Unravelling the Black Box: Enhancing Virtual Reality Network Security with Interpretable Deep Learning-Based Intrusion Detection System

Urslla Uchechi Izuazu, Dong-Seong Kim, Jae Min Lee
Department of IT Convergence Engineering, Kumoh National Institute of Technology, Gumi, South Korea
(uursla8, dskim, ljmpaul)@kumoh.ac.kr

***Abstract*—This study addresses the critical need to secure VR network communication from non-immersive attacks, employing an intrusion detection system (IDS). While deep learning (DL) models offer advanced solutions, their opacity as "black box" models raises concerns. Recognizing this gap, the research underscores the urgency for DL-based explainability, enabling data analysts and cybersecurity experts to grasp model intricacies. Leveraging sensed data from IoT devices, our work trains a DL-based model for attack detection and mitigation in the VR network, Importantly, we extend our contribution by providing comprehensive global and local interpretations of the model's decisions post-evaluation using SHAP-based explanation.**

***Index Terms*—Virtual Reality, XAI, Metaverse, Deep learning, Machine Learning, Intrusion Detection.**

I. Introduction

The emergence of virtual reality (VR) technology has ushered in a transformative phase of immersive and interactive user experience. Through the amalgamation of visual, haptic, and auditory stimuli, VR endeavors to transport individuals into simulated environments that evoke a profound sense of realism and engagement. [1]. The domain of VR technology is undergoing extensive adoption and is positioned for notable progress in the global market. Forecasts suggest that by 2024, the VR industry is expected to attain a remarkable valuation of $44.07 billion.

However, the rapid evolution of VR technology presents notable challenges that expose VR networks' vulnerability to malicious interference from within the network, raising paramount concerns about security threats [2]. Illegitimate users frequently employ various tactics to exploit the vulnerable architecture of VR high-speed networks.

In addition, attackers engage in eavesdropping on private conversations or perpetrate other criminal activities within VR environments. In more extreme cases, they resort to an immersive attack known as the "Human Joystick Attack." In this attack, the attackers manipulate the VR experience of users by superimposing images in their field of vision, leading to potential collisions with physical objects and walls [3].

Furthermore, through the exploitation of critical system vulnerabilities and compromised devices, malicious entities can infiltrate real-world devices, encompassing household appliances, thereby posing risks to personal safety and compromising critical infrastructure like water supply networks, power grids, and high-speed rail systems. These threats manifest in the form of advanced persistent attacks (APTs) as highlighted in [4], and numerous other attack vectors of similar nature.

A. Explainable DL-based IDS

Despite progress in DL-based Intrusion Detection Systems (IDS), their intricate models pose interpretation challenges, especially for non-experts [5]. This opacity hinders trust and user implementation of DL-based NIDS. Black-box models lack explanations, limiting optimization based on outputs [6]. In crucial domains like medical diagnosis or threat detection, blind reliance on models can lead to severe consequences [7]. Before deployment, assessing model performance and alignment with goals is vital. Traditional metrics may not capture real-world variations or purpose, making predictions and their explanations valuable for reliability assessment.

To address this, Artificial Intelligence (AI) embraces eXplainable AI (XAI), enhancing model interpretability [8]. XAI makes IDSs more understandable for cybersecurity experts.

AI model interpretability divides into intrinsic and post hoc types. Intrinsic integrates interpretability into model architecture, e.g., rule-based and decision tree models. Post hoc creates simpler surrogate models approximating complex ones [9]. Employing both enhances understanding and transparency in various applications, including cybersecurity.

In [10], model interpretability is further classified into two types, local and global aspects. Local explains individual predictions, revealing rationales. Global offers insights into overall model behavior and feature interactions [9].

As the field of XAI gains traction and finds applications in various domains like natural language processing, and computer vision, it becomes essential to extend these advancements to IDS to demystify its internal mechanism, to foster practical deployment.

