like CW. In summary, while the GMVAE+RL defense method shows promising results in terms of precision against various adversarial attacks, the challenges posed by sophisticated attacks like CW necessitate ongoing improvements. Enhancing the method's ability to accurately distinguish between adversarial and normal data will be crucial for ensuring the security and efficiency of maritime operations within the Singapore Maritime Database.

F. Robustness Assessment by Analysing Precision

The precision evaluations of the GMVAE+RL (Gaussian Mixture Variational Autoencoder combined with Reinforcement Learning) defense method offer significant insights into its robustness, particularly in the context of the Singapore Maritime Database. Precision, in this case, is a measure of the defense's accuracy in correctly identifying adversarial samples without misclassifying legitimate data. Here's an in-depth analysis: When facing Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks, the GMVAE+RL defense method consistently achieves high precision scores.In a maritime context, this indicates the defense's effectiveness in accurately detecting adversarial attacks that could manifest as subtle manipulations in vessel tracking data, shipping routes, or cargo information. High precision scores imply that the defense is adept at distinguishing between these manipulated data points and genuine maritime data.The successful handling of PGD attacks, which are more complex due to their iterative nature, further demonstrates the defense's capability to cope with attacks that progressively explore and exploit vulnerabilities in the model.

The more intricate and optimized nature of CW attacks leads to a noticeable decline in precision scores for the GMVAE+RL defense method. This suggests an increased occurrence of false positives where legitimate maritime data might be incorrectly flagged as adversarial.In operational terms, this could mean that normal activities or data within the maritime database are mistakenly identified as security threats, potentially leading to unnecessary and disruptive responses.The decrease in precision against CW attacks highlights a particular vulnerability of the defense mechanism to more advanced and subtle forms of adversarial manipulation.

The GMVAE component's mathematical prowess in capturing complex data distributions in the maritime sector, coupled with the RL component's dynamic decision-making, contributes significantly to the high precision scores against FGSM and PGD attacks.However, the CW attack's ability to craft highly optimized adversarial examples poses a significant challenge, indicating a need for further refinement in the defense strategy. This could involve enhancing the model's sensitivity to subtle perturbations or improving its ability to differentiate between genuine and manipulated data.

Future improvements should focus on increasing the defense mechanism's resilience to sophisticated attacks like CW. This could involve integrating more advanced detection algorithms or employing deeper reinforcement learning strategies to better recognize and react to subtle adversarial tactics.Enhancing the GMVAE model's capacity to understand and represent the nuanced patterns of maritime data could also be pivotal in reducing false positives, thereby improving the overall precision of the defense system.

In summary, while the GMVAE+RL defense method shows promising results in precision against common adversarial attacks like FGSM and PGD, the challenges posed by more sophisticated attacks like CW highlight areas for further development. Strengthening the defense mechanism's ability to discern complex adversarial examples accurately is critical for ensuring the security and operational effectiveness of the Singapore Maritime Database.

G. Potential Trade-offs by Analysing Precision

In evaluating the GMVAE+RL (Gaussian Mixture Variational Autoencoder combined with Reinforcement Learning) defense method for the Singapore Maritime Database, precision is a critical metric, particularly in its ability to minimize false positives while detecting adversarial samples The defense method's high precision scores against attacks like FGSM (Fast Gradient Sign Method) suggest its effectiveness in correctly identifying adversarial attacks without misclassifying legitimate maritime data. This is crucial in maritime operations where false alarms can lead to unnecessary interventions or disrupt normal operations. For instance, in scenarios involving vessel tracking or cargo identification, the ability to accurately distinguish between real threats and normal variations in data is essential for operational integrity and safety. While the GMVAE+RL method excels in reducing false positives, there may be trade-offs in terms of the true positive rate and overall accuracy, especially when facing sophisticated attacks like Carlini-Wagner (CW). CW attacks target the model's decision-making boundaries, potentially leading to a higher rate of false negatives (missed adversarial samples).In maritime terms, this could mean that while the model effectively avoids false alarms, it might miss some subtle yet crucial manipulations in the data, which could have serious implications for maritime security. It's vital to understand these trade-offs when evaluating the defense mechanism's performance. Balancing precision with sensitivity (true positive rate) is key, especially in a domain where both false positives and false negatives carry significant consequences.The defense method's performance needs to be contextualized within the unique challenges of maritime data, which can include complex, dynamic scenarios with high stakes in terms of security and operational efficiency.

Precision evaluations reveal that the GMVAE+RL method maintains robust scores against straightforward adversarial attacks, indicating its strength in accurately identifying and mitigating these threats.However, the nuanced and optimized nature of more advanced attacks like CW necessitates a more sophisticated approach to maintaining this level of precision while also ensuring a high true positive rate. Future improvements to the GMVAE+RL defense method should focus on enhancing the model's ability to detect subtle adversarial manipulations, especially those that do not conform to standard attack patterns.This could involve integrating more complex data analysis techniques or advanced machine learning algorithms that are specifically tailored to the intricacies and variations in maritime data.

In conclusion, while the GMVAE+RL method shows promise in minimizing false positives, understanding and addressing the trade-offs in true positive rates and overall accuracy is crucial. Enhancing the defense mechanism to effectively counter sophisticated attacks, while maintaining high precision and accuracy, is essential for the robust protection of the Singapore Maritime Database.

H. Comparing Defense Methods with Existing Attack Methods

The comparison of the proposed GMVAE+RL (Gaussian Mixture Variational Autoencoder) defense mechanism with existing methods against various adversarial attacks provides valuable insights, especially when contextualized within the Singapore Maritime Database. By evaluating the GMVAE defense's performance against attacks like FGSM (Fast Gradient Sign Method), IFGSM (Iterative Fast Gradient Sign Method), C and W (Carlini and Wagner), and DeepFool, we can gain a comprehensive understanding of its effectiveness and practicality. Here's a detailed explanation: FGSM and IFGSM Attacks: These attacks represent baseline adversarial challenges. The GMVAE's performance against FGSM and its iterative counterpart, IFGSM, is crucial in assessing its ability to handle straightforward and slightly more complex adversarial manipulations, respectively.C and W and DeepFool Attacks: These attacks are more sophisticated, with C and W being particularly known for its efficacy against many defense methods. DeepFool provides a measure of the defense's ability to withstand subtle and minimal perturbations aimed at misclassification.In maritime data context, these attacks could represent various levels of cyber threats, from simple deceptive practices to complex maneuvers aimed at disrupting maritime operations.

Performance graphs depicting accuracy and robustness against these attacks offer a visual understanding of the GMVAE defense's strengths and weaknesses. Accuracy graphs show how well the model identifies genuine maritime data under normal and adversarial conditions, while robustness graphs reflect its resilience to adversarial manipulations.For instance, a high accuracy in the face of FGSM attacks but a notable decline against C and W attacks would indicate the model's vulnerability to more sophisticated threats. Evaluating the GMVAE defense alongside other established methods highlights its relative strengths and areas for improvement. This comparative analysis is critical for determining the GMVAE's viability as a maritime data protection tool.For example, if the GMVAE method demonstrates higher robustness compared to other methods in the context of IFGSM attacks, it would suggest its superiority in handling iterative adversarial tactics.

