electronics

Article

Zero-Trust Marine Cyberdefense for IoT-Based Communications: An Explainable Approach

Ebuka Chinaechetam Nkoro [1], Judith Nkechinyere Njoku [1], Cosmas Ifeanyi Nwakanma [2], Jae-Min Lee [1] and Dong-Seong Kim [1,*]

[1] IT Convergence Engineering, Kumoh National Institute of Technology, Gumi 39177, Republic of Korea; nkorochinaechetam@gmail.com (E.C.N.); judithnjoku24@kumoh.ac.kr (J.N.N.); ljmpaul@kumoh.ac.kr (J.-M.L.)

[2] ICT-Convergence Research Center, Kumoh National Institute of Technology, Gumi 39177, Republic of Korea; cosmas.ifeanyi@kumoh.ac.kr

[*] Correspondence: dskim@kumoh.ac.kr

Abstract: Integrating Explainable Artificial Intelligence (XAI) into marine cyberdefense systems can address the lack of trustworthiness and low interpretability inherent in complex black-box Network Intrusion Detection Systems (NIDS) models. XAI has emerged as a pivotal focus in achieving a zero-trust cybersecurity strategy within marine communication networks. This article presents the development of a zero-trust NIDS framework designed to detect contemporary marine cyberattacks, utilizing two modern datasets (2023 Edge-IIoTset and 2023 CICIoT). The zero-trust NIDS model achieves an optimal Matthews Correlation Coefficient (MCC) score of 97.33% and an F1-score of 99% in a multi-class experiment. The XAI approach leverages visual and quantitative XAI methods, specifically SHapley Additive exPlanations (SHAP) and the Local Interpretable Model-agnostic Explanations (LIME) algorithms, to enhance explainability and interpretability. The research results indicate that current black-box NIDS models deployed for marine cyberdefense can be made more reliable and interpretable, thereby improving the overall cybersecurity posture of marine organizations.

Keywords: cybersecurity; zero-trust security; marine cybersecurity; Explainable Artificial Intelligence (XAI); SHAP; LIME; network intrusion detection; deep learning; IoT; IoUT; communications

check for updates

Citation: Nkoro, E.C.; Njoku, J.N.; Nwakanma, C.I.; Lee, J.-M.; Kim, D.-S. Zero-Trust Marine Cyberdefense for IoT-Based Communications: An Explainable Approach. *Electronics* **2024**, *13*, 276. https://doi.org/10.3390/electronics13020276

Academic Editor: Krzysztof Szczypiorski

Received: 15 November 2023
Revised: 28 December 2023
Accepted: 3 January 2024
Published: 8 January 2024

1. Introduction

Approximately 71% of the Earth's surface is subaqueous, significantly influencing the territorial, geographical, and economic landscapes of nation-states. In the current era of cyberwarfare [1], maritime organizations bear the crucial responsibility of establishing, managing, and securing marine networks to mitigate the risks of breaches and vulnerabilities. The frequency of cyber incidents in the maritime sector has witnessed a notable increase in recent years, exemplified by the 2020 ransomware attack on industry giants such as MAERSK. This attack resulted in substantial financial losses, estimated to be between 200 and 300 million USD. Additionally, intranet breaches targeting the International Maritime Organization (IMO) have raised security and reputational concerns [2]. The genesis of major marine cyberattacks often stems from vulnerabilities in Internet of Things (IoT) and Internet of Underwater Things (IoUT) sensors, which malicious actors exploit to initiate and perpetuate intrusions into maritime systems [3].

To strengthen marine cyberdefense systems, previous research has investigated the use of Artificial Intelligence (AI) frameworks to improve maritime Network Intrusion Detection Systems (NIDS) [4], thus guaranteeing faster and more reliable detection of cyberattacks such as Distributed Denial of Service (DDoS) attacks, ransomware, phishing, and backdoor attacks. Strong learning algorithms, such as deep neural networks, have been used to guarantee highly accurate predictions in marine NIDS, due to their ability to capture the spatial relations of IoT/IoUT network traffic data and detect malicious threats [5].

Major challenges within the introduction of AI in marine NIDS are outlined below:

(i) The prevalence of false alarm rates, fake distress calls, and especially the lack of explanations regarding the black-box AI algorithms used to predict marine cyberattacks [6]. Meanwhile, marine cyberdefense systems now require Explainable AI (XAI) frameworks and human-in-the-loop interactions for security experts to provide reliable and trustworthy predictions of marine cyberthreats.

(ii) Most XAI interpretation methods reported in the current literature focus majorly on visual explanations and still lack quantitative XAI metrics that can aid expert decisions or methods.

To overcome the above-highlighted challenges, visual, quantitative, and human-in-the-loop XAI can be employed to salvage the challenges of reliability and transparency in marine NIDS. A current cyberdefense paradigm, Zero-trust Architecture (ZTA), as proposed by the United States Department of Defence (DoD) in 2022, highlights a holistic approach that embodies real-time network traffic monitoring, strong authentication, and continuous evaluation of the confidence levels of AI-based NIDS models to address transparency and reliability in cybersecurity issues. ZTA adopts the "trust no one, verify everything" principle, thus providing NIDS experts with better understanding, reliability, and authentication of network users and mitigation of security threats just-in-time [7]. Within this strategy, explainable NIDS are layered to understand and prevent the stealthy advances of attackers whose aim is to tamper with the confidentiality, integrity, and availability of marine cyberspace.

For example, a marine cybersecurity expert might wish to investigate the following question: "How certain is this NIDS model's prediction of a normal or DDoS attack, and what training features led to the NIDS prediction?" To address the lack of transparency and model trustworthiness of most NIDS models, the growing area of XAI aims to address the major reasons for model distrust and provide security experts with insight-driven feedback for the improved security posture of their organizations [8]. Although recent works have begun studying XAI, only a few of them have addressed cybersecurity concerns related to marine cyberdefense. Other works have not provided quantitative and secure methods that help to differentiate malicious alerts and improve expert decisions and model trustworthiness [9].

Motivation—Previous research on cyber-resilience for marine networks using AI algorithms has not considered explainable methods that can provide transparency for predicting marine threats. By ignoring model explainability, a lack of better validation of classification results may be missing and thus can be linked to the growing rate of false alarms and attacks in marine networks.

Contribution—This paper's contributions are two-fold:

(I) We employ a hybrid neural network architecture that combines two popular types of neural network: a Convolutional Neural Network (CNN) and a Bidirectional Long Short-term Memory (BiLSTM) NIDS model with proper feature selection using the decision tree (filter) algorithm, capable of effectively detecting IoT marine cyberattacks.

(II) An exploration of the SHapley Additive exPlanations (SHAP) explainability method and the Local Interpretable Model-agnostic Explanations (LIME) methods are employed to yield more visual and quantitative insights towards the predictions of the marine NIDS model.

The rest of the paper is organized as follows: Section 2 provides a background study, Section 3.1 summarizes the relevance of the proposed system, and Section 3 deals with the methodology. The results of experimental findings are provided in Section 3, followed by conclusions in Section 5.

2. RelatedWorks

2.1. Cyberdefense in Marine Networks

The broad term “marine cyberdefense” represents the security of a broad range of marine sectors, including vessels, offshore and onshore facilities, navigation and transport systems, and cargo systems that rely on networked IoT and IoUT technologies that facilitate day-to-day marine operations [10,11]. Each marine system, depending on the type of application, is enabled with peripherals such as microphones, cameras, sound and image processing units, GPS units, and a collaborative communication mechanism where sensor nodes broadcast their data packets to neighboring nodes until data exchange is achieved. Due to the tremendous amount of data generated from the integration of these multiple marine systems, cybercriminals now leverage the vulnerabilities in IoT and IoUT communications to perpetuate marine cyberattacks [12]. Meanwhile, the surface attacks witnessed in marine organizations may vary specifically from regular cyber scenarios in terms of the underlying network infrastructure compromising of complexities in marine supply chain systems, marine GPS vulnerabilities, and even marine communication jamming attacks, which ordinary businesses do not usually witness [13].

Severe cyberattacks have been reported by shipping industries and the IMO [14] concerning the prevalence of attackers (hacktivists, terrorists, digital pirates, and ransomware groups) who disrupt marine networks in the form of DDoS attacks or steal confidential information for financial gain [15]. As shown in Figures 1 and 2, there have been several cyberattacks detected by leading marine cybersecurity experts in 2023 alone, showing a continual increase in cyber incidents within marine environments (shipping, supply chain, energy, yard, port, defense, marine organizations, and vessel operations) [16]. These proliferated attacks can be linked to the fast-paced stealthiness of modern attackers, vulnerabilities of IoT and IoUT technologies, and most especially, lack of real-time defense mechanisms such as NIDS [2].

As shown in Figure 3, marine networks employ automated IoT and IoUT systems that foster ship-to-ship communication, which optimizes marine productivity while reducing operational costs. Marine network communication is characterized by interoperable nodes such as base stations, coastal units, and the Software-Defined Network (SDN) controller. The network control center transmits and receives several different wireless technologies, such as Long-term Evolution Advanced (LTE-A), Wireless Fidelity (Wi-Fi) networks, satellites, and acoustic communications (buoys) [17,18]. All marine communication nodes are geographically distributed among the different regions, including coastal, offshore, open-sea, or underwater communication endpoints to facilitate continuous communications and the running of the marine industry.

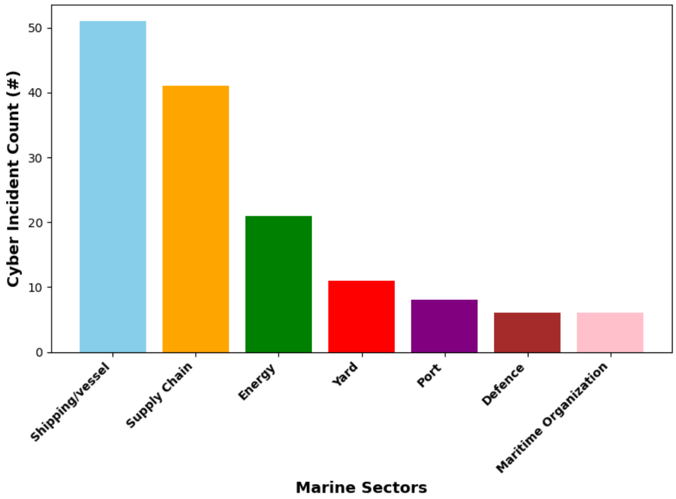


Figure 1. Cyberattack incidents within various marine sectors in 2023 where # signifies the cyber incident count.

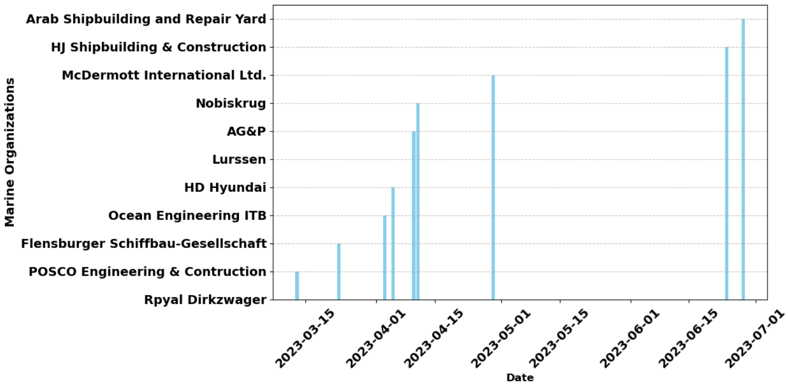


Figure 2. Timeline of cyber incidents in yard marine industry 2023.

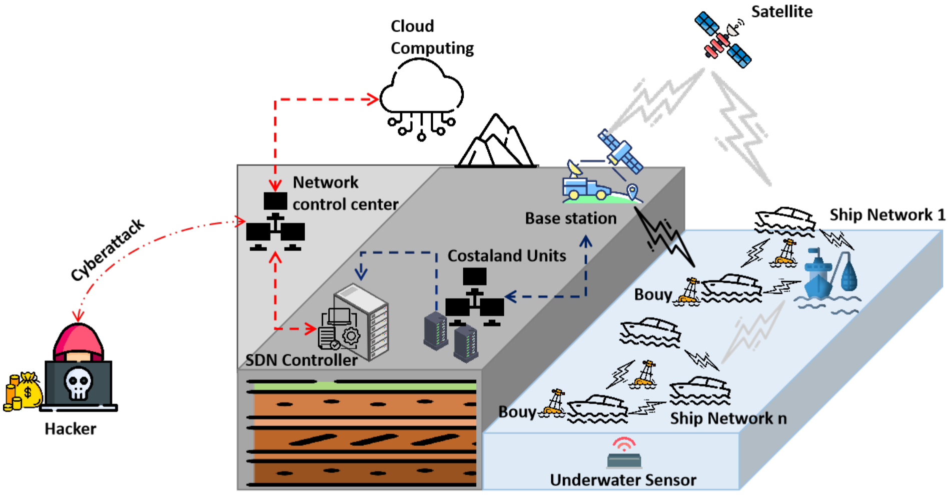


Figure 3. Illustration of marine network communications and cyberthreats.