In exploring model explainability, various approaches exist [11]. This study prioritizes the Shapley Additive Explanations (SHAP) method for model interpretation.

Concretely, this paper presents the following contributions:

979-8-3503-1327-7/23/$31.00 ©2023 IEEE

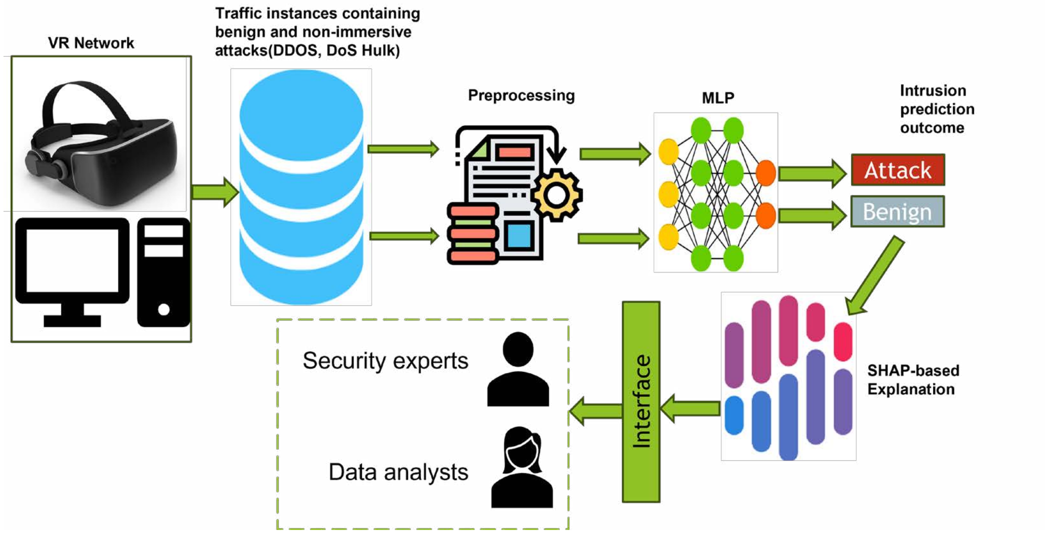
VR Network
Traffic instances containing benign and non-immersive attacks(DDOS, DoS Hulk)
Preprocessing
MLP
Intrusion prediction outcome
Attack
Benign
SHAP-based Explanation
Interface
Security experts
Data analysts

Fig. 1. Overview of System Design

1) The design of a Deep Learning-based Intrusion Detection Framework to protect VR networks against emerging non-immersive attacks.
2) The integration of SHAP-based explainability techniques with the aim of providing data analysts and security experts, with a systematic means of explaining decisions made by DL-based IDS.
3) The validation of our framework with novel publicly available datasets, containing well-known and contemporary attacks which are good representatives of the complex and diverse nature of the current threat landscape, and the associated cyber-security challenges.

The paper's structure is outlined as follows: Section II reviews research on XAI in IDS. Section III outlines the proposed framework. In Section IV, we assess our model's performance, compare it with state-of-the-art, and offer SHAP-based explanations for the decisions. Lastly, Section V provides conclusions.

B. Background Information on XAI in IDS and Research Gaps

In [9], the authors proposed an interpretable DL-based IDS for detecting and mitigating adversarial attacks in IoT networks. They employed adversarial training, addressing untargeted and white-box attacks, and integrated SHAP-based explanations to reveal the model's decision rationale, emphasizing critical classification features.

Similarly, [8] introduced a SHAP-based XAI framework for producing local and global model explanations. They justified their SHAP usage due to its strong theoretical basis and versatility across various models, unlike other XAI methods. They constructed two classifiers and compared their interpretations.

In [6], the authors employed three XAI methods (Rulefit, LIME, SHAP) to interpret their DL-based model designed for IoT network threat detection. They aimed to address the question of trustworthiness in their IDS by exploring both linear and non-linear techniques for local and global interpretation.