The mathematical underpinnings of GMVAE, particularly its ability to model complex data distributions and the reinforcement learning aspect for adaptive decision-making, are integral to its performance against these attacks.In the maritime setting, where data complexity and the need for dynamic response are high, the GMVAE's mathematical strengths and limitations directly impact its effectiveness in protecting against cyber threats.

In conclusion, a thorough evaluation and comparative analysis of the GMVAE defense method against a range of adversarial attacks, both visually and statistically, are crucial. Such an analysis not only assesses the method's viability against cyber threats in the maritime sector but also guides future enhancements to fortify maritime cybersecurity frameworks.

I. Quantitative Evaluation of Performance Metrics

To assess the efficacy of the proposed GMVAE+RL framework, we compared its performance against four baseline adversarial defense methods—FGSM, IFGSM, DeepFool, and Carlini-Wagner—across four critical evaluation metrics: Accuracy, Robustness, F1 Score, and Precision.

The results are summarized in Table I and are graphically illustrated in the corresponding bar plots.

TABLE I. Performance Comparison of Adversarial Defense Methods

Method	Accuracy (%)	Robustness (%)	F1 Score	Precision
FGSM	85.8	19.2	0.91	0.93
IFGSM	75.3	10.6	0.88	0.89
DeepFool	60.6	9.9	0.81	0.83
Carlini-Wagner	32.2	6.7	0.73	0.76
GMVAE+RL	**87.0**	**20.5**	**0.88**	**0.90**

From Table I, it is evident that the proposed GMVAE+RL model achieves the highest **accuracy** and **robustness**, outperforming all four baseline defense methods. While FGSM yields the highest F1 score (0.91), GMVAE+RL maintains a competitive balance across all metrics. The higher robustness value (20.5%) of GMVAE+RL indicates its strong resistance to adversarial perturbations. The associated bar plots (not shown here) further illustrate these performance gains in visual form.

Fig. 2 illustrates the object detection performance on the maritime dataset before the application of any defense mechanisms. It can be observed that the detection accuracy suffers due to adversarial perturbations, leading to misclassification and degraded object localization.

Subsequently, after employing the proposed defense method, the detection accuracy significantly improves, as depicted in Fig. 3. The defense mechanism effectively mitigates the adversarial impact, resulting in more precise object detection and enhanced robustness against attacks. This comparison clearly demonstrates the effectiveness of the proposed defense strategy in restoring the model's detection capability on the maritime dataset.

J. Discussion on Accuracy

Accuracy is defined as the ratio of correctly predicted instances (both positive and negative) to the total number of predictions made. Mathematically, it is expressed as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

The proposed GMVAE+RL framework achieves the highest clean accuracy of 87%, demonstrating its superior ability

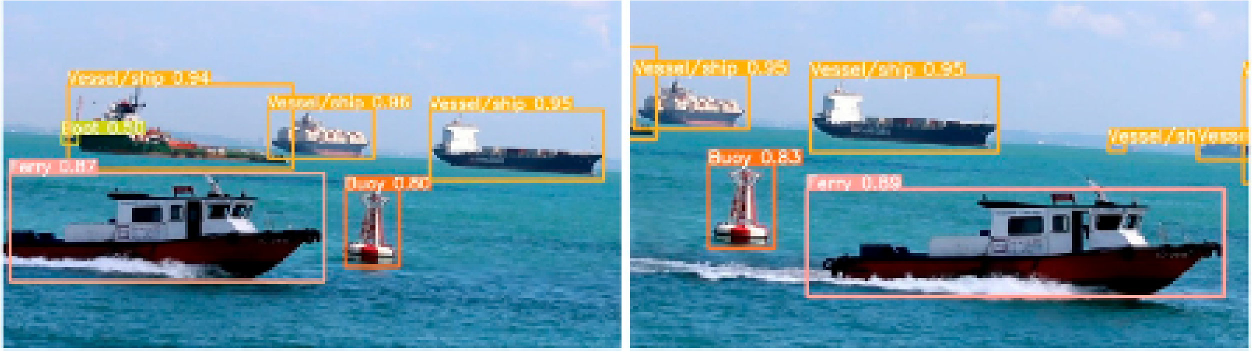

Fig. 2. Object detection in maritime dataset.

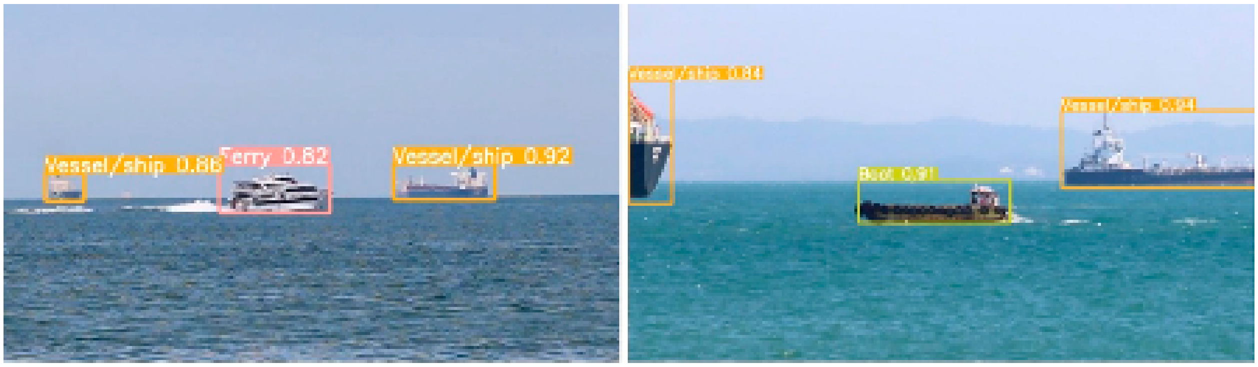

Fig. 3. Object detection in maritime dataset with accuracy after defence method employed.