However, stealthy hacker groups can leverage network system vulnerabilities to perpetuate cyber incidents such as vessel engine failures, manipulation of control systems, and jamming attacks [19]. Recovering from cyberattacks has become very expensive. In January 2023 a hacker group 'PLAY' published the private data of four European Union marine IT companies on the darkweb after December 2022's ransomware infection. Within the same year, another recognized advisor for the maritime industry, Det Norske Veritas (a leading classification society and a recognized advisor for the maritime industry), was reported as a fresh target, where hackers compromised the data of ship management companies, which account for 21% of the total share of the marine industry [20].

2.2. NIDS for Cyberdefense in Marine Networks

To overcome marine cyber incidents in a very responsive manner, the use of network intrusion detection systems has been employed by security organizations and in the domain of marine security to detect and automate potential attacks, thus preserving marine security postures and organizational reputation. NIDS methods, as explored by previous works [21], can be categorized as follows: signature-based, anomaly-based, flow-based, and machine-learning-enabled NIDS aimed at detecting anomalous network traffic while minimizing the number of false-positive predictions.

Traditional machine-learning algorithms have been explored in the field of IoT-enabled NIDS for secure marine network operations. In a study on IoT botnet attack detection, Alqahtani, Mathkour, and Ismail (2020) proposed a method based on the optimized extreme Light Gradient Boosting (LGB) algorithm for detecting and protecting IoT devices from dangerous large-scale botnet attacks. Our previous work [22] also utilized the LGB due to its fast computation, which is vital for fast and cost-efficient computation in IoT networks. The results of the multi-class results yielded a 95% accuracy while predicting 12 diverse cyberattack types using the 2023 EdgeIIoT dataset.

Unlike traditional machine-learning classifiers, which are limited in their ability to extract features from massive data, considering the extensive cyber traffic in real life, the use of deep-learning algorithms for network traffic classification has been preferred in modern research [5].

This work is an extension of our previous approach presented in [22]. Earlier work addressed the gap in explainable NIDS models within the domain of marine cyberdefense. Furthermore, this study was supplemented with an additional dataset, well-investigated feature selection methods, visual XAI, and a significant quantitative XAI interpretation of the proposed "black-box" neural network model. In comparison with the tree-based algorithm used for the classification task in our prior work [22], neural network NIDS models are inherently not easily interpretable [8].

Meanwhile, Hou et al. in [23], proposed an intrusion detection framework for hydrographic station network anomalies. The proposed approach utilized a hybrid CNN and BiLSTM method using the NSL KDD dataset while obtaining an F1-score of 87.35%. Although their approach was effective in identifying deep features, the low accuracy of their results cannot be ignored, taking into consideration efficiency requirements and the need for the low false-positive rates required for a zero-trust model in marine networks.

Xin et al. in [24], proposed a Generative Adversarial Network (GAN) approach to process the imbalanced NSL-KDD dataset [25] for IDS in marine networks. Within their method, a data generation module was initiated to improve minority class samples, using the OPTICS denoising algorithm. The classification accuracy of the authors' proposed data augmentation method yielded a micro-average accuracy of 95% with five classes of network traffic. A decentralized training method using a federated learning approach for marine IDS was investigated by authors in [4]. Their federated learning technique was designed to save computing and storage overhead, with an accuracy of 87%, 500 rounds of training, and the use of the old NSL-KDD dataset. Dataset dimensionality in the domain of NIDS availability, suitability, and dimensionality in the domain of NIDS has become a bottleneck for the efficient and effective correlation of network traffic for improved model accuracy. Therefore, the use of obsolete datasets such as NSL-KDD may not fit the current demands of modern networks.

2.3. Zero-Trust Cyberdefense in IoT

The zero-trust security architecture, as recently published by the National Institute of Standards and Technology (NIST), is a paradigm shift towards rethinking the network security and protection of organizational assets. The strength of ZTA principles in IoT and marine cyberdefense lies in its skepticism [26], i.e., "assume breach, verify explicitly, privilege access only", and not blind trust, thus supporting multi-level authorization/scrutiny to achieve fine-grained security controls. ZTA embraces five core tenets, as shown in Figure 4, namely:

i Resource segmentation;
ii Ubiquitous authentication;
iii Strong encryption;
iv Principle of least privilege;
v Intelligent real-time threat monitoring.

The US DoD released a ZTA framework for integrated threat intelligence and remediation [27]. Therein, machine-learning analytics, real-time network traffic monitoring, and orchestration capabilities were employed to enforce the DoD's data/enterprise security against cyberthreats. In addition, evaluating the confidence levels of the DoD's ML models, devices, users, and resources is routinely performed to ensure minimal security vulnerabilities. Recent advancements towards the Industry 5.0 paradigm now require zero-trust network-based access using AI for effective cybersecurity and real-time monitoring, controlling, and allocations of production sequences to prevent false rates and maximize productivity [28].

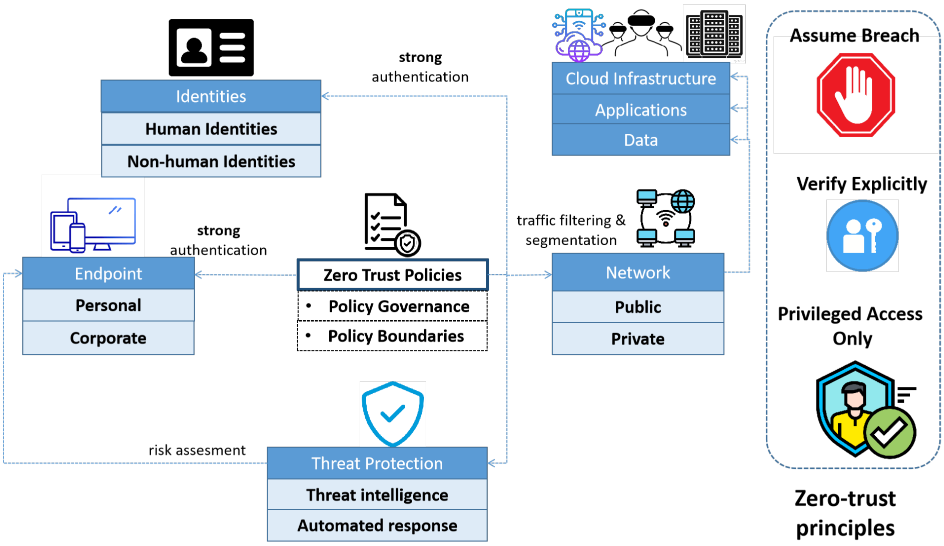


Figure 4. A visual illustration of the zero-trust cybersecurity principles, governed by strong authentication, filtering, threat intelligence, and zero-trust policies.

Current academic research has also investigated the employment of the ZTA for strong IoT security. Recent studies in [29] have proposed that the ZTA model will address most of the security concerns in 5G networks, where security models can dynamically detect/identify the malicious activities of users, devices, or applications. Proof-of-concept experiments using blockchain have also been employed to satisfy security requirements in edge computing networks [30]. The simulation results showed that ZTA in edge networks can satisfy successful edge node authentication with good time constraints. A recent trend towards adopting ZTA NIDS using deep learning has also yielded the increased security of network devices by calculating the security scores/awareness of imminent security threats [31], thus satisfying the ZTA demand for real-time monitoring and mitigation against threats. Syed et al. [32] presented a comprehensive survey, highlighting the relevance of AI-based NIDS for full ZTA realization in IoT-based networks. The survey, like previous ZTA policies in [27], outline the need to evenly distribute the zero-trust principle within machine–machine communications, IoT devices, security protocols, and even AI models. The authors in [33] have also emphasized how AI-based NIDS methods can be employed for resilient zero-trust IoT defense. Here, AI models can establish a probabilistic relation between a Cyber–Physical System (CPS) hypothesis (i.e., likelihood of attacks) and the supporting evidence (i.e., signs of attack activities); thus, even the slightest malicious activities can still be detected in real-time, and with high confidence, as long as enough evidence is accumulated. However, the development of AI algorithms requires scrutiny and interpretability to ensure that they make predictions as required.

2.4. *XAI for Cyberdefense*

The development of trustworthy, transparent, and reliable algorithms has gained tremendous momentum in modern AI development. Recently, in October 2023, President Biden issued an executive order on safe, secure, and trustworthy AI, which requires that developers provide interpretable and trustworthy models during training to satisfy the safety/security of AI-enabled CPSs, software, and networks [34]. To address the potential risks (bais, transparency, privacy) and challenges associated with the widespread adoption of AI in IoT, there has been a huge interest in the field of XAI models by cybersecurity experts and researchers in IoT-enabled CPSs. In the domain of NIDS, for example, diverse questions such as “Why should we implicitly trust the predictions of NIDS?” [9] and “How certain is this model's prediction of a cyberattack?". These are the basic answers that an explainable NIDS seeks to address.

Within the domain of XAI, two categories of XAI periods (time of model explanation) post-hoc and ad-hoc, have been adopted for interpretable NIDS [8]. As illustrated in Figure 5, the current XAI taxonomy can be classified within time, complexity, and scope requirements. Concerning timely requirements, the ad-hoc explainability method provides model explanations during its decision process, while the post-hoc methods offer explainability information after model prediction in terms of intrinsic explanations, such as feature contributions to the model output. Commonly used post-hoc explainability methods in NIDS are SHAP and LIME. The SHAP and LIME explainers have become a favorite choice for explainable NIDS due to their model-agnostic features, while providing explanations to deep-learning-based NIDS models, which are fairly opaque in their decision-making process. Meanwhile, the complexity of interpretation methods can be classified as either extrinsic (model-agnostic) or intrinsic (model-specific). For example, tree-based (rule-based) NIDS models are inherently interpretable depending on the dataset or complex nature of training. As regards the scope of interpretations, XAI techniques can be categorized as either global or local [8].

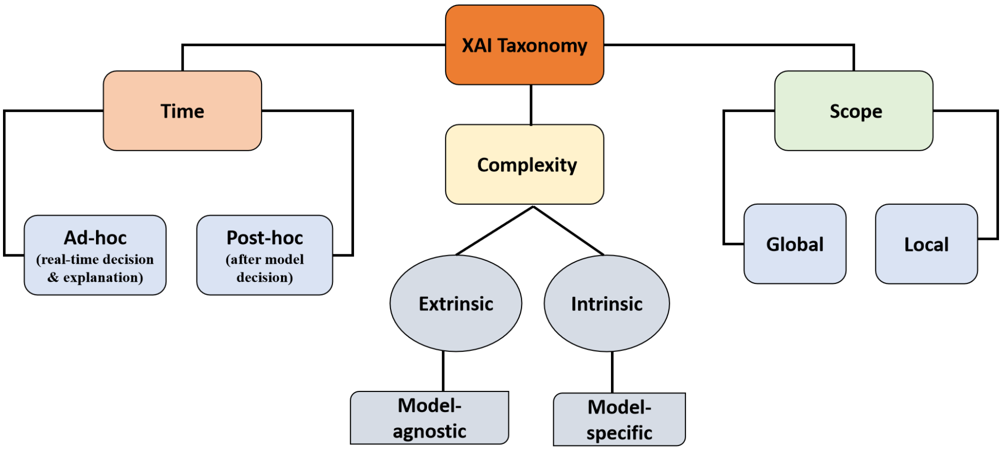


Figure 5. A summarized visual taxonomy of XAI methods.

Within current XAI frameworks for NIDS, the SHAP and LIME XAI methods possess specific suitability and efficiency with network traffic data, and they are an alternative to computer vision-based XAI methods such as Generalization of the Class Activation mapping (Grad-CAM), Guided Grad-CAM, and axiom-based Grad-CAM, which may be computationally expensive while converting network traffic (text) to images, given the real-time demands for security, efficiency, and interpretability. The SHAP explainer provides the marginal value of contributions made by a feature or subset of features within a model's prediction. Similarly, the LIME explainer can generate local surrogate models to approximate the decision-making process of a complex model, providing interpretable explanations for individual predictions by highlighting important model features [35].

Current research into various cyberattacks, such as phishing attacks, botnets, and fraud, is gaining better insights, proper visualizations, and deeper forensics into the nature of these attacks. Additionally, significant features for model training can be identified to perform effective/trustworthy cyberdefense [8]. As depicted in Figure 6, NIDS approaches in IoT CPSs can model explainability and trustworthiness to evaluate the credibility of the predicted cyberattacks. Other layers of the zero-trust model cover areas such as physical barriers or mechanisms, which can help marine organizations to prevent, monitor, or detect unauthorized access to their assets through the use of locks, rails, CCTVs, badges, PC locks, turnstiles, and alarms. Important perimeter defenses includes identity access, perimeter security, compute, application security, and data integrity.

