1. INTRODUCTION

Our daily lives are changing significantly with the increase in the popularity of IoT. Products such as cars, televisions, and other home appliances relate to the internet and cutting-edge features that change how our daily lives work [1]. IoT devices can interact with one another over the internet, creating a network that enables their interconnection and easier control [2].

With the increasing number of IoT devices, the number of devices that can be used in a potential attack is increasing [3]. This increasing potential attack serves as a hurdle for the privacy and security of IoT systems[4]. These devices can indirectly take valuable personal information, conduct monitoring activities, or ransomware attacks. Due to the broader adoption of IoT, security concerns are already common today. The more traditional security goals of information technology consist of ensuring the system's confidentiality, integrity, and accountability [5].

The Intrusion Detection System (IDS) created by the cybersecurity community is a solution that can detect and monitor attacks continuously. This capability enables responses to dangers bypassing current protection [6]. Intrusion Detection Systems (IDSs) powered by artificial intelligence are becoming more attractive and consistently demonstrate outstanding results in identifying attacks due to their capability to detect new threats [7]. Despite their potential, AI-based methods encounter several challenges. One such significant drawback is the black-box nature of AI models. Due to the nature of the black box, the AI Model often makes it difficult for individuals to understand the outcome because of its lack of reasoning and justification. In this case, the black-box system would hinder security and make the system highly susceptible to data breaches and other threats [8]. However, we can use interpretable machine learning to comprehend how models generate predictions and address questions such as which features are most influential in driving the predictions [9].

Research [7] proposes a method of identifying DDoS attacks based on explainable artificial intelligence (XAI). The proposed method identifies strange behavior in the internet network by analyzing traffic at the network layer. The results of this research show that the proposed method offers better detection accuracy and attack reliability compared to other methods.

Research [10] proposes a framework capable of classifying ordinary traffic and malicious traffic due to DDoS attacks using a Multi-Layer Perceptron network (MLP) and provides an understanding of model decision-making through XAI techniques such as Shapley Additive Explanations (SHAP). The proposed framework's accuracy results are more than 99%.

Research [11] builds a model based on an autoencoder that can detect strange behavior in computer network traffic. The data used comes from CICIDS201 datasets. Based on the dataset, two models were created, with the second model (OPT_Model), which was created using some features based on Shapley values, beating the first model (Model_1), which was created using all the features.

Research [12] presents a new ensemble model that can identify DDoS attacks. The method utilizes machine learning algorithms such as Logistic Regression, Random Forest, Decision Tree, and Extreme Gradient Boosting classifiers to identify and categorize malicious attacks effectively. The XAI models used in this research are Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) to improve readability and transparency. The results show that the XGBoost ensemble model outperforms Logistic Regression, Random Forest, and Decision Tree with 97% accuracy.

Research [13] implements XAI-based solutions using machine learning frameworks. The data used is the CIRA-CIC-DoHB-rw-2020 dataset. For the classification task, this proposed method of balanced and stacked random forest reached high scores on precision (99.91%), recall (99.92%), and F1 score of 99.91%. Using the XAI method, the researcher highlighted the feature's contribution to the model for transparency and explainability.

This research analyzes the method of selecting features to detect attacks on IoT systems. Extreme Gradient Boosting (XGBoost) is used to detect attacks, and the XAI SHAP model is used to improve interpretability and explainability. In this study, the model's results are analyzed.