To enable dynamic access control in an SDN, authors in [12] employed an Anomaly-based RNN. and explained prediction outcomes via linear regression model coefficients.

Despite significant efforts made by the aforementioned research works, the field of VR has remained largely unexplored in the context of IDS explainability. Additionally, the utilization of small-dimensional datasets in these endeavors raises concerns about their suitability for comprehensive representation.

Addressing the identified research gap, we introduce an XAI-driven system for non-immersive threat detection in VR networks. To overcome the limitations of older datasets, we employ a more representative dataset. This system integrates SHAP-based explanations, enhancing transparency for improved decision-making and risk mitigation.

II. System Model.

A. Model Architecture.

The proposed AI-driven framework, designed for threat detection and mitigation for non-immersive VR communication networks, is depicted in Fig. 1. Leveraging data from IoT devices, which includes normal traffic instances and non-immersive attacks like DDos and Dos Hulk, we constructed a DL model for intrusion prediction. The model consists of 5 layers: an input layer with 17 dimensions derived from feature engineering on the CICIDS-2017 dataset, and an output layer with 2 dimensions representing class labels (Benign or Attack). Hidden layers are composed of 100 and 50 neurons respectively, utilizing the rectified linear unit (ReLU).

The self-defense framework is integrated into users' head-mounted displays (HMDs), analyzing incoming network traffic for deviations and triggering alarms preemptively for early threat detection. To enhance confidence in our model, we use SHAP for XAI, offering global and instance-specific explanations. SHAP employs game theory to estimate feature importance, selecting them via forward selection or backward elimination. Its strong theoretical foundation and alignment with human intuition make it a reliable choice.

B. Description of Dataset/ Preprocessing

Due to the lack of a VR-specific cybersecurity dataset, the widely used CIC-IDS2017 dataset was adopted for model training. It comprises eight files with traffic data for five days, encompassing various attack types (e.g., DOS Hulk, Portscan, DDoS) and normal traffic. Preprocessing was crucial to optimize DL model efficiency [9]. The data was transformed into a binary classification format by grouping attacks and benign instances. Feature importance-based thresholding led to a 17-dimensional dataset from the original 80 features. Label encoding converted categorical variables to numerical format. The dataset was partitioned into training (70%), testing (20%), and validation (10%) sets. Standard scaling was applied to ensure uniform feature scaling within the range of 0 to 1.

TABLE I
THE DIFFERENT TRAFFIC TYPES, THEIR SAMPLE AND SPLIT SIZES

Traffic_Type	Sample_size	Training_set	Test_set
Benign	654771	523816	130954
FTP Patator	230124	18409	46024
DoS Hulk	158804	12704	32176
SSH Patator	128025	10242	2560
Port Scan	10293	8234	2058
DoS Slowris	7935	6348	1587
DDoS	5897	4717	1179
DoS Slowhttptest	5796	4636	1159
DoS Goldeneye	5499	4399	1099
Malicious	4183	3346	836

C. Experimental Set-up/Hyperparameters

The experiment utilized Python with Tensorflow 2.9.0 on Windows 10 platform. The hardware setup included an Intel(R) Core(TM) i5-7400 CPU @ 3.00GHz processor, 8GB RAM, and a Tesla K80 GPU. Hyper-parameter tuning was manually performed to identify optimal settings listed in TableII.

TABLE II
HYPERPARAMETER USED FOR PROPOSED SCHEME

S/n	Hyperparameters	Value
1	number of layers	5
2	activation function	relu/
3	batch size	32
4	optimizer	adam
5	learning rate	0.001
6	epoch	20
7	loss function	binary cross-entropy

III. PERFORMANCE EVALUATION/RESULT DISCUSSION

A. Performance Evaluation

The XAI-driven framework's evaluation used essential metrics: accuracy, precision, recall, and f1-score. The results showcase its strong performance with 99.0% accuracy, 99.6% precision, 98.6% recall, and 99.0% f1-score respectively.