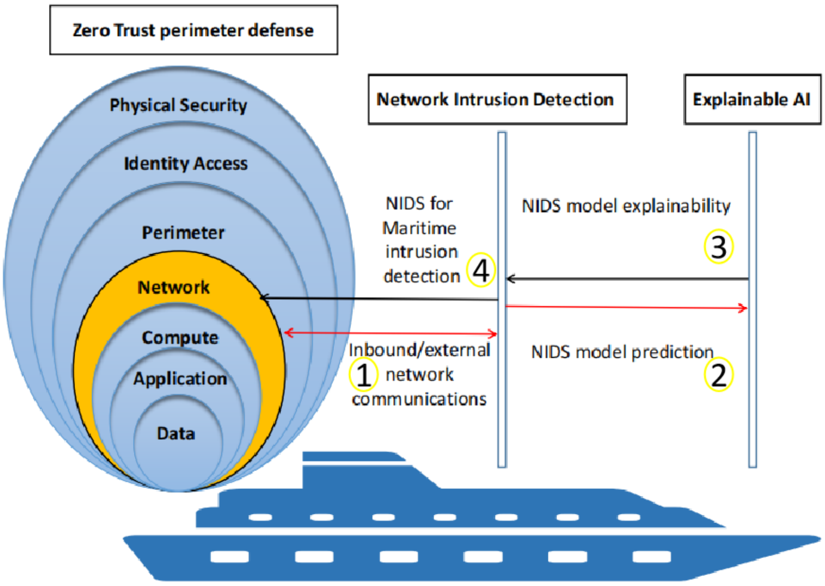

Zero Trust perimeter defense
Physical Security
Identity Access
Perimeter
Network
Compute
Application
Data
Network Intrusion Detection
Explainable AI
NIDS for Maritime intrusion detection
4
NIDS model explainability
3
Inbound/external network communications
1
NIDS model prediction
2

Figure 6. Sequence diagram of the zero-trust perimeter defense strategy for marine networks, with insight-driven feedback using an Explainable AI (XAI) Network Intrusion Detection (NIDS) model.

A malware detection method in [36] investigated an explainable malware detection method to understand the "outstanding" performance of their proposed model in real-world environments. One drawback in their approach was a lack of visual interpretations or XAI methods, for better interpretability. A lack of visual explainability of the XAI frameworks hinders the understanding, debugging, and decision-making within the proposed system. To address the lack of XAI in NIDS, [37] presented an explainable ANN DL model using the CICIDS 2017 dataset [38]. The authors utilized the oracle module, which also showed limited explainability of the model results.

To provide additional insights and forensics into NIDS models, Shruti et al. [39] only explored the visual LIME explainability using naturally transparent machine-learning algorithms, such as decision trees, random forests, and SVM. Another approach [40] using the SHAP explainer investigated a deep-learning approach named the trustworthy explainable artificial intelligence and enhanced krill herd optimisation intrusion detection system to detect breaches in IoT-enabled CPSs. Using the NSL KDD dataset [41] and the CICIDS 2018 dataset [38], their explanations using SHAP yielded insights towards the significant impact of training features in terms of the proposed NIDS model classification accuracy. Mohammed [42] proposed packet-based efficient and explainable IoT botnet detection using machine learning. The SHAP discussions using the Shapley additive explanation also provided transparency to the classifier's prediction process.

Zakaria et al. [9] designed a deep neural network XAI-based framework using the SHAP, LIME, and RuleFit XAI methods to explain their proposed NIDS framework, which was aimed at detecting IoT-related intrusions. However, the proposed system included a non-informative and redundant network traffic feature—source IP address ('srcip')—which is only meaningful within the dataset explored (NSL KDD). The 'srcip' would, by default, gain a high weight, thus dominating the SHAP plot and the model's predictions. This dominance led to a false conclusion that the source IP address is the most critical feature for the proposed NIDS, which may not be the case in a general scenario. The proposed approach does not generalize to the robust defense, interpretation, and transparency of NIDS models. A spoofing detection method in [43] provided both the LIME and SHAP explainability results of cyberattacks in IoT networks. Their work still lacks state-of-the-art evaluation metrics and a better discussion of the explainability results in terms of quantitative decisions and confidence in model predictions. It is therefore expedient, as required by modern defenses in the domain of NIDS, to provide XAI interpretations to bolster the trust and reliability of NIDS prediction.

3. Methodology

As shown in Figure 7, the proposed cross-silo marine NIDS framework comprises a deep-learning-based NIDS model and XAI interpretations for classifying malicious traffic within marine CPSs. First, marine network traffic from IoT/IoUT nodes, devices, data, and application planes consisting of anomalous and normal traffic managed by the Software-Defined Network (SDN) controller is parsed into a CSV format using traffic-capturing tools. Next, the statistical features of the captured network traffic are extracted using feature-extraction tools such as the CiCFlowmeter.

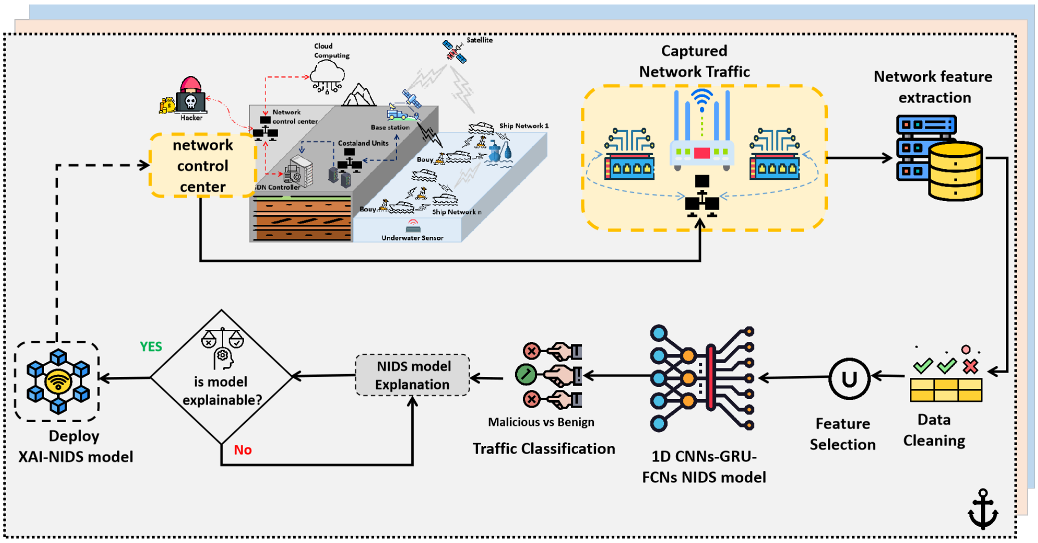


Figure 7. Proposed workflow of the zero-trust marine NIDS with insight-driven feedback using an Explainable AI.

3.1. Relevance of Proposed System

This work builds from the existing body of knowledge that has laid strong emphasis on the need for zero-trust cyberdefense in IoT-enabled marine communications. We design a ZTA policy where marine security experts can employ proactive defense for marine cyberattack detection. To start with, a proper feature selection method is employed for model training and enhanced performance. Next, this work proposes a deep-learning-based NIDS model that can detect network traffic patterns more effectively. Essentially, a combination of quantitative and qualitative XAI interpretation of the NIDS model is performed, enabling marine network experts to handle, visualize, and mitigate cyberthreats more effectively. The overall cross-silo approach is conveyed in Figure 7.

3.2. Dataset Selection

This work employs the most recent cybersecurity datasets for the proposed NIDS marine cyberdefense. The 202 EdgeIIoT dataset, published by Amine et al. [44], which includes diverse forms of cyberattacks, such as DoS/DDoS attacks, information gathering, man-in-the-middle attacks, injection attacks, and other malware attacks (15 classes), as shown in Figure 8, along with a 7-layer test-bed for ML-based NIDS tasks, is employed for our experiments. It comprises 64 features generated from a test-bed of the cloud computing layer, network functions virtualization layer, blockchain network layer, fog computing layer, software-defined networking layer, edge computing layer, and IoT and IIoT perception layer, which satisfies the critical requirements of IoT communications.

3.3. Investigating Class In-Balance

Representative network traffic datasets for standard modern IoT network communications datasets come inherently unbalanced, with more significant volumes of benign traffic, since it is unlikely that IoT networks are persistently under attack. Balancing network traffic data before training may pose the risk of oversimplifying the complexity of actual network dynamics and cause sub-optimal model generalization. However, we equally investigate an oversampling technique using the EdgeIIoT dataset [45] to evaluate more realistic/generalized model performance with oversampling or without. This approach is deemed suitable for our proposed ZTA framework and is attainable in real-world scenarios. As depicted in Figure 9, this study addresses the issue of class imbalance in the EdgeIIoT dataset by using the Synthetic Minority Over-sampling Technique (SMOTE) to oversample the minority class in the training set, creating a balanced dataset for subsequent training.

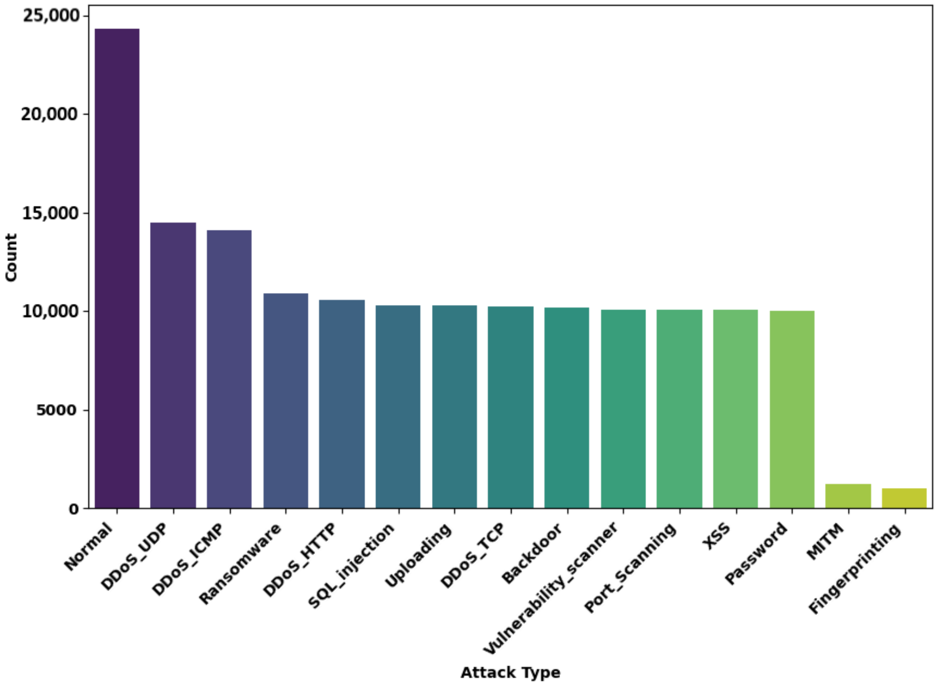


Figure 8. Class distribution of the Edge IIoT 2023 dataset employed for experimental analysis.

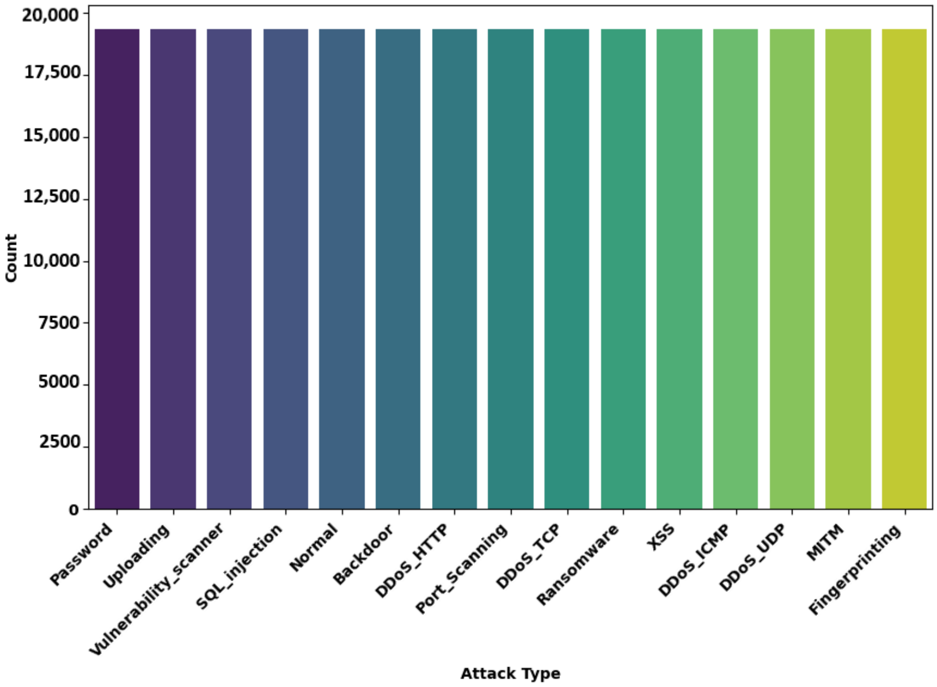


Figure 9. Class distribution of the Edge IIoT 2023 dataset after oversampling using SMOTE.