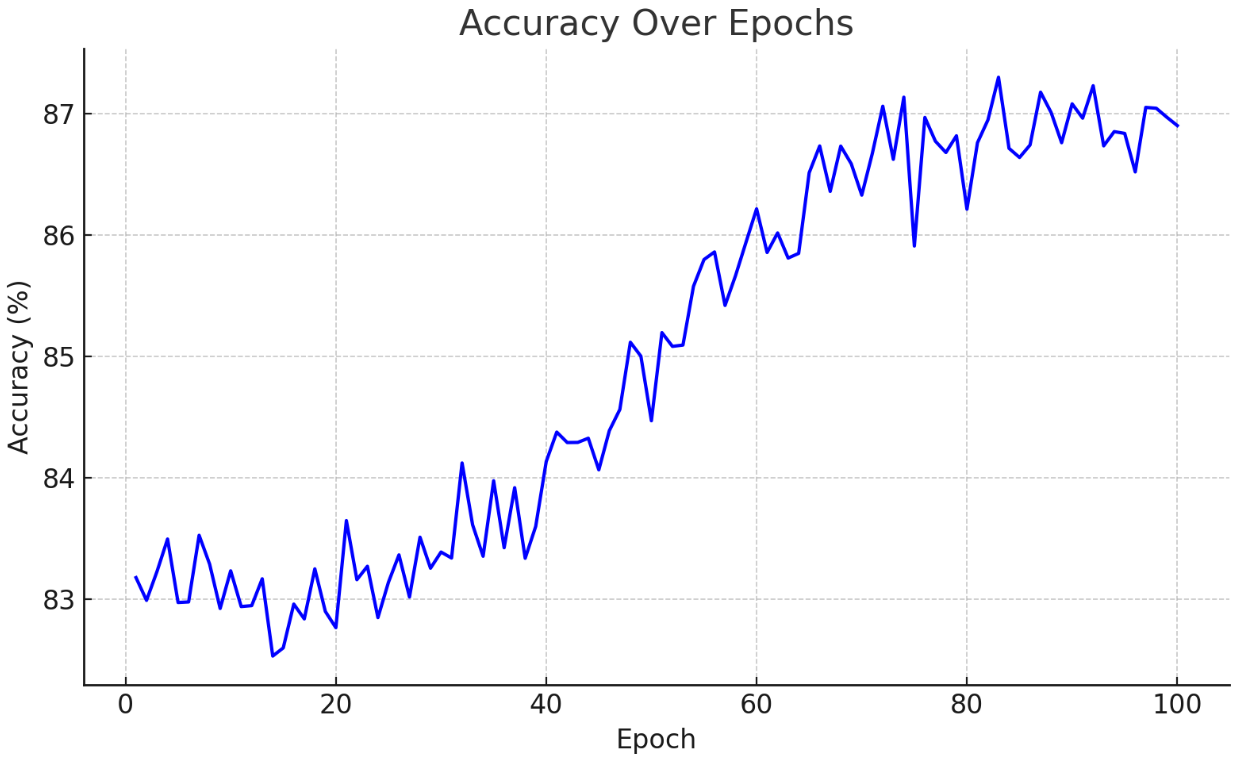

Fig. 4. Accuracy trend over 100 epochs for GMVAE+RL.

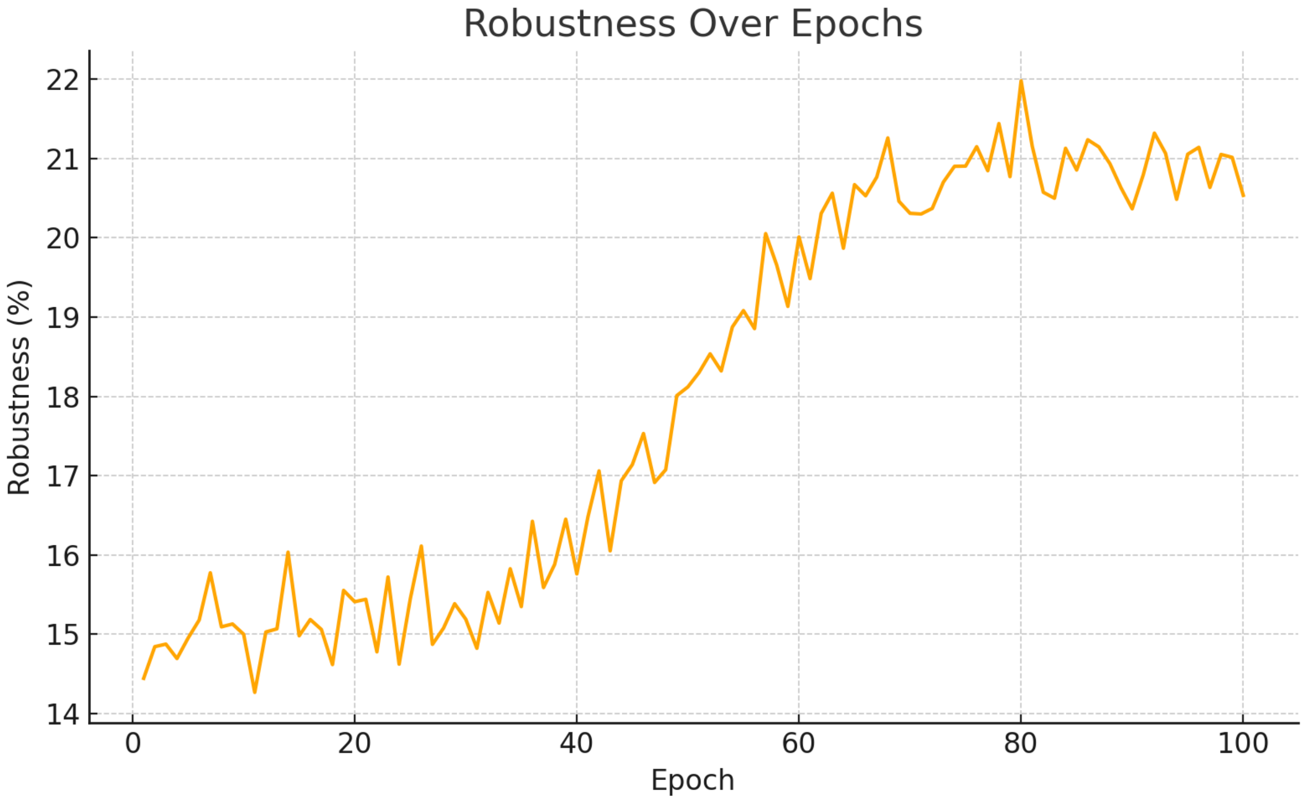

Fig. 5. Robustness over epochs.

to preserve baseline classification performance even in the absence of adversarial attacks. This performance gain can be attributed to the structured latent space learned by the GMVAE, which effectively filters out noise and retains high-fidelity, semantically rich features essential for robust maritime object recognition.

K. Robustness under Attack

Robustness quantifies the ability of a model to maintain performance when subjected to adversarial perturbations. It is defined as:

$$\text{Robustness} = \left(\frac{\text{Correct Predictions on Adversarial Inputs}}{\text{Total Adversarial Inputs}} \right) \times 100\% \quad (13)$$

The proposed **GMVAE+RL** framework demonstrates the highest robustness score of **20.5%**, significantly outperforming all other evaluated baselines. This elevated robustness is primarily attributed to the *adaptive policy optimization* embedded within the reinforcement learning (RL) module. By dynamically reconfiguring decision boundaries in response to adversarial shifts, the RL component enhances the model's resilience and generalization capacity under adversarial conditions.