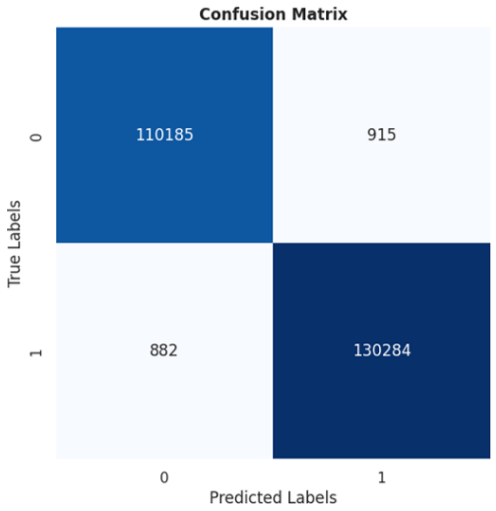

Fig. 2. Confusion Matrix of proposed XAI-driven Model

The confusion metrics in Fig.2 shows the total number of accurate classification, versus the few numbers of misclassified instances. This indicates that the proposed model is efficient in making accurate predictions with fewer errors.

B. Comparison Analysis

The proposed XAI-driven model gave better results compared to the CNN using the same dataset. As shown in Fig.3, our model achieved a 99.2% detection rate, while the CNN achieved 91.9% accuracy. This signifies a substantial 6.1% enhancement in accuracy, despite the models being tested under identical conditions.

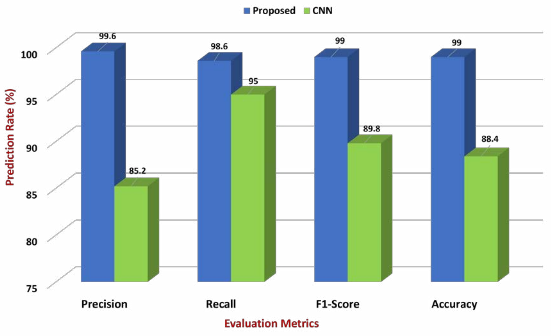

Fig. 3. Performance Comparison of proposed Model XAI-Driven Model and CNN model

C. SHAP-Global Interpretation of XAI-Driven Model

Following model training and evaluation, the subsequent phase focused on providing global and local explanations for predicted outcomes based on SHAP. The CIC-IDS2017 dataset is quite a voluminous dataset and generating an explanation of model decisions on a large dataset can be computationally expensive, therefore to compute SHAP values within a minimal time, a sample size of 10 was selected from the X_train since the goal of XAI is to gain insight on the decisions of a complex model.

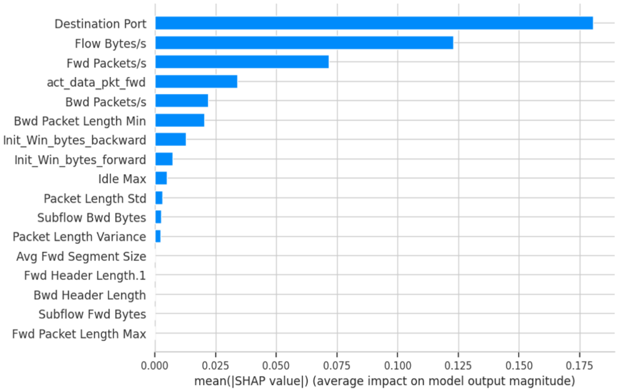

Fig. 4. Global interpretation of model highlighting the most important features and the magnitude of their impact on the model

Fig. 4 provides a global explanation through a summary plot, which highlights the most important features and the magnitude of their impact on the model, offering valuable insight into their significance in the decision-making process.