For further experimental evaluation, the significance of the proposed zero-trust NIDS model is evaluated with the CICIoT2023 dataset [46], as shown in Figure 10. The CICIoT2023 dataset was recently developed in a more extensive network of over 4 million entries and 33 attack types and over 105 IoT devices, including DDoS attacks, network traffic features of legacy industrial cyberattacks related to unauthorized commands, and denial of service attacks widely witnessed in critical IoT infrastructures. The two recent datasets, comprising the most closely representative real-world network traffic deliverables, offer valuable resources for advancing experimentation in the domain of NIDS.

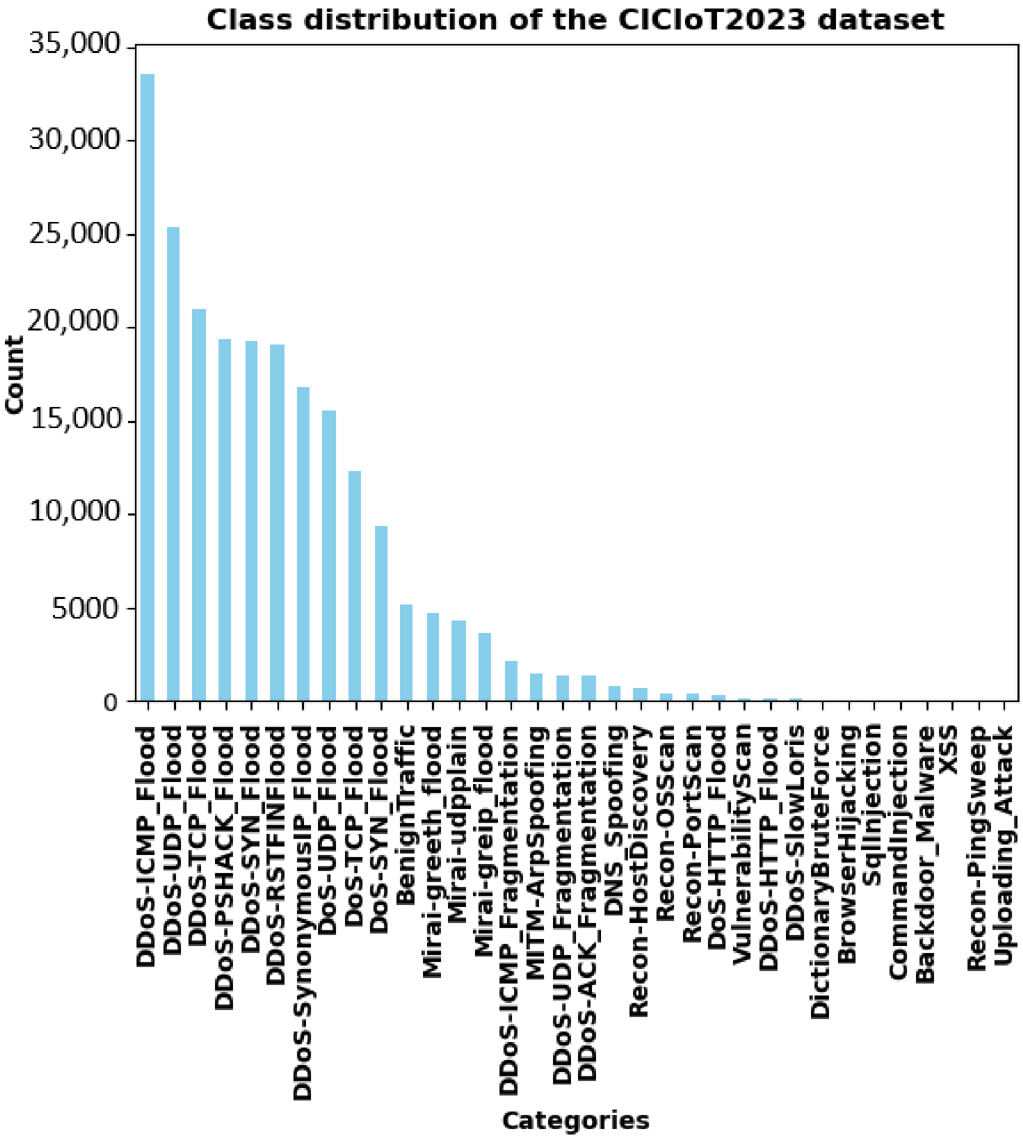


Figure 10. Class distribution of the CICIoT 2023 datasets employed for experimental analysis.

3.4. Feature Selection

To satisfy the requirements of an efficient NIDS model, the Feature Selection (FS) method, comprising viable features from the Pearson Correlation Coefficient (PCC) and Decision Tree (DT) filter algorithms, is adopted. The Filter FS method, in the domain of network traffic classification, outperforms other methods such as wrapper and embedded due to its computational efficiency and accuracy yielding [47]. Employing each filter FS method, i.e., the PCC and DT in this work, exhibits a distinctive approach and classification performance in terms of accuracy. By carefully measuring the natural cutoff points in the feature distributions, 0.8 is chosen as an appropriate threshold to select the top 30 traffic features of the DT and PCC methods using the Scikit learn Min–Max Scaling function.

The PCC represented in Equation (1) calculates the congruence between network traffic features, thereby removing the "uncorrelated" features within the set variance threshold of K-highest scores (0 to 1), where 0 denotes no correlation and 1 is a positive linear correlation.

$$r = \frac{N\sum XY - (\sum X \sum Y)}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}}, \tag{1}$$

The DT feature importance is determined by using a decision tree classifier to reduce the dimensionality of the input data and equalize the performance of ML models by removing redundant features. The "entropy" criterion is used for splitting the tree until it is pure (having only one class), ranking top features based on their information gain. Equation (2) shows the Entropy criterion H used for impurity measurement.

$$H(Q_m) = -\sum_k (p_{mk}) log(p_{mk}). \tag{2}$$

where m is the terminal node, Q_m represents the data at node m, and P_{mk} shows the arrays of predicted probabilities for each class where the sum of probabilities for all classes is 1.0. The feature importances is computed to select the top 30 traffic features in a ranking order. The PCC omits very important traffic features such as the flow duration, while the DT includes important features dropped by the PCC. Evaluating each FS (PCC and DT) method helps a security expert to select and measure the impact of viable traffic features when performing a multi-class or binary classification task. Adopting only one FS method may result in potential consequences, such as low detection accuracy, especially in different types of network attacks that rely heavily on flow-based features [48].

3.5. Proposed Hybrid NIDS Model

This study employs state-of-the-art DL algorithms to tackle the proposed marine NIDS classification task. This choice of DL method is predicated on its exceptional ability to capture intricate traffic features within the NIDS domain, as substantiated in previous studies [5]. The models under consideration encompass a foundational deep neural network (DNN) with four layers, a three-layered recurrent neural network (RNN), a 1D Convolutional Neural Network, and Gated Recurrent Unit, followed by Fully Connected Networks (1D CNNs-GRU-FCNs), CNN-LSTM, and CNN-BiLSTM models. These models were employed in both the training and evaluation phases. Each model is trained on IoT network traffic data that have been preprocessed, standardized, and encoded using one-hot encoding.

The best performing NIDS model (CNN-BiLSTM) in terms of more accurate predictions, as shown in Figure 11, distinctly detects diverse malicious from benign IoT traffic. The NIDS model consists of a sequential neural network designed for classification tasks. It begins with two 1D convolutional layers (64 filters, ReLU activation) for feature extraction. The core element is a BiLSTM layer with 64 units, using 'concat' merge mode for sequential pattern capturing. After flattening, there are two dense layers with ReLU activation, followed by a softmax classification layer. This architecture effectively extracts and classifies patterns in the input network traffic data. The model was trained using an Adam optimizer (0.001) for 20 epochs to minimize the loss function while optimizing the weights and biases.

Equation (3) represents the NIDS model:

$$\hat{y} = \sigma(w_l^T \cdot X_i + b_l). \tag{3}$$

where a data sample, X_i, is passed to the lth layer, given as $X_i \in R^{1\times d}$, and d signifies the number of features for the model prediction.

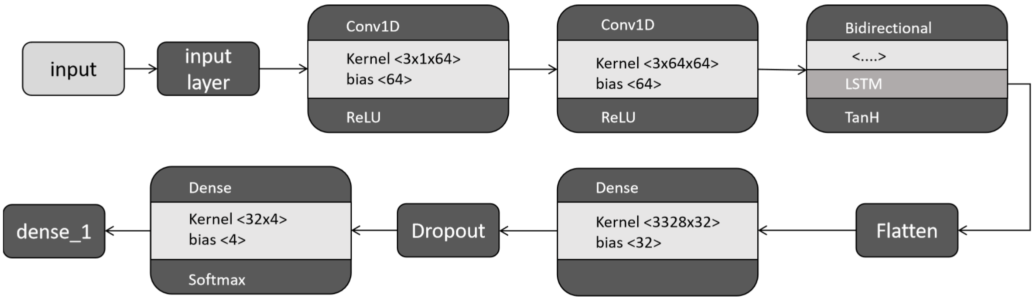


Figure 11. Illustrates the CNN BiLSTM NIDS model for marine IoT traffic classification.

4. Result Discussion

4.1. Model Performance and Evaluation Metrics

The performance of the NIDS model is evaluated adequately to the degree of correctness and model cost savings. The evaluation metrics within the experiment include the Mathew correlation coefficient (MCC) score [49], a weighted F1-Score, accuracy, test-loss value, and training time (computational resource requirements).

The MCC score is represented in Equation (4) and serves as an instinctive performance evaluation metric for the classification task, which highlights the performance of base classifiers on balanced or imbalanced datasets and is not as misleading as the normal F1-Score or the accuracy, which fails to consider the ratio between positive and negative elements. Accuracy, on the other hand, is the simple mean of model correctness obtained from the difference in predictions from the labeled ground-truth data.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}}, \tag{4}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \tag{5}$$

$$Recall = \frac{TP}{TP+FN}, \tag{6}$$

$$Precision = \frac{TP}{TP+FP}, \tag{7}$$

F1-Score is denoted as

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{8}$$

Implementation Environment: The IDS framework was carried out using the scikit learn library, which provides tree-based APIs. The experiment was run on a Jupyter Notebook with an Intel(R) Core(TM) i3-7100 CPU @ 3.90 GHz, and RAM 8.00 GB.

4.2. Discussion and Analysis Conducted with the EdgeIIoT Dataset

4.2.1. Impact of FS Methods and Oversampling Using SMOTE with the CNN-BiLSTM NIDS Model

As shown in Table 1, the quality FS of network traffic features yields an improved detection accuracy. A comparison of the filter FS method with the CNN-BiLSTM model is evaluated and demonstrates the performance of 30 features of the individual PCC and DT-FS algorithms. The DT technique shows greater improvement than the PCC FS method in more accurately predicting the 15 classes of cyberattacks, with an MCC score of 78.96, minimal loss of 0.4796, and an F1-score of 80%, while yielding an accuracy, precision, and recall of 80%, 80%, and 83%, respectively.

Table 1. Evaluation of the PCC and DT-FS methods with oversampling, using the CNN-BiLSTM NIDS model.

Feature Selection	No. of Classes	MCC %	Loss %	F1-Score %	Accuracy %	Precision %	Recall %	Train Time (s)
PCC	15	76.26	0.5535	77	78	80	78	1477
DT	15	78.96	0.4796	80	80	83	81	1834
DT with SMOTE	15	79.29	0.4707	80	81	83	81	1509

The result of the balanced dataset using SMOTE is equally evaluated in Table 1 using the better performing DT-FS method. Comparatively, the analysis of model performance between the balanced dataset using SMOTE yields very marginal MCC, loss, F1 score, accuracy, precision, and recall values. The decision not to leverage SMOTE in the proposed ZTA marine framework is justified by existing NIDS literature in [50,51] and the acknowledgment of the intrinsic imbalances within network traffic data, inherently reflective of authentic network dynamics, which fosters a trustworthy and adaptable ML-enabled NIDS solution.