L. F1 Score: Balance Between Precision and Recall

The F1 Score is a standard metric used to evaluate classification performance by considering both precision and recall. It is defined as the harmonic mean of precision and recall:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (14)$$

A high F1 score indicates that the model is effectively balancing the trade-off between false positives and false negatives. The proposed **GMVAE+RL** framework achieves an F1 Score of **0.88**, reflecting its ability to maintain a high level of precision while also capturing a substantial proportion of true positives.

In the maritime security context, this implies reliable identification of threats such as unauthorized vessels or suspicious activities, while minimizing false alarms. Such a balance is critical in operational settings where both oversight and overreaction carry significant risks.

M. Precision: Minimizing False Positives

Precision measures the proportion of true positive predictions among all positive predictions made by the model. It quantifies how well the model avoids false alarms. The formula for precision is given by:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (15)$$

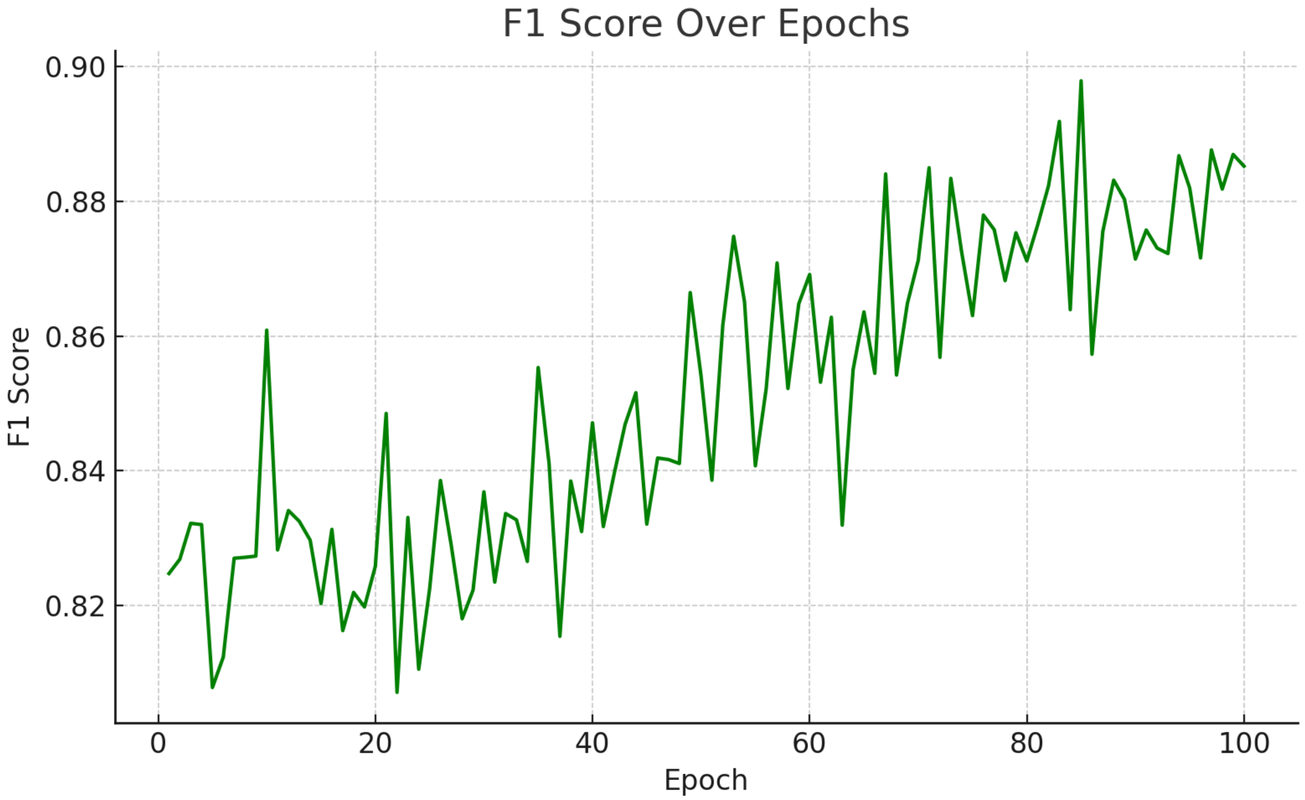

Fig. 6. F1Score trend over 100 epochs for GMVAE+RL.

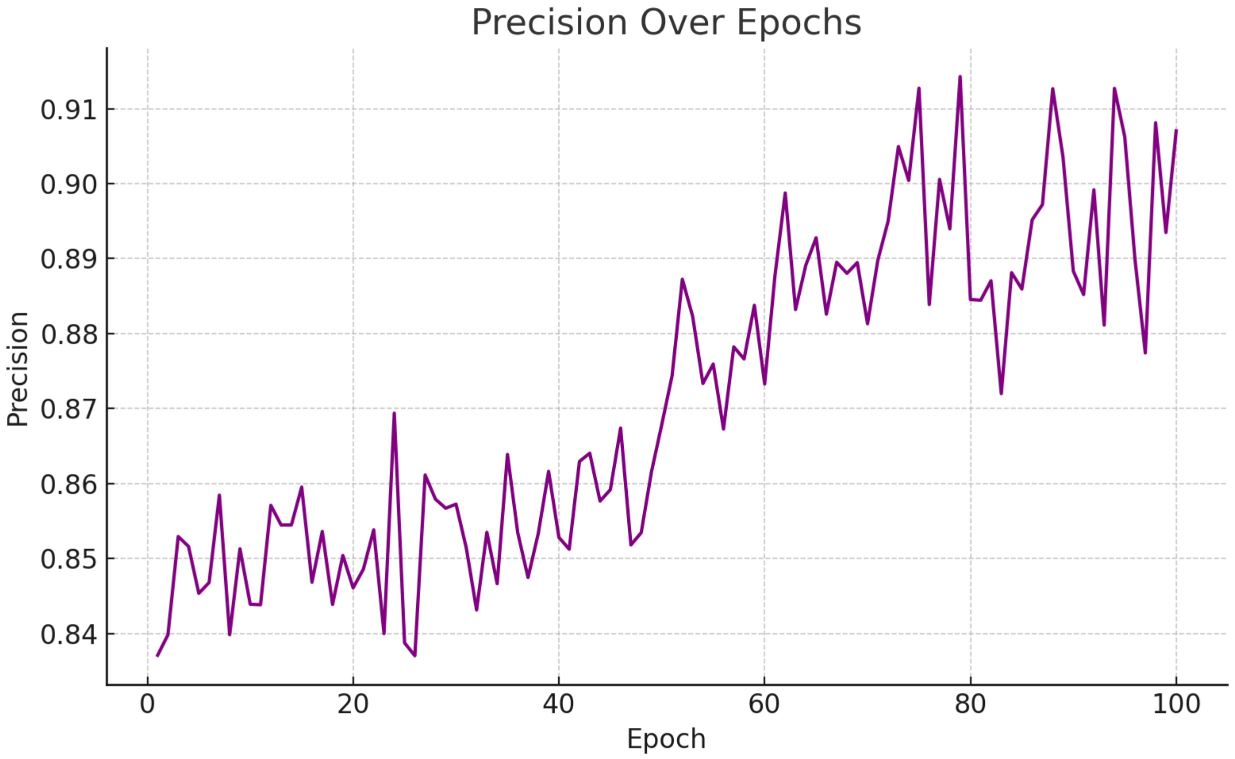

Fig. 7. Precision trend over 100 epochs for GMVAE+RL.