D. SHAP-Instance Interpretation of XAI-Driven Model

Local interpretation involves comprehending a model's predictions for individual instances. Fig. 5 displays a forced plot, utilized for visualizing the alignment between the "output value" and the "base value". The plot also illustrates which features exert a positive impact (red) or a negative impact (blue) on the prediction, along with their respective magnitudes. A "benign instance" was selected from a sample index value in this particular instance. The result indicates that with a confidence level of 1.00, the predicted output remains benign. Most importantly, the prominently highlighted features in red significantly contributed to this prediction.

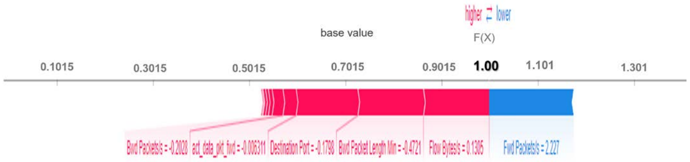

Fig. 5. An instance-wise interpretation of a single datapoint with the positive contributing features and their values displayed in red and the negative contributing feature in blue.

IV. Conclusion and Future Work

In this study, we introduce an innovative XAI-driven model aimed at real-time non-immersive threat detection within the VR environment. Our approach integrates the SHAP-based XAI technique, augmenting our design with enhanced explainability and transparency. This infusion of explainability renders our proposed framework highly credible and dependable, catering to the needs of both cybersecurity and data science analysts for optimal decision-making. Our future endeavors involve investigating alternative XAI techniques on diverse datasets to discern computational efficiency.

V. Acknowledgment

This research 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) and the MSIT(Ministry of Science and ICT), Korea, under the Innovative Human Resource Development for Local Intellectualization support program (IITP-2023-2020-0-01612) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation.

References

[1] M. Chen, W. Saad, and C. Yin, "Virtual reality over wireless networks: Quality-of-service model and dl learning-based resource management," *IEEE Transactions on Communications*, vol. 66, no. 11, pp. 5621–5635, 2018.

[2] D. A. Reddy, V. Puneet, S. S. R. Krishna, and S. Kranthi, "Network attack detection and classification using ann algorithm," in *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 2022, pp. 66–71.

[3] P. Casey, I. Baggili, and A. Yarramreddy, "Immersive virtual reality attacks and the human joystick," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 2, pp. 550–562, 2019.

[4] P. Hu, H. Li, H. Fu, D. Cansever, and P. Mohapatra, "Dynamic defense strategy against advanced persistent threat with insiders," in *2015 IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 2015, pp. 747–755.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[6] Z. Abou El Houda, B. Brik, and L. Khoukhi, ""why should i trust your ids?": An explainable deep learning framework for intrusion detection systems in internet of things networks," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 1164–1176, 2022.

[7] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 1721–1730.

[8] S. Wang, M. Atif Qureshi, L. Miralles-Pechuan, T. Reddy Gadekallu, M. Liyanage *et al.*, "Explainable ai for b5g/6g: technical aspects, use cases, and research challenges," *arXiv e-prints*, pp. arXiv–2112, 2021.

[9] K. Sauka, G.-Y. Shin, D.-W. Kim, and M.-M. Han, "Adversarial robust and explainable network intrusion detection systems based on deep learning," *Applied Sciences*, vol. 12, no. 13, p. 6451, 2022.

[10] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Communications of the ACM*, vol. 63, no. 1, pp. 68–77, 2019.

[11] C. I. Nwakanma, L. A. C. Ahakonye, J. N. Njoku, J. C. Odirichukwu, S. A. Okolie, C. Uzondu, C. C. Ndubuisi Nweke, and D.-S. Kim, "Explainable artificial intelligence (xai) for intrusion detection and mitigation in intelligent connected vehicles: A review," *Applied Sciences*, vol. 13, no. 3, p. 1252, 2023.

[12] H. Li, F. Wei, and H. Hu, "Enabling dynamic network access control with anomaly-based ids and sdn," in *Proceedings of the ACM international workshop on security in software defined networks & network function virtualization*, 2019, pp. 13–16.