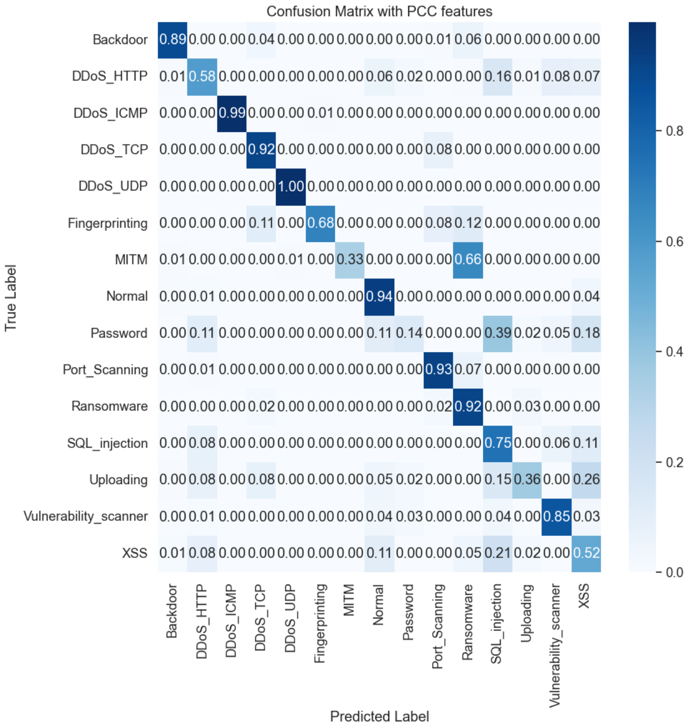
4.2.2. CNN-BiLSTM NIDS Model with 15 Classes

A comparative analysis of the confusion matrix in Figures 12 and 13 summarizes the performance of the DT-FS method while predicting 15 classes of cyberattacks (including benign traffic). As shown, the DT-FS technique can identify more classes correctly than the PCC method. The backdoor, DDoSHTTP, DDoSICMP, DDoSTCP, DDoSUDP, fingerprinting, MITM, Normal, password, port scanning, ransomware, SQL injection, Uploading, vulnerability scanner, and XSS classes are well identified, with a good performance of 0.91%, 0.62%, 0.99%, 1.0%, 0.69%, 0.34%, 1.0%, 0.22%, 0.96%, 0.88%, 0.66%, 0.41%, 0.90%, and 0.72%, respectively. There still exists some poorly identified classes in the multi-class results. In general, most machine-learning algorithms assume data is equally distributed; therefore, in the presence of a class imbalance the classifier tends to be more biased towards the majority class, which is the reason for the poor classification of the minority class.

4.2.3. CNN-BiLSTM NIDS Model with 4 Classes

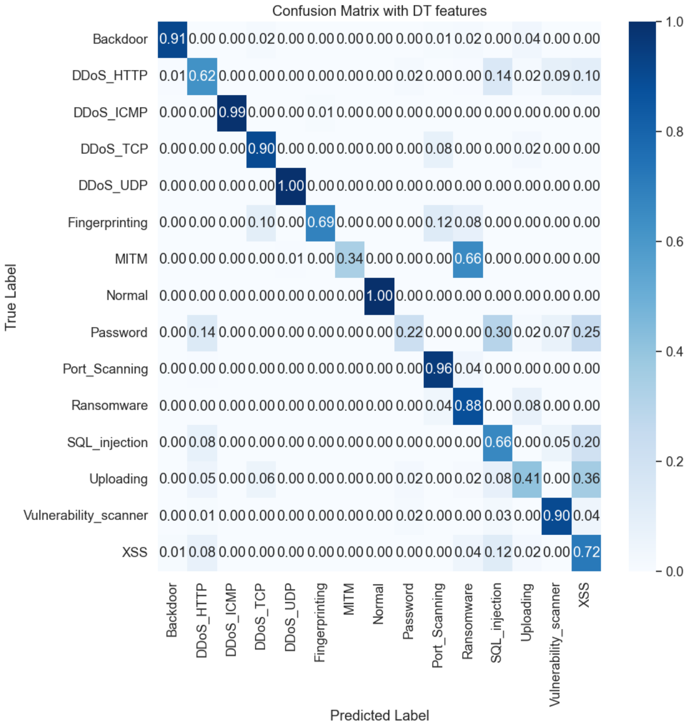
The EdgeIIoT dataset has been categorized into four main classes to address challenges associated with the minority of the 15 classes. Notably, attacks such as Backdoor, password, MITM, ransomware, Uploading, XSS (cross-site scripting), and SQL injection are grouped under the umbrella of malware. Meanwhile, DDoSHTTP, DDoSICMP, DDoSTCP, and DDoSUDP are collectively classified as DDoS attacks. The remaining traffic types, including fingerprinting, enumeration, port scanning, and vulnerability scanner, are considered as enumeration attacks. The normal class remains standalone, resulting in the four primary categories: malware, DDoS, enumeration, and normal.

Table 2 presents the experimental results for the EdgeIIoT dataset with these four classes. The CNN-BiLSTM model, employing the DT-FS method, outperforms the other four models in terms of a higher MCC score and minimal loss. Achieving an impressive MCC score of 88.38% and a minimal loss of 0.1991, the CNN-BiLSTM model demonstrates its superiority in identifying all four classes of the predicted network traffic. The weighted F1-score, accuracy, precision, and recall all stand at a consistent 92%. The yielded F1-score, in the context of marine NIDS, highlights the mode's ability to detect network anomalies while maintaining a balance between false-positive and false-negative intrusion detection. This is particularly important in ZTA marine cyberdefense, where the consequences of missing a genuine threat or triggering false alarms could have significant implications for maritime security and operations. In addition, the high precision results in Table 2 indicate that the proposed marine cyberdefense model truly flags marine incidents, while minimizing the number of false positives that could lead to unnecessary disruptions or investigations. Additionally, the NIDS model recall, or sensitivity, reflects the NIDS model's ability to capture and correctly identify all instances of malicious or anomalous behavior within the maritime network.



	Backdoor	DDoS_HTTP	DDoS_ICMP	DDoS_TCP	DDoS_UDP	Fingerprinting	MITM	Normal	Password	Port_Scanning	Ransomware	SQL_injection	Uploading	Vulnerability_scanner	XSS
Backdoor	0.89	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.00	0.00	0.00	0.00
DDoS_HTTP	0.01	0.58	0.00	0.00	0.00	0.00	0.00	0.06	0.02	0.00	0.00	0.16	0.01	0.08	0.07
DDoS_ICMP	0.00	0.00	0.99	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DDoS_TCP	0.00	0.00	0.00	0.92	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00
DDoS_UDP	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Fingerprinting	0.00	0.00	0.00	0.11	0.00	0.68	0.00	0.00	0.00	0.08	0.12	0.00	0.00	0.00	0.00
MITM	0.01	0.00	0.00	0.00	0.01	0.00	0.33	0.00	0.00	0.00	0.66	0.00	0.00	0.00	0.00
Normal	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.04
Password	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.11	0.14	0.00	0.00	0.39	0.02	0.05	0.18
Port_Scanning	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.93	0.07	0.00	0.00	0.00	0.00
Ransomware	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.92	0.00	0.03	0.00	0.00
SQL_injection	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.75	0.00	0.06	0.11
Uploading	0.00	0.08	0.00	0.08	0.00	0.00	0.00	0.05	0.02	0.00	0.00	0.15	0.36	0.00	0.26
Vulnerability_scanner	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.04	0.03	0.00	0.00	0.04	0.00	0.85	0.03
XSS	0.01	0.08	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.05	0.21	0.02	0.00	0.52

Figure 12. Confusion matrix showing the benchmarking of the 15 classes of the EdgeIIoT dataset with the PCC FS methods.



	Backdoor	DDoS_HTTP	DDoS_ICMP	DDoS_TCP	DDoS_UDP	Fingerprinting	MITM	Normal	Password	Port_Scanning	Ransomware	SQL_injection	Uploading	Vulnerability_scanner	XSS
Backdoor	0.91	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.04	0.00	0.00
DDoS_HTTP	0.01	0.62	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.14	0.02	0.09	0.10
DDoS_ICMP	0.00	0.00	0.99	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DDoS_TCP	0.00	0.00	0.00	0.90	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.02	0.00	0.00
DDoS_UDP	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Fingerprinting	0.00	0.00	0.00	0.10	0.00	0.69	0.00	0.00	0.00	0.12	0.08	0.00	0.00	0.00	0.00
MITM	0.00	0.00	0.00	0.00	0.01	0.00	0.34	0.00	0.00	0.00	0.66	0.00	0.00	0.00	0.00
Normal	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Password	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.00	0.00	0.30	0.02	0.07	0.25
Port_Scanning	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96	0.04	0.00	0.00	0.00	0.00
Ransomware	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.88	0.00	0.08	0.00	0.00
SQL_injection	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.66	0.00	0.05	0.20
Uploading	0.00	0.05	0.00	0.06	0.00	0.00	0.00	0.00	0.02	0.00	0.02	0.08	0.41	0.00	0.36
Vulnerability_scanner	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.03	0.00	0.90	0.04
XSS	0.01	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.12	0.02	0.00	0.72

Figure 13. Confusion matrix showing the benchmarking of the 15 classes of the EdgeIIoT dataset with the DT-FS methods.

Table 2. Performance of the evaluated methods on the EdgeIIoT dataset (4 classes).

Model	Feature Selection	MCC %	Loss	F1-Score %	Accuracy %	Precision %	Recall %	Train Time (s)
DNN	DT	78.96	0.2472	91	91	91	91	224
RNN	DT	86.90	0.2606	91	91	92	91	270
1D-CNN-GRU-FCN	DT	88.30	0.2078	92	92	92	92	1458
CNN-LSTM	DT	88.32	0.2008	92	92	92	92	1283
CNN-BiLSTM	DT	88.38	0.1991	92	92	92	92	1841

In comparison, the other four models, including DNN, RNN, ID-CNN-GRU-FCN, and CNN-LSTM, exhibit lower MCC scores than the CNN-BiLSTM model, with scores of 78.96%, 86.90%, 88.30%, and 88.32%, respectively. The MCC score proves to be a reliable metric for evaluating the proposed zero-trust NIDS model, offering accurate predictions and performance insights for each explored model.

The CNN-BiLSTM model significantly contributes to enhanced security by detecting a broader range of attacks, despite having the longest training time (1841s). Designed to detect various attack scenarios, including complex and evolving ones, the CNN-BiLSTM model's parameters demonstrate depth and intricacy, resulting in a more accurate model for cyberattack detection. Meanwhile, other models with fewer parameters and lower MCC scores lead to faster training times.

To gain a deeper understanding of the CNN-BiLSTM NIDS model's performance in multi-class classification (4 classes), the confusion matrices in Figure 14 provide valuable insights into how the model classifies different categories. The model accurately identifies DDoS, enumeration, malware, and normal attack types with percentages of 0.82%, 0.91%, 0.96%, and 1.0%, respectively. The high accuracy aligns with the principles of zero-trust detection, emphasizing vigilance, continuous monitoring, optimal threat coverage, and reduced false positives.

4.3. Extended Analysis with the CICIoT2023 Dataset

CNN-BiLSTM NIDS Model with 4 Classes

The CICIoT2023 dataset employed in the experiment contains 30 classes of cyber-attacks (BackdoorMalware, BenignTraffic, BrowserHijacking, CommandInjection, DDoS-ACKFragmentation, DDoS-HTTPFlood, DDoS-ICMPFlood, DDoS-ICMPFragmentation, DDoS-PSHACKFlood, DDoS-RSTF-IN-Flood, DDoS-SYNFlood, DDoS-SlowLoris, DDoS-Synonymous-IPFlood, DDoS-TCPFlood, DDoS-UDPFlood, DDoS-UDPFragmentation, DNS-Spoofing, DictionaryBruteForce, DoS-HTTP-Flood, DoS-SYN-Flood, DoS-TCP-Flood, DoS-UDP-Flood, MITM-ArpSpoof, Mirai-greeth-flood, Mirai-greip-flood, Mirai-udpplain, Recon-HostDiscovery, Recon-OSScan, Recon-PingSweep, Recon-PortScan, SqlInjection, Uploading-Attack, Vulnerability-Scan, and XSS).

Cyberattacks targeting IoT CPSs exhibit a common pattern and are thus categorizable into five overarching types: benign attacks, Distributed Denial of Service (DDoS) attacks, DoS attacks, and enumeration attacks. Table 3 presents a further analysis of the zero-trust NIDS model using state-of-the-art deep-learning models and the CICIoT dataset. Within the multi-class experiment, the CNN-BiLSTM NIDS model with the DT-FS method outperforms other models (DNN, RNN, ID-CNN-GRU-FCN, and CNN-LSTM)in terms of the highest MCC score and least loss value. The CNN-BiLSTM model attains an MCC score of 97.33 and 0.0318 while yielding a weighted F1-score, accuracy, precision, and recall of 99% respectively.

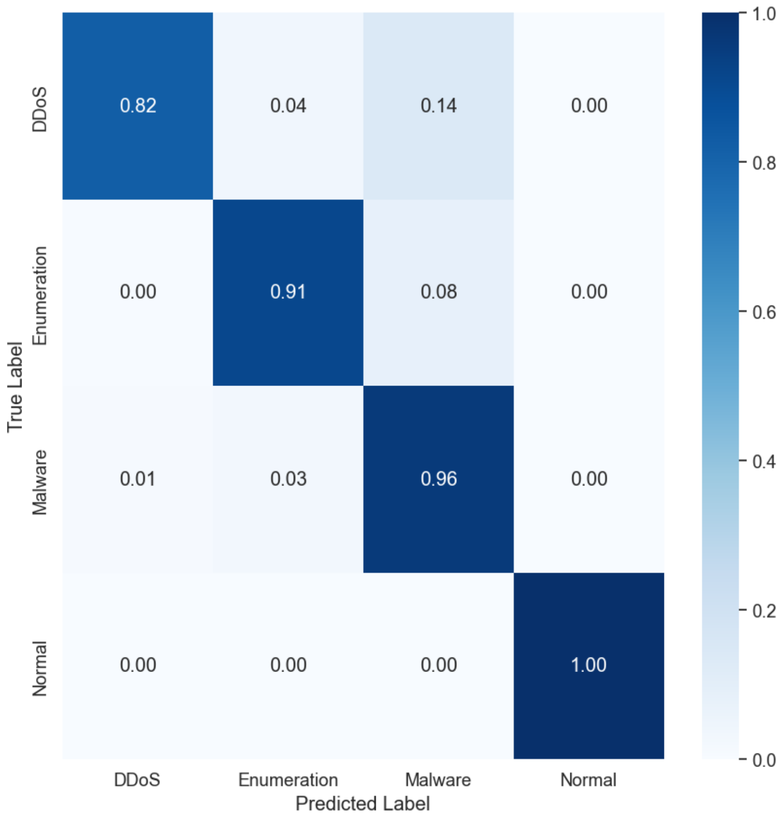


Figure 14. Confusion matrix of the multi-class (4 classes) prediction with the 2023 EdgeIIoT dataset.