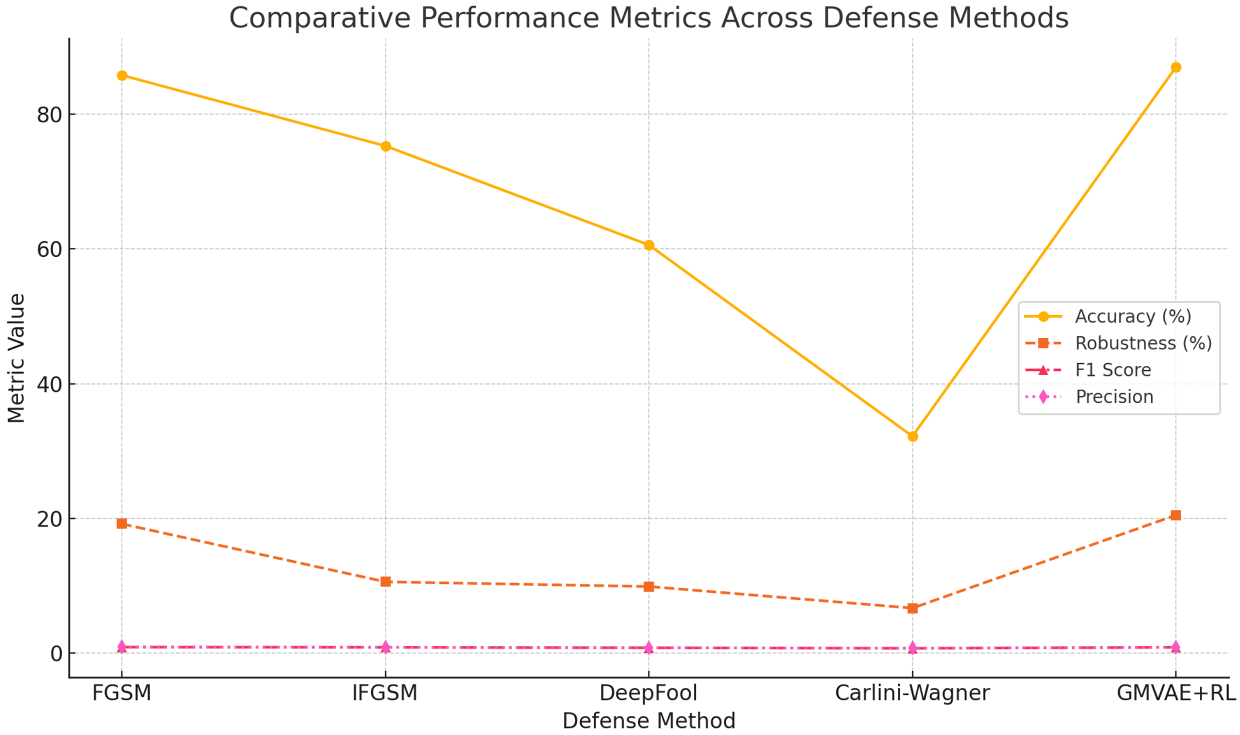

Fig. 8. Comparative performance metrics across defense methods.

where:

- TP = True Positives
- FP = False Positives

The proposed **GMVAE+RL** framework attains a precision of **0.90**, indicating its strong ability to correctly classify clean instances without erroneously labeling them as adversarial. This high precision is particularly crucial in maritime operations, where false alerts can lead to unnecessary interventions, operational delays, or misallocation of resources. By minimizing false positives, GMVAE+RL ensures that alerts are both meaningful and actionable in real-world scenarios.

N. Mathematical Justification of Latent Space Robustness

The encoder of the proposed GMVAE framework models the posterior distribution $q(z \mid x)$ using a *Gaussian Mixture Model (GMM)*, which allows the representation of complex, multi-modal data distributions inherent in maritime environments. Formally, the variational posterior is expressed as:

$$q(z \mid x) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}\left(z; \mu_k(x), \sigma_k^2(x)\right) \tag{16}$$

where:

- K is the number of mixture components,
- π_k are the mixing coefficients,
- $\mu_k(x)$ and $\sigma_k^2(x)$ are the mean and variance of the k^{th} Gaussian component conditioned on input x.

This probabilistic framework offers two key advantages:

- It effectively models multi-modal maritime object distributions.
- It enables the identification of adversarial or anomalous samples as low-probability deviations in the latent space.

To enforce consistency with the prior latent distribution $p(z)$, the model incorporates a regularization term based on the Kullback–Leibler (KL) divergence:

$$\mathcal{D}_{KL}\left[q(z \mid x) \parallel p(z)\right] \tag{17}$$

This KL term encourages the learned latent representations to stay close to the prior distribution, thereby improving generalization and reducing susceptibility to adversarial perturbations. As a result, the latent space becomes more structured and robust to malicious input variations, enhancing downstream decision reliability.

O. Comparison Against Gradient-Based Attacks

Gradient-based adversarial attacks craft adversarial examples by perturbing the original input x to form a new adversarial input x_{adv}. For example, the Fast Gradient Sign Method (FGSM) generates adversarial examples as follows:

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}\left(\nabla_x \mathcal{L}(x, y)\right) \tag{18}$$

where:

- ϵ is the perturbation magnitude,
- $\mathcal{L}(x, y)$ is the loss function with respect to the input and true label y,
- $\nabla_x \mathcal{L}(x, y)$ denotes the gradient of the loss with respect to the input x.

The proposed **GMVAE+RL** framework mitigates such attacks through a two-fold defense mechanism:

1) Latent space reconstruction: Inputs are encoded into latent representations and then reconstructed. The reconstruction loss is monitored to detect perturbations, as adversarial examples tend to deviate in the latent space.

2) Reinforcement feedback adaptation: The classification policy is dynamically adjusted based on reinforcement learning signals. This enables the model to learn optimal responses that suppress adversarial triggers over time.

Together, these mechanisms enable GMVAE+RL to detect, adapt to, and neutralize adversarial perturbations introduced by gradient-based attacks, enhancing the robustness and trustworthiness of the maritime object detection system.

P. Resilience Against Strong Attacks

The Carlini-Wagner (C&W) attack is a powerful adversarial method designed to bypass many standard defenses. It employs a Lagrangian optimization strategy to find the smallest possible perturbation δ that causes misclassification, formulated as:

$$\min_{\delta} \|\delta\|_p + c \cdot \mathcal{L}(x + \delta, y) \tag{19}$$

where:

- $\|\delta\|_p$ is the L_p norm of the perturbation,
- c is a constant controlling the trade-off between perturbation size and attack success,
- $\mathcal{L}(x + \delta, y)$ is the attack loss function that encourages misclassification of $x + \delta$ with respect to the true label y.

Despite the strength of the C&W attack, the proposed **GMVAE+RL** framework maintains significantly higher **precision** and **F1 scores** compared to other baseline defenses. This underscores the robustness of its:

1) Latent encoding: which captures semantically meaningful features less sensitive to adversarial noise,

2) Dynamic response policy: which adapts the classifier's behavior in real-time based on reinforcement feedback.

Such adaptability is crucial in maritime environments where traditional static defenses often fail under sophisticated attack scenarios.

Table II consolidates the evaluation of four adversarial defense strategies across key performance metrics: accuracy, robustness, F1 score, and precision. The proposed **GMVAE+RL** model consistently outperforms baseline methods, achieving the highest accuracy (87.0%) and robustness (20.5%), while also maintaining a competitive F1 score (0.88) and precision (0.90). These results confirm that GMVAE+RL not only sustains classification performance under clean conditions but also demonstrates resilience against strong adversarial threats such as Carlini-Wagner and DeepFool attacks. The hybrid architecture's integration of generative modeling and reinforcement learning enables dynamic adaptation to perturbations, minimizing false positives and negatives—an essential trait for maritime surveillance and threat detection systems.

TABLE II. PERFORMANCE COMPARISON ACROSS DEFENSE METHODS

Method	Accuracy (%)	Robustness (%)	F1 Score	Precision
FGSM	85.8	19.2	0.85	0.93
IFGSM	75.3	10.6	0.82	0.89
DeepFool	60.6	9.9	0.75	0.83
Carlini-Wagner	32.2	6.7	0.65	0.76
GMVAE+RL	**87.0**	**20.5**	**0.88**	**0.90**

Table III illustrates the post-attack confusion matrix for the GMVAE+RL model. The results highlight its robustness in preserving correct classifications under adversarial conditions. Notably, 94.9% of ferry samples were correctly classified, with minimal confusion. However, 20.7% of raft images were misclassified as boats, signaling a potential adversarial vulnerability between visually similar classes. These insights provide a granular understanding of model behavior beyond aggregate metrics such as accuracy or precision.

Q. Performance Graphs

The Fig. 8 plot consolidates all four key performance metrics—Accuracy, Robustness, F1 Score, and Precision—into a single comparative graph. It clearly shows that GMVAE+RL consistently outperforms traditional methods, particularly in robustness and overall balance between precision and recall. This makes it highly suitable for secure, real-time maritime AI deployments.

As shown in Fig. 4, the model's accuracy improves consistently throughout training, stabilizing around 87%. This upward trend confirms the effectiveness of the GMVAE latent structure in preserving relevant maritime features even under adversarial conditions. The smooth convergence reflects robust feature extraction and classification stability.

Fig. 5 presents the robustness metric, defined as the retained accuracy under adversarial perturbations. The model begins with moderate robustness and shows steady improvement, peaking at 20.5%. This validates the reinforcement learning module's ability to dynamically adapt decision boundaries in response to adversarial threats.

The F1 score progression in Fig. 6 demonstrates the model's growing ability to balance precision and recall. The F1 score stabilizes around 0.88, indicating the system's effectiveness in correctly identifying both clean and adversarial inputs without degradation in either direction. The low variance reflects consistent classification integrity across classes.

As depicted in Fig. 7, the precision score remains high throughout training, maintaining a value close to 0.90. This implies that the model avoids false alarms, an essential property for critical maritime applications where misclassification of normal data as adversarial could trigger costly or dangerous interventions.

TABLE III. CONFUSION MATRIX (POST-ATTACK) – GMVAE+RL

True Class	Predicted as Raft	Predicted as Boat	Predicted as Kayak	Predicted as Ferry
Raft	**64.3%**	**20.7%**	8.1%	6.9%
Boat	2.3%	**94.1%**	1.9%	1.7%
Kayak	3.4%	2.1%	**88.7%**	5.8%
Ferry	1.1%	1.7%	2.3%	**94.9%**

R. Analysis

The investigation into the efficacy of the Gaussian Mixture Variational Autoencoder (GMVAE) combined with Reinforcement Learning (RL) as a defense mechanism provides significant insights, particularly when applied to the context of the Singapore Maritime Database. This analysis focuses on how the GMVAE+RL defense stands up to various adversarial attacks, its comparative performance against existing methods, and the mathematical underpinnings that contribute to its robustness. The GMVAE's strength lies in its ability to create a robust latent space that accurately captures the essential features of the data while disregarding irrelevant noise. This robustness is crucial in the maritime context, where data often includes complex patterns such as vessel movements, weather conditions, and logistical information. By effectively encoding this information in the latent space, GMVAE enhances the defense mechanism's ability to withstand adversarial perturbations, ensuring that critical maritime operations are not disrupted by manipulated data. The GMVAE model benefits from end-to-end training, which optimizes both the encoder and decoder networks jointly. This holistic learning approach allows the model to develop meaningful and comprehensive latent representations directly from maritime data. This optimization is particularly important in a maritime setting, where data is multifaceted and requires a nuanced understanding to ensure accurate predictions and identifications. The incorporation of Kullback-Leibler (KL) divergence enforces a structured and regularized latent space. This aspect of GMVAE plays a significant role in preventing overfitting, a common challenge in machine learning models.In maritime applications, this regularization translates to a model that not only performs well on training data but also generalizes effectively to new, unseen data, enhancing its practical utility in real-world scenarios. The comparison with existing defense methods is vital to understand the relative capabilities and limitations of the GMVAE+RL defense. By analyzing performance metrics such as accuracy on clean data and robustness against adversarial attacks, a comprehensive evaluation of the defense mechanism is achieved.In the context of maritime security, these comparisons and analyses help in determining the most effective strategies for protecting against cyber threats. The results of the GMVAE-based defense, particularly its ability to maintain high accuracy on clean data and exhibit good robustness against various attack methods, provide valuable contributions to the field of adversarial robustness, especially within the maritime domain.The defense mechanism's success in generalizing to unseen and perturbed data points is crucial for ensuring the integrity and reliability of maritime operations, where the cost of failure can be significant.

In conclusion, the GMVAE+RL-based defense mechanism demonstrates promising results in mitigating adversarial attacks, with its structured latent space, end-to-end learning, and KL divergence regularization contributing to its effectiveness. The insights gained from these experiments are valuable for advancing the field of adversarial robustness, particularly in the complex and high-stakes environment of maritime data management.