Table 3. Performance of the evaluated methods on the CICIoT 2023 dataset (5 classes).

Model	Feature Selection	MCC %	Loss	F1-Score %	Accuracy %	Precision %	Recall %	Train Time (s)
CNN-LSTM	DT	69.06	0.3155	86	87	87	87	182.74
DNN	DT	71.34	0.3009	87	88	88	88	214.72
RNN	DT	85.06	0.1596	93	93	93	93	471.73
1D-CNN-GRU-FCN	DT	97.00	0.0341	99	99	99	99	746.98
CNN BiLSTM	DT	97.33	0.0318	99	99	99	99	1021.9

The remarkable results of the CNN-BiLSTM model, particularly regarding MCC score and other performance metrics, strongly suggest its potential utility within a zero-trust marine NIDS framework. The confusion matrix in Figure 15 illustrates the performance of the CNN-BiLSTM model. The NIDS model can identify all DDoS and DOS attacks very accurately (1.0), followed by benign traffic (0.91), malware traffic (0.92), and enumeration attacks (0.67).

The sustained optimal performance of the NIDS model across various standard datasets underscores the essential requirement for robust threat detection in marine cyberdefense. While the model consistently demonstrates high efficacy, the imperative for XAI methods becomes apparent, particularly in unraveling the intricacies of the black-box NIDS model, such as the CNN-BiLSTM architecture (i.e., its certainty and intrinsic model nature). This pursuit of transparency is crucial for enhancing marine cyberdefense capabilities.

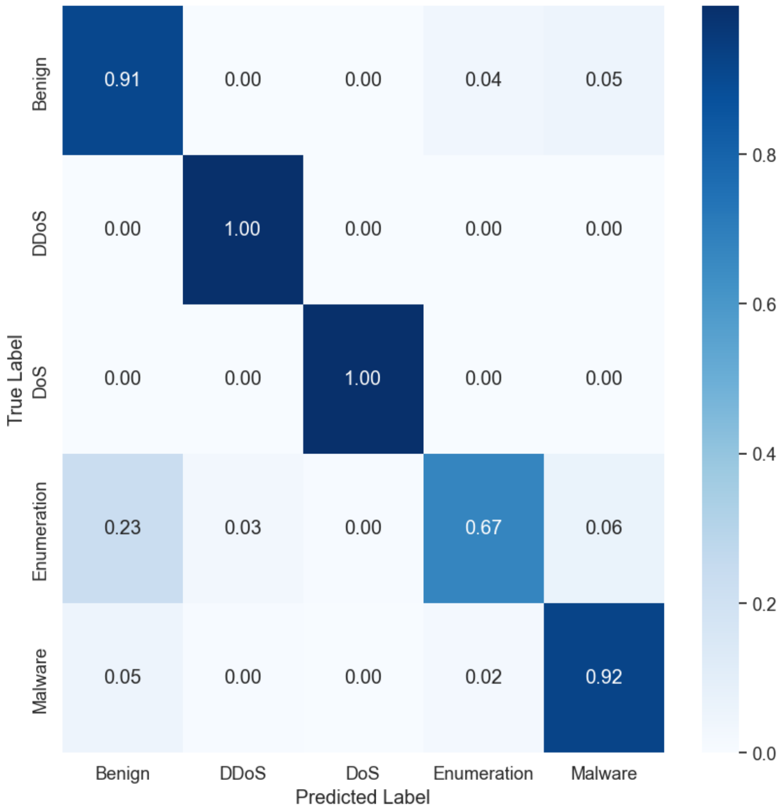


Figure 15. Confusion matrix of the multi-class (5 classes) prediction with the 2023 CICIoT dataset.

4.4. Model Explainability

The SHAP explainer, as represented in Equation (9), provides the marginal value of contributions made by a subset of features. The SHAP explainability method is a cooperative game theory method introduced by authors in [52], which basically provides the marginal value of contributions made by a feature or subset of features within a model's prediction. Since most models with large data are black-box in nature, the SHAP method addresses the issue of clarity and reason for model predictions. SHAP values can be calculated by the equation provided below:

$$ {}_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'!|(M-|z'|-1)!}{M!}[f_x(z'-f_x(z')i)], \tag{9} $$

where M is the total number of features, ${}_i$ denotes the shapely value for feature i in model (f), with input data points (x), and z', $f_x(z')i$ represents the model output with and without target features.

The LIME explainer generates local surrogate models to approximate the decision-making process of a complex model, providing interpretable explanations for individual predictions by highlighting important features. LIME explanatory values can be calculated by Equation (10):

$$ \xi(x) =^{argmin}_{g \in G} L(f,g,\pi_x) + \sigma(g), \tag{10} $$

where $\xi(x)$ represents the explanation, x is an input value, g is a simple interpretable model, G is a class of potentially interpretable models, f is a complex (blackbox) model, π_x is the proximity, and $\sigma(g)$ is a complexity measure.

To facilitate informed decision-making for marine security experts, a comprehensive quantitative evaluation of the SHAP and LIME explainers is conducted. This assessment employed decision and confidence impact ratios to provide deeper insight and foster confidence in the effectiveness of the XAI methods. XAI professionals can leverage the decision impact ratio and confidence impact ratio to assess the significance of decisions and the reliability (confidence) of an explanation method [53], thus enhancing the security and trustworthiness of AI models.

Equation (11) evaluates a quantitative SHAP and LIME explanation of the CNN-BiLSTM NIDS model using the following metrics:

Decision Impact Ratio (DIR): DIR refers to the rate of change in decisions owing to the omission of critical network traffic features in the interpretation method.

$$DIR = \sum_{i}^{N} \frac{1D(x_i) \neq D(x_i - c_i)}{N}, \tag{11}$$

where x_i denotes the ith original sample and c_i denotes the critical area marked by the model for the ith sample.

Confidence impact ratio (CIR): CIR signifies the percentage decline in confidence due to the omissions of critical network traffic features in the interpretation method, as in Equation (12):

$$CIR = \sum_{i}^{N} \frac{max(C(x_i) - C(x_i - c_i), 0)}{N}. \tag{12}$$

An evaluation of the explainability methods helps obtain a subjective assessment of the security expert's trust and assessments of the CNN-BiLSTM model's trustworthiness.

4.4.1. Visualization of XAI Results and Deductions

A visual XAI of NIDS models can create a more user-friendly interpretation, fostering trust and understanding in the deployment of NIDS models in marine cybersecurity. By incorporating visual XAI techniques in marine NIDS, cybersecurity professionals can better understand and trust the decisions made by AI models, making the overall system more effective and accountable. It also facilitates collaboration between AI systems and human analysts, leading to improved network security. Understanding the explanations provided by XAI tools in the context of NIDS can vary depending on the complexity of the explanations and the technical background of the end-user. Such users must be familiar with the fundamental concepts of cybersecurity, networking, machine learning, and critical thinking. The visual XAI experiment in this study was carried out on the 2023 CICIoT dataset since it has a higher volume of traffic samples than the 2023 EdgeIIoT dataset.

Visual XAI with SHAP:

This study employs the SHAP values to explain the predictions of a CNN-BiLSTM NIDS model on 10 selected instances since calculating SHAP values for over 30,000 instances can be computationally expensive. Next, the Kernel SHAP explainer is wrapped with the predict-wrapper, which is used to estimate the Shapley values for the selected test data samples. The SHAP XAI method is model-agnostic and is used by aggregating the results obtained locally.

The SHAP visualization in Figure 16 provides insights into the influence of individual training features on the model's prediction of a malware (blue) and DoS (purple) attack. As depicted, the descending order of IAT ($\geq$0.4), UDP ($\geq$0.1), Magnitude, Total size, protocol type, etc., illustrates the significance of these network traffic features on the CNN-BiLSTM NIDS model. The plot shows that the IAT feature, with the longest bar, had the greatest influence on the model's prediction of the malware and DoS attack. Leveraging on the SHAP explanations, the marine NIDS expert can now make room for feature importance prioritization and guidance for feature engineering in a case of low detection and can satisfy the black-box question of "what training features dominated the NIDS model's prediction?"

To gain a global behavior of the NIDS model, the feature importance plot in Figure 17 provides summary feature importance visualizations across instances in the selected test sample. Each dot represents the aggregated impact of a feature. If precomputed SHAP values are available and the goal is to provide a global summary of feature importance, the plot in Figure 17 becomes more suitable. This visualized interpretation also underscores the essence of proper feature selection techniques before model training. The participating features show reasonable/almost equal strength (0.4–0.7) on the model output. Overall, the key highlight of the first snippet in Figure 16 shows the convenience of using precomputed SHAP values for global feature importance summaries, while the second snippet in Figure 17 unveils the dynamic computation of SHAP values for a more detailed, instance-specific analysis.

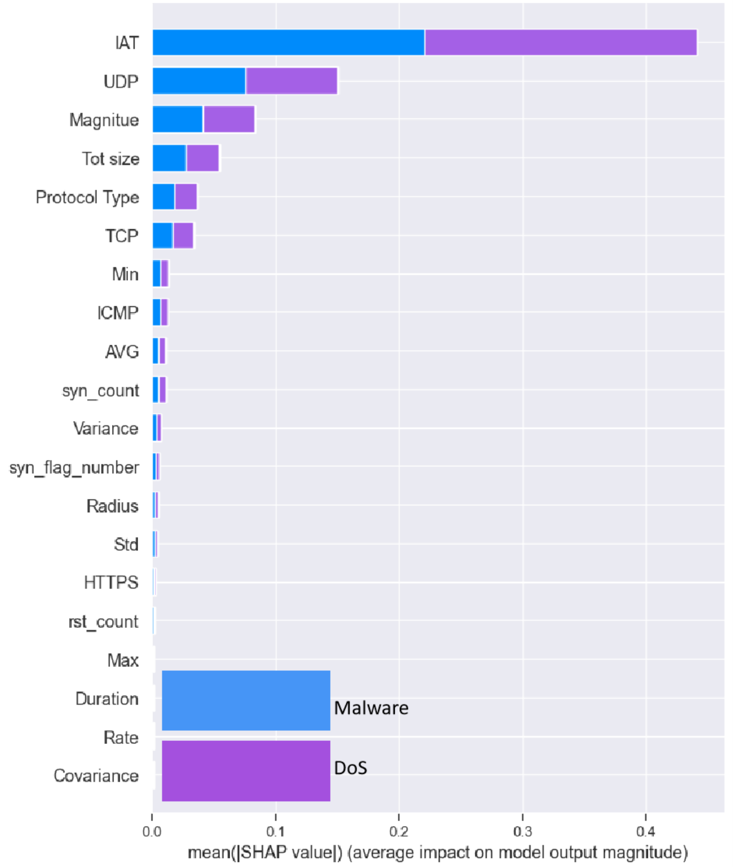


Figure 16. SHAP plot showing the feature contribution to the model's output magnitude for 10 traffic instances.

Visual XAI with LIME:

The LIME visual XAI explainer in Figure 18 provides local prediction probabilities for a single instance of test data to justify the NIDS model reliability, i.e., "how certain is the NIDS model's prediction of an attack?" The visual plot can aid in modeling reliability, transparency, and judgment of cyberattack instances from benign instances by marine security experts.

A random visualized explanation in Figure 18 is investigated to interpret the NIDS model's prediction within a specific instance to ascertain if it is either a DDoS, DoS, malware, or enumeration attack (multi-class). However, the visual explanations using the LIME explainer reveal a strong prediction of an enumeration attack and less likelihood of a DDoS, DoS, malware, or benign attack. The enumeration attack is strongly predicted, with a high probability rate of 94%, while the DDoS, DoS, malware, and benign attacks have low prediction probabilities of 0.04%, 0.00%, 0.00%, and 0.02%, respectively. The features and corresponding values that led to the model's prediction of an enumeration attack is also highlighted, with different weighted feature values. The provided LIME XAI visualization can aid a cybersecurity expert's understanding of collaborative human-in-the-loop NIDS monitoring and explanation while preventing false alarm/distress calls and ambiguity in addressing cyberattacks.