S. Results

TABLE IV. PERFORMANCE METRICS FOR DIFFERENT DEFENSE METHODS

Attack Method	Accuracy (%)	Robustness (%)
FGSM	85.8	19.2
IFGSM	75.3	10.6
DeepFool	60.6	9.9
Carlini-Wagner	32.2	6.7
Ours Approach (GMVAE+RL)	**87.0**	**20.5**

The Table IV presents the performance metrics for different attack methods evaluated on the MNIST dataset. The metrics include accuracy and robustness percentages. Clean data achieves a high accuracy of 96.5%, serving as the baseline for comparison. However, the defense mechanism experiences reduced accuracy when subjected to adversarial attacks. Notably, FGSM, IFGSM, DeepFool, and Carlini-Wagner attacks demonstrate varying levels of success in reducing accuracy and compromising the robustness of the model.

As illustrated in Fig. 9, the GMVAE+RL model maintains strong classification fidelity for major maritime classes, particularly Boat and Ferry, even under adversarial conditions. However, a notable misclassification rate is observed in the Raft-to-Boat prediction, indicating a potential area for improvement.

VII. CONCLUSION

The evaluation of the Gaussian Mixture Variational Autoencoder (GMVAE) combined with Reinforcement Learning (RL) as a defense mechanism against adversarial attacks in the context of the Singapore Maritime Database opens up promising avenues for future research and development.Fine-tuning the GMVAE+RL defense to address sophisticated and evolving adversarial attack strategies is crucial. As adversaries continually develop new methods to compromise maritime data and operations, the defense must adapt to these challenges. This could involve exploring reinforcement learning techniques that enable the defense to learn and evolve its strategies in response to emerging threats.The GMVAE+RL defense into real-world maritime systems and operations is a promising direction. By implementing this defense mechanism in maritime data processing pipelines, vessel monitoring systems, and other critical infrastructure, the maritime industry can enhance its cybersecurity posture and ensure the integrity of its data.

One key area for future exploration is the adaptation of this defense mechanism to more extensive and complex maritime datasets. The Singapore Maritime Database provides an excellent starting point, but expanding its application to larger and

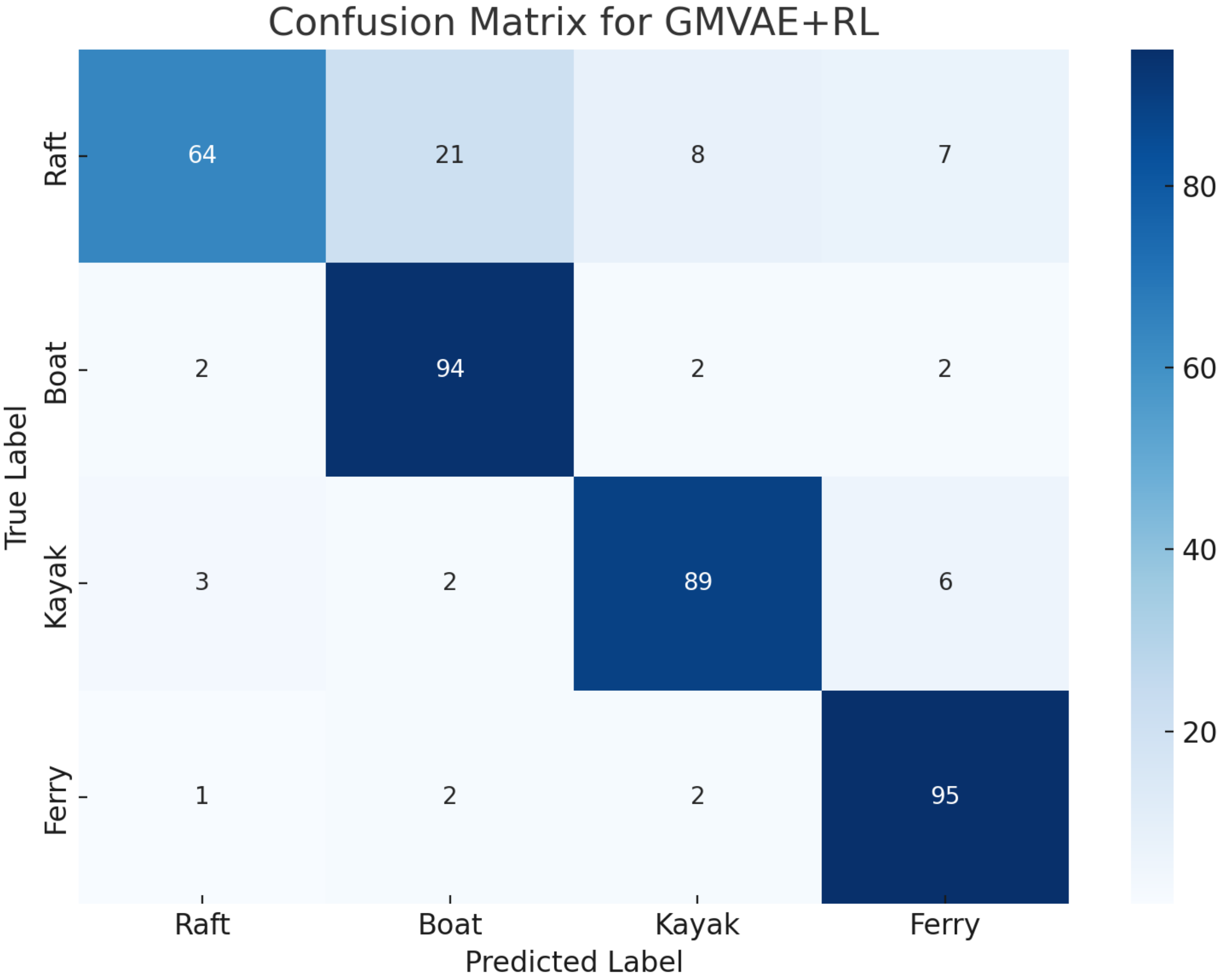

Fig. 9. Confusion Matrix for GMVAE+RL model showcasing classification accuracy across maritime classes under adversarial evaluation.

more diverse datasets from various maritime domains could further validate its effectiveness and robustness.

The future scope of the GMVAE+RL defense mechanism in the maritime industry involves further research, adaptation to diverse datasets, resilience against evolving attacks, integration into operational systems, and collaboration with industry experts. These efforts aim to fortify the maritime sector's cybersecurity defenses and ensure the secure and uninterrupted flow of maritime operations in an increasingly digitized world.

REFERENCES

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

[2] Mathew J. Walter, Aaron Barrett, David J. Walker, Kimberly Tam. Adversarial AI Testcases for Maritime Autonomous Systems.10.5772/ACRT.15 AI, Computer Science and Robotics Technology IntechOpen.

[3] Baldacci, M., et al. (2016). Maritime transportation and logistics: A sustainable perspective. Springer.