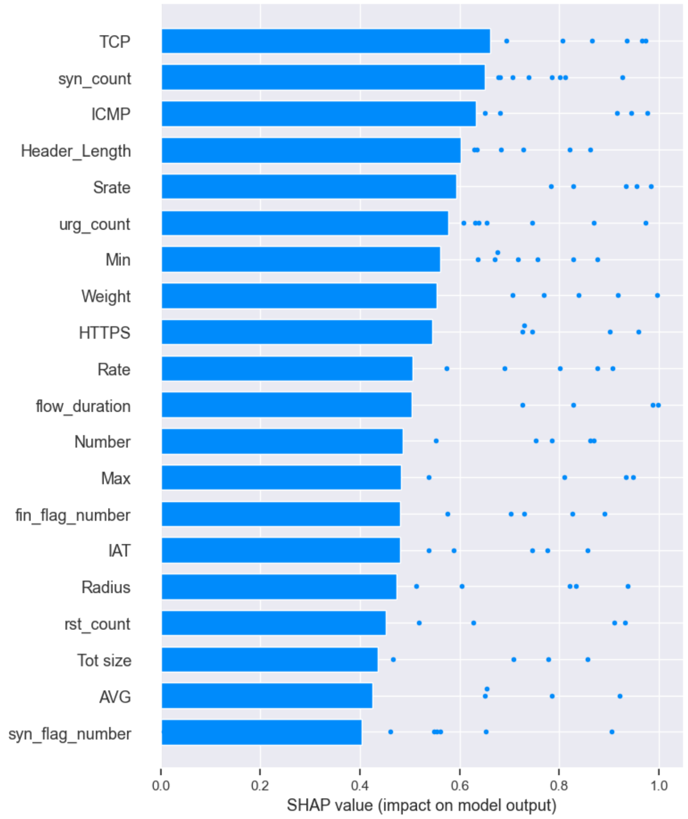
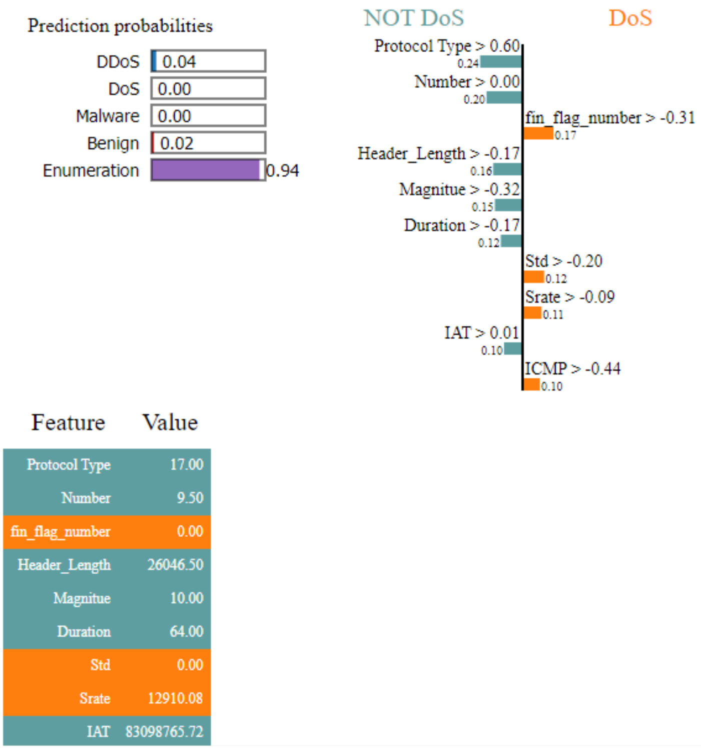


Figure 17. SHAP feature importance.



Feature	Value
Protocol Type	17.00
Number	9.50
fin_flag_number	0.00
Header_Length	26046.50
Magnitue	10.00
Duration	64.00
Std	0.00
Srate	12910.08
IAT	83098765.72

Figure 18. LIME probability prediction.

Furthermore, another instance of sample data is selected to investigate the feature(s) value's contribution to a model's prediction, within a single prediction. The plot in Figure 19 reveals the features that have positive values (green bars) and those with negative values (red bars). The visualizations of strong features such as the fin-count, Rate, Duration, Protocol Type can be valuable for understanding and validating the model's behavior in specific instances. This visualization enhances model interpretability and supports decision-making in not only marine cyberdefense but also in real-world applications by providing insights into why the model made specific predictions.

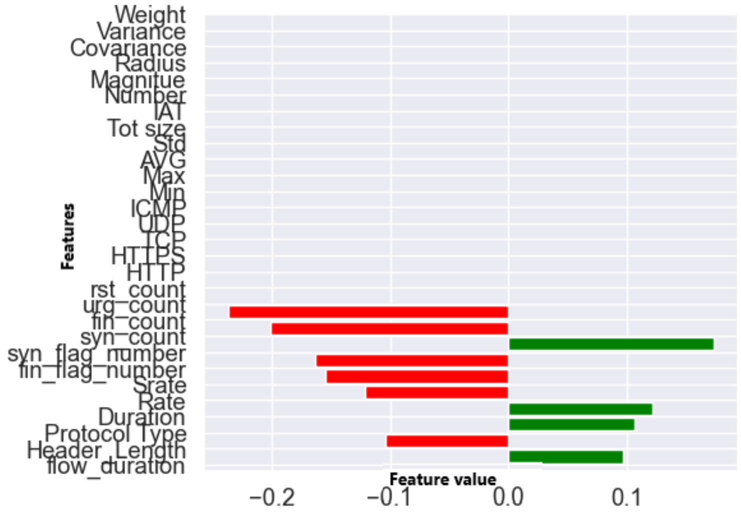


Figure 19. LIME for local feature contributions for an instance where the positive, and negative contributions are represented with green and red colour respectively.

4.4.2. Quantitate Performance Evaluation of the XAI Models Using DIR and CIR

The CIR and DIR calculations in Table 4 provide quantitative explanations of the explored XAI methods. Specifically, the CIR and DIR values provide more confidence to a security expert on the potential use of either the SHAP or LIME XAI methods for NIDS model interpretations [53]. As shown in Table 4, in the quantitative calculations of each method the LIME yields higher DIR and CIR values than the SHAP explainer, within a single subset of test data calculated. A high DIR by the LIME means that changes in that feature have a substantial effect on the model's prediction. A low DIR by the SHAP suggests that variations in that feature have minimal impact on the prediction.

In terms of confidence impact by the SHAP and LIME explainer, the lower CIR by the SHAP implies that the model is less certain than the LIME XAI, and its prediction may be more uncertain or influenced by various factors such as model features or architecture. The experimental results in this study revealed similar values of DIR/CIR within the LIME XAI method, which suggests that there could be a strong relationship between some network traffic feature(s), which reflects consistently on the confidence/decision impact of the LIME explainer.

Table 4. Quantitative explainability of the CNN-BiLSTM NIDS model using the CIR/DIR.

Quantitative XAI	Decision Impact Ratio (DIR)	Confidence Impact Ratio (CIR)
SHAP	0.696	0.129
LIME	1.476	1.476

4.5. Comparision with SOA Methods

Previous frameworks designed to address IoT marine cyberthreats have gained notable acceptance within the NIDS domain. However, contemporary research trends are increasingly oriented toward zero-trust defenses. This study introduces a novel framework aspiring to establish a zero-trust marine NIDS by leveraging AI-enabled NIDS and XAI methodologies.

As presented in Table 5, some researchers in the NIDS domain have utilized XAI techniques to enhance model trustworthiness in IoT and marine cyberdefense domains. However, our novel approach, detailed in Section 3.1, and the corresponding experimental results discussed in Section 4, incorporating both quantitative and visual XAI methods, hold significant potential for the development of a robust zero-trust marine NIDS.

Table 5. Comparison with existing studies.

Authors	Model	Dataset	No. of Classes	Accuracy %	Visual XAI	Quantitative XAI
[54]	MAGRU	EdgeIIoT	-	99	X	X
[24]	CVAE-GAN	NSL-KDD	5	95	X	X
[9]	DNN	NSL-KDD UNSW-BB15	2	88, 99	Y	X
Ours	CNN-BiLSTM	2023 EdgeIIoT 2023 CICIoT	4, 5	92, 99	Y	Y

5. Conclusions

This study identified the fast-growing concerns of cyberattacks in marine communications and the need for adopting a zero-trust security paradigm—"trust but verify all"—for enhanced NIDS model security against marine cyberthreats. By employing a deep-learning-based NIDS model (CNN-BiLSTM) with a proper feature selection, the proposed method was capable of identifying diverse cyberthreat patterns in marine networks using recent cybersecurity datasets. Next, the employment of XAI methods was investigated to satisfy model debugging and trustworthy predictions of cyberattacks to prevent false alarms. Specifically, a visual and quantitative XAI approach using the SHAP and LIME explainer was explored to ease interpretability and XAI concerns in the domain of marine network intrusion detection.

Future directions worth exploring include reducing the time complexity of SHAP processes to enable faster and more efficient model interpretability. Additionally, in an adversarial scenario, bad actors can perpetuate anomalous XAI interpretations to gain unauthorized explanations of NIDS models, thereby compromising model security. Preventing anomalous XAI should be of key interest in our future work. Furthermore, the integration of privacy-inspired federated learning models with blockchain could provide stronger security for zero-trust NIDS deployment.

Author Contributions: Conceptualization, E.C.N., J.N.N. and C.I.N.; methodology, E.C.N., J.N.N. and C.I.N.; software, E.C.N.; validation, E.C.N., J.N.N., C.I.N. and D.-S.K.; formal analysis, E.C.N., J.N.N. and C.I.N.; investigation, E.C.N., J.N.N. and C.I.N.; resources, J.-M.L., and D.-S.K.; data curation, E.C.N., J.N.N. and C.I.N.; writing—original draft preparation, E.C.N., J.N.N. and C.I.N.; writing—review and editing, E.C.N. and C.I.N.; visualization, E.C.N. and J.N.N.; supervision, C.I.N., J.-M.L. and D.-S.K.; project administration, C.I.N., J.-M.L. and D.-S.K.; funding acquisition, J.-M.L. and D.-S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Ministry of Education, Science and Technology (MEST) and the Ministry of Science and ICT (MIST), Korea.

Data Availability Statement: Data is contained within the article.

Acknowledgments: This work was supported by Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (MEST) (2018R1A6A1A03024003) & Ministry of Science and ICT (MIST), Korea, under the Innovative Human Resource Development for Local Intellectualization support program (IITP-2024-2020-0-01612) supervised by the Institute for Information & communications Technology Planning & Evaluation (IITP).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ANN	Artificial Neural Network
Bi-LSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
CNN-BiLSTM	CNNs with Bidirectional Long Short-Term Memory (BiLSTM) network
CNN-FRU-FCN	Convolutional Neural Network with Fully-Connected
CNN-LSTM	Convolutional Neural Network-Long Short-Term Memory
CPS	Cyber–Physical System
CIR	Confidence Impact Ratio
DIR	Decision Impact Ratio
DNN	Deep Neural Network
DoS	Denial of Service
DT	Decision Tree
DDoS	Distributed Denial of Service
CAM	Class Activation Mapping
FS	Featurue Selection
GAN	Generative Adversarial Network
Grad-CAM	Gradient-weighted Class Activation Mapping
IMO	International Maritime Organization
IAT	Inter-Arrival Time
UDP	User Datagram Protocol
IoUT	Internet of Underwater Things
IoT	Internet of Things
LIME	Local Interpretable Model-agnostic Explanations
LSTM	Long Short-Term Memory
LTE	Long-Term Evolution
MCC	Mathew Correlations Coefficient
NIDS	Network Intrusion Detection System
NIST	National Institute of Standards and Technology
PCC	Pearson Correlation Coefficient
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
SDN	Software-Defined Networking
SHAP	SHapley Additive exPlanations
TCP/IPV6	Transmission Control Protocol/Internet Protocol Version 6
XAI	Explainable Artificial Intelligence
XGrad-CAM	Axiom-based Grad-CAM
ZTA	Zero-Trust Architecture