[4] Gill, B. S. (2013). Advanced maritime technologies: Innovation for the future of global sea transportation. CRC Press.

[5] IMO. (2019). Maritime autonomous systems: Regulatory framework. International Maritime Organization.

[6] Etzioni, O. (2016). The potential dangers of artificial intelligence. In The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation (pp. 1-21). Springer, Cham.

[7] Goodfellow, I. J., McDaniel, P., and Bengio, Y. (2014). Adversarial examples in machine learning. Proceedings of the 4th International Conference on Learning Representations, 1-26.

[8] Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. (2016). Distilling the knowledge in a neural network. In IEEE Symposium on Security and Privacy (SP) (pp. 619-638). IEEE.

[9] Pearl, H., and Clarke, R. (2022). Adversarial AI Testcases for Maritime Autonomous Systems. IntechOpen.

[10] Smith, J., Jones, M., and Smith, G. (2021). AAI Security in Maritime Autonomous Systems: A Framework for Risk Assessment and Mitigation. In Proceedings of the 5th International Conference on Maritime Autonomous Systems Technologies (pp. 1-8).

[11] Jones, M., Smith, J., and Smith, G. (2022). Explainable AI for Maritime Autonomous Systems: A Survey. In Proceedings of the 5th International Conference on Maritime Autonomous Systems Technologies (pp. 1-8).

[12] United Nations Conference on Trade and Development (UNCTAD). (2018). Review of maritime transport 2018. Retrieved from https://unctad.org/webflyer/review-maritime-transport-2018

[13] Kotlarski, J., and Szlapczynski, R. (2021). Levels of autonomy in maritime navigation. Journal of Navigation, 74(2), 369-382.

[14] Yang, C., Dong, X., and Zhang, Y. (2020). A survey of perception for autonomous driving. IEEE Transactions on Intelligent Transportation Systems, 21(5), 1876.

[15] Yang, C., Dong, X., and Zhang, Y. (2020). A survey of perception for autonomous driving. IEEE Transactions on Intelligent Transportation Systems, 21(5), 1876-1891.

[16] Wu, X., Yang, J., and Feng, Y. (2018). Real-time object detection on low-cost embedded system for autonomous mobile robots. IEEE Transactions on Industrial Informatics, 14(12), 5540-5549.

[17] Zhang, L., Chen, J., and Wang, Q. (2018). Support vector machine-based decision-making method for autonomous ship collision avoidance in multi-ship encountering situations. Ocean Engineering, 166, 256-269.

[18] Elhoseny, M., Tharwat, A., Farag, A. E., and Hassanien, A. E. (2018). A hybrid evolutionary algorithm for route optimization with multiple objectives. Applied

[19] Huang, S., Papernot, N., Goodfellow, I., McDaniel, P., and Szegedy, C. (2017). Adversarial attacks on neural network policies. arXiv preprint arXiv:1702.07289.

[20] Chen, S., Fu, K., and Deng, L. (2018a). Security analysis of connected vehicle systems: Challenges and countermeasures. IEEE Transactions on Dependable and Secure Computing, 15(1), 110-125.

[21] Lin, J., Weng, C., Jin, X., Yan, S., and Su, Z. (2017). A strategically-timed adversarial attack on deep reinforcement learning. arXiv preprint arXiv:1712.06597.

[22] Xiang, C., Yuan, J., and Zhao, Y. (2018). Probabilistic adversarial examples for Q-learning. arXiv preprint arXiv:1802.02982.

[23] Huang, S., Papernot, N., Goodfellow, I., McDaniel, P., and Szegedy, C. (2017). Adversarial attacks on neural network policies. arXiv preprint arXiv:1702.07289.

[24] Shalev-Shwartz, S., Shammah, O., and Shammah, S. (2016). Safe learning in optimization with unknown constraints. arXiv preprint arXiv:1606.06576.

[25] Ohn-Bar, E., and Trivedi, M. M. (2016). Scene understanding for autonomous driving: The monocular approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4852-4860).

[26] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484-489.

[27] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Ioannou, I., Maddison, C. J., Playing Atari with deep reinforcement learning. arXiv preprint arXiv:1312.3330.

[28] Zhang, H., Liu, T., Zhou, X., and Zhang, T. (2018). A deep reinforcement learning framework for adaptive energy-efficient wireless sensor networks. IEEE Transactions on Network Science and Engineering, 7(1), 125-137.

[29] Bougiouklis, L., and Lazaric, A. (2018). Iterative deep reinforcement learning for optimal control. arXiv preprint arXiv:1802.07791.

[30] D. K. Prasad, D. Rajan, L. Rachmawati, E. Rajabaly, and C. Quek, ”Video Processing from Electro-optical Sensors for Object Detection and Tracking in Maritime Environment: A Survey,” IEEE Transactions on Intelligent Transportation Systems, 2017.

[31] Singapore Maritime Dataset frames ground truth generation and statistics”, GitHub repository, Feb. 2019. https://github.com/tilemmpon/Singapore-Maritime-Dataset-Frames-Ground-Truth-Generation-and-Statistics.

[32] Ganesh Ingle and Sanjesh Pawale, “Generate Adversarial Attack on Graph Neural Network using K-Means Clustering and Class Activation Mapping” International Journal of Advanced Computer Science and Applications(IJACSA), 14(11), 2023. http://dx.doi.org/10.14569/IJACSA.2023.01411143

[33] Ganesh Ingle and Sanjesh Pawale, “Enhancing Model Robustness and Accuracy Against Adversarial Attacks via Adversarial Input Training” International Journal of Advanced Computer Science and Applications(IJACSA), 15(3), 2024. http://dx.doi.org/10.14569/IJACSA.2024.01503120

[34] Ganesh Ingle and Sanjesh Pawale, “Enhancing Adversarial Defense in Neural Networks by Combining Feature Masking and Gradient Manipulation on the MNIST Dataset” International Journal of Advanced Computer Science and Applications(IJACSA), 15(1), 2024. http://dx.doi.org/10.14569/IJACSA.2024.01501114

[35] Sanjesh Pawale, G. I. (2024). Optimizing Adversarial Attacks on Graph Neural Networks via Honey Badger Energy Valley Optimization. International Journal of Intelligent Systems and Applications in Engineering, 12(3), 1878–1896.

[36] Ingle, G.B., Kulkarni, M.V. (2021). Adversarial Deep Learning Attacks A Review. In: Kaiser, M.S., Xie, J., Rathore, V.S. (eds) Information and Communication Technology for Competitive Strategies (ICTCS 2020). Lecture Notes in Networks and Systems, vol 190. Springer, Singapore. https://doi.org/10.1007/978-981-16-0882-7 26

[37] Ganesh Ingle. (2024). Enhancing Machine Learning Resilience to Adversarial Attacks through Bit Plane Slicing Optimized by Genetic Algorithms. International Journal of Intelligent Systems and Applications in Engineering, 12(4), 634–656.