References

1. Serpanos, D.; Komninos, T. The Cyberwarfare in Ukraine. *Computer* **2022**, *55*, 88–91. [CrossRef]
2. Park, C.; Kontovas, C.; Yang, Z.; Chang, C.H. A BN driven FMEA approach to assess maritime cybersecurity risks. *Ocean Coast. Manag.* **2023**, *235*, 106480. [CrossRef]
3. Mohsan, S.A.H.; Li, Y.; Sadiq, M.; Liang, J.; Khan, M.A. Recent Advances, Future Trends, Applications and Challenges of Internet of Underwater Things (IoUT): A Comprehensive Review. *J. Mar. Sci. Eng.* **2023**, *11*, 124. [CrossRef]
4. Liu, W.; Xu, X.; Wu, L.; Qi, L.; Jolfaei, A.; Ding, W.; Khosravi, M.R. Intrusion Detection for Maritime Transportation Systems With Batch Federated Aggregation. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 2503–2514. [CrossRef]
5. Dong, B.; Wang, X. Comparison deep-learning method to traditional methods using for network intrusion detection. In Proceedings of the 2016 8th IEEE International Conference on Communication Software and Networks (ICCSN), Beijing, China, 4–6 June 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 581–585.
6. Nwakanma, C.I.; Ahakonye, L.A.C.; Njoku, J.N.; Odirichukwu, J.C.; Okolie, S.A.; Uzondu, C.; Ndubuisi Nweke, C.C.; Kim, D.S. Explainable Artificial Intelligence (XAI) for Intrusion Detection and Mitigation in Intelligent Connected Vehicles: A Review. *Appl. Sci.* **2023**, *13*, 1252. [CrossRef]
7. Shore, M.; Zeadally, S.; Keshariya, A. Zero Trust: The What, How, Why, and When. *Computer* **2021**, *54*, 26–35. [CrossRef]
8. Capuano, N.; Fenza, G.; Loia, V.; Stanzione, C. Explainable Artificial Intelligence in CyberSecurity: A Survey. *IEEE Access* **2022**, *10*, 93575–93600. [CrossRef]
9. Houda, Z.A.E.; Brik, B.; Khoukhi, L. "Why Should I Trust Your IDS?": An Explainable Deep Learning Framework for Intrusion Detection Systems in Internet of Things Networks. *IEEE Open J. Commun. Soc.* **2022**, *3*, 1164–1176. [CrossRef]
10. Ali, E.S.; Saeed, R.A.; Eltahir, I.K.; Khalifa, O.O. A systematic review on energy efficiency in the internet of underwater things (IoUT): Recent approaches and research gaps. *J. Netw. Comput. Appl.* **2023**, *213*, 103594. [CrossRef]
11. Khan, Z.U.; Gang, Q.; Muhammad, A.; Muzzammil, M.; Khan, S.U.; Affendi, M.E.; Ali, G.; Ullah, I.; Khan, J. A comprehensive survey of energy-efficient MAC and routing protocols for underwater wireless sensor networks. *Electronics* **2022**, *11*, 3015. [CrossRef]
12. Heering, D.; Maennel, O.; Venables, A. Shortcomings in cybersecurity education for seafarers. In *Maritime Technology and Engineering 5 Volume 1*; CRC Press: Boca Raton, FL, USA, 2021; pp. 49–61.
13. Jacq, O.; Boudvin, X.; Brosset, D.; Kermarrec, Y.; Simonin, J. Detecting and hunting cyberthreats in a maritime environment: Specification and experimentation of a maritime cybersecurity operations centre. In Proceedings of the 2018 2nd Cyber Security in Networking Conference (CSNet), Paris, France, 24–26 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–8.
14. Lin, W.C. Maritime Environment Assessment and Management Using through Balanced Scorecard by Using DEMATEL and ANP Technique. *Int. J. Environ. Res. Public Health* **2022**, *19*, 2873. [CrossRef] [PubMed]
15. Akpan, F.; Bendiab, G.; Shiaeles, S.; Karamperidis, S.; Michaloliakos, M. Cybersecurity Challenges in the Maritime Sector. *Network* **2022**, *2*, 123–138. [CrossRef]
16. Jo, Y. Cyberattack Incidents in Maritime Sector. Available online: https://https://www.cytur.net/ (accessed on 10 October 2023).
17. Koulouras, I.; Bobotsaris, I.; Margariti, S.V.; Stergiou, E.; Stylios, C. Assessment of SDN Controllers in Wireless Environment Using a Multi-Criteria Technique. *Information* **2023**, *14*, 476. [CrossRef]
18. Liang, M.; Su, X.; Liu, X.; Zhang, X. Intelligent ocean convergence platform based on iot empowered with edge computing. *J. Internet Technol.* **2020**, *21*, 235–244.
19. Chen, H.; Yin, F.; Huang, W.; Liu, M.; Li, D. Ocean Surface Drifting Buoy System Based on UAV-Enabled Wireless Powered Relay Network. *Sensors* **2020**, *20*, 2598. [CrossRef] [PubMed]
20. Jongwoo, A. KR Maritime Cyber Safety News & Report. Available online: https://www.krs.co.kr/Common/Com_Popup/Com_FileDown.aspx?DATA1=7rF67H0cjeYuxn6YdejCySra1U5wS9J0jjGzbttW1YbZqalp5CIKgYVcAVRi6k!_!_!V&DATA2=W241p64Xg7ER4wTHluR9Dw==&DATA3=v5dA4mdXiDVTVUw536GDwhpm0u4qvoFnDtpDCl6AfYnL8GSQ3DqomHVFddy6UekCDDqQiK1aHiIRfNeXSsIong== (accessed on 9 September 2023).
21. Rehman, M.H.U.; Dirir, A.M.; Salah, K.; Damiani, E.; Svetinovic, D. TrustFed: A Framework for Fair and Trustworthy Cross-Device Federated Learning in IIoT. *IEEE Trans. Ind. Inform.* **2021**, *17*, 8485–8494. [CrossRef]
22. Nkoro, E.C.; Njoku, J.N.; Nwakanma, C.I.; Lee, J.M.; Kim, D.S. SHAP-Based Intrusion Detection Framework for Zero-Trust IoT Maritime Security. In Proceedings of the 2023 the 2nd International Conference on Maritime IT Convergence (ICMIC), Jeju Island, Korea, 23–25 August 2023; pp. 1–8.
23. Hou, T.; Xing, H.; Liang, X.; Su, X.; Wang, Z. A Marine Hydrographic Station Networks Intrusion Detection Method Based on LCVAE and CNN-BiLSTM. *J. Mar. Sci. Eng.* **2023**, *11*, 221. [CrossRef]
24. Su, X.; Tian, T.; Cai, L.; Ye, B.; Xing, H. A CVAE-GAN-based Approach to Process Imbalanced Datasets for Intrusion Detection in Marine Meteorological Sensor Networks. In Proceedings of the 2022 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), Melbourne, Australia, 17–19 December 2022; pp. 197–203. [CrossRef]
25. Kalluri, R.; Mahendra, L.; Kumar, R.S.; Prasad, G.G. Simulation and Impact Analysis of Denial-of-Service Attacks on Power SCADA. In Proceedings of the 2016 National Power Systems Conference (NPSC), Bhubaneswar, India, 19–21 December 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–5. [CrossRef]

26. Stafford, V. Zero trust architecture. *NIST Spec. Publ.* **2020**, *800*, 207.
27. Freter, R. Department of Defence (DoD) Zero Trust Reference Architecture, Version 2.0. In Proceedings of the Defense Information Systems Agency (DISA) and National Security Agency (NSA), July 2022. Available online: https://dodcio.defense.gov/Portals/0/Documents/Library/(U)ZT_RA_v2.0(U)_Sep22.pdf (accessed on 9 September 2023).
28. Abuhasel, K.A. A Zero-Trust Network-Based Access Control Scheme for Sustainable and Resilient Industry 5.0. *IEEE Access* **2023**, *11*, 116398–116409. [CrossRef]
29. Li, S.; Iqbal, M.; Saxena, N. Future industry internet of things with zero-trust security. *Inf. Syst. Front.* **2022**, 1–14. [CrossRef]
30. Ali, B.; Hijjawi, S.; Campbell, L.H.; Gregory, M.A.; Li, S. A maturity framework for zero-trust security in multiaccess edge computing. *Secur. Commun. Netw.* **2022**, *3178760*, 1–14. [CrossRef]
31. Lee, B.; Vanickis, R.; Rogelio, F.; Jacob, P. Situational awareness based risk-adapatable access control in enterprise networks. *arXiv* **2017**, arXiv:1710.09696.
32. Syed, N.F.; Shah, S.W.; Shaghaghi, A.; Anwar, A.; Baig, Z.; Doss, R. Zero Trust Architecture (ZTA): A Comprehensive Survey. *IEEE Access* **2022**, *10*, 57143–57179. [CrossRef]
33. Restuccia, F.; D'Oro, S.; Melodia, T. Securing the Internet of Things in the Age of Machine Learning and Software-Defined Networking. *IEEE Internet Things J.* **2018**, *5*, 4829–4842. [CrossRef]
34. House, W. FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. Available online: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/ (accessed on 1 October 2023).
35. Dieber, J.; Kirrane, S. Why model why? Assessing the strengths and limitations of LIME. *arXiv* **2020**, arXiv:2012.00093. Available online: http://arxiv.org/abs/2012.00093 (accessed on 21 July 2023).
36. Liu, Y.; Tantithamthavorn, C.; Li, L.; Liu, Y. Explainable AI for Android Malware Detection: Towards Understanding Why the Models Perform So Well? In Proceedings of the 2022 IEEE 33rd International Symposium on Software Reliability Engineering (ISSRE), Charlotte, NC, USA, 31 October–3 November 2022; pp. 169–180. [CrossRef]
37. Szczepański, M.; Choraś, M.; Pawlicki, M.; Kozik, R. Achieving Explainability of Intrusion Detection System by Hybrid Oracle-Explainer Approach. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8. [CrossRef]
38. Sharafaldin, I.; Lashkari, A.H.; Ghorbani, A.A. Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp* **2018**, *1*, 108–116.
39. Patil, S.; Varadarajan, V.; Mazhar, S.M.; Sahibzada, A.; Ahmed, N.; Sinha, O.; Kumar, S.; Shaw, K.; Kotecha, K. Explainable Artificial Intelligence for Intrusion Detection System. *Electronics* **2022**, *11*, 3079. [CrossRef]
40. Sivamohan, S.; Sridhar, S.; Krishnaveni, S. TEA-EKHO-IDS: An intrusion detection system for industrial CPS with trustworthy explainable AI and enhanced krill herd optimization. *Peer Peer Netw. Appl.* **2023**, *16*, 1993–2021. [CrossRef]
41. Tavallaee, M.; Bagheri, E.; Lu, W.; Ghorbani, A.A. A detailed analysis of the KDD CUP 99 data set. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, ON, Canada, 8–10 July 2019; IEEE: Piscataway, NJ, USA, 2009; pp. 1–6.
42. Alani, M.M. BotStop: Packet-based efficient and explainable IoT botnet detection using machine learning. *Comput. Commun.* **2022**, *193*, 53–62. [CrossRef]
43. Alani, M.M.; Awad, A.I.; Barka, E. ARP-PROBE: An ARP spoofing detector for Internet of Things networks using explainable deep learning. *Internet Things* **2023**, *23*, 100861. [CrossRef]
44. Ferrag, M.A.; Friha, O.; Hamouda, D.; Maglaras, L.; Janicke, H. Edge-IIoTset: A New Comprehensive Realistic Cyber Security Dataset of IoT and IIoT Applications for Centralized and Federated Learning. *IEEE Access* **2022**, *10*, 40281–40306. [CrossRef]
45. Xu, B.; Sun, L.; Mao, X.; Ding, R.; Liu, C. IoT Intrusion Detection System Based on Machine Learning. *Electronics* **2023**, *12*, 4289. [CrossRef]
46. Neto, E.C.P.; Dadkhah, S.; Ferreira, R.; Zohourian, A.; Lu, R.; Ghorbani, A.A. CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment. *Sensors* **2023**, *23*, 5941. [CrossRef]
47. Fahad, A.; Tari, Z.; Khalil, I.; Habib, I.; Alnuweiri, H. Toward an efficient and scalable feature selection approach for internet traffic classification. *Comput. Netw.* **2013**, *57*, 2040–2057. [CrossRef]
48. Oh, B.H.; Vural, S.; Wang, N.; Tafazolli, R. Priority-Based Flow Control for Dynamic and Reliable Flow Management in SDN. *IEEE Trans. Netw. Serv. Manag.* **2018**, *15*, 1720–1732. [CrossRef]
49. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef]
50. He, K.; Kim, D.D.; Asghar, M.R. Adversarial Machine Learning for Network Intrusion Detection Systems: A Comprehensive Survey. *IEEE Commun. Surv. Tutor.* **2023**, *25*, 538–566. [CrossRef]
51. Buczak, A.L.; Guven, E. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 1153–1176. [CrossRef]
52. Shapley, L.S.; Shubik, M. The Assignment Game I: The core. *Int. J. Game Theory* **1971**, *1*, 111–130. [CrossRef]

53. Zou, L.; Goh, H.L.; Liew, C.J.Y.; Quah, J.L.; Gu, G.T.; Chew, J.J.; Prem Kumar, M.; Ang, C.G.L.; Ta, A. Ensemble Image Explainable AI (XAI) Algorithm for Severe Community-Acquired Pneumonia and COVID-19 Respiratory Infections. *IEEE Trans. Artif. Intell.* **2022**, *4*, 242–254. [CrossRef]
54. Ullah, S.; Boulila, W.; Koubâa, A.; Ahmad, J. MAGRU-IDS: A Multi-Head Attention-Based Gated Recurrent Unit for Intrusion Detection in IIoT Networks. *IEEE Access* **2023**, *11*, 114590–114601